

# **Understanding L2 Proficiency**

Theoretical and  
meta-analytic investigations

EDITED BY

**Eun Hee Jeon and Yo In'nami**

**John Benjamins Publishing Company**

## Understanding L2 Proficiency

# *Bilingual Processing and Acquisition (BPA)*

ISSN 2352-0531

Psycholinguistic and neurocognitive approaches to bilingualism/multilingualism and language acquisition continue to gain momentum and uncover valuable findings explaining how multiple languages are represented in and processed by the human mind. With these intensified scholarly efforts come thought-provoking inquiries, pioneering findings, and new research directions. The *Bilingual Processing and Acquisition* book series seeks to provide a unified home, unlike any other, for this enterprise by providing a single forum and home for the highest-quality monographs and collective volumes related to language processing issues among multilinguals and learners of non-native languages. These volumes are authoritative works in their areas and should not only interest researchers and scholars investigating psycholinguistic and neurocognitive approaches to bilingualism/multilingualism and language acquisition but also appeal to professional practitioners and advanced undergraduate and graduate students.

For an overview of all books published in this series, please see [benjamins.com/catalog/bpa](http://benjamins.com/catalog/bpa)

## **Executive Editor**

John W. Schwieter  
Wilfrid Laurier University

## **Associate Editor**

Aline Ferreira  
University of California, Santa Barbara

## **Editorial Advisory Board**

Jeanette Altarriba  
*University at Albany, State  
University of New York*

Panos Athanasopoulos  
*Lancaster University*

Laura Bosch  
*Universitat de Barcelona*

Marc Brysbaert  
*Ghent University*

Kees de Bot  
*University of Groningen*

Yanping Dong  
*Zhejiang University*

Mira Goral  
*Lehman College, The City  
University of New York*

Roberto R. Heredia  
*Texas A&M International University*

Arturo E. Hernandez  
*University of Houston*

Ludmila Isurin  
*Ohio State University*

Janet G. van Hell  
*Pennsylvania State University*

Walter J.B. van Heuven  
*University of Nottingham*

Iring Koch  
*RWTH Aachen University*

Li Wei  
*UCL IOE*

Gerrit Jan Kootstra  
*Radboud University Nijmegen &  
Windesheim University of Applied  
Sciences*

Gary Libben  
*Brock University*

Silvina Montrul  
*University of Illinois at Urbana-  
Champaign*

Kara Morgan-Short  
*University of Illinois at Chicago*

Greg Poarch  
*University of Groningen*

Leah Roberts  
*University of York*

Norman Segalowitz  
*Concordia University*

Antonella Sorace  
*University of Edinburgh*

## **Volume 13**

Understanding L2 Proficiency. Theoretical and meta-analytic investigations  
Edited by Eun Hee Jeon and Yo In'nam

# Understanding L2 Proficiency

Theoretical and meta-analytic investigations

*Edited by*

Eun Hee Jeon

University of North Carolina at Pembroke

Yo In'nami

Chuo University

John Benjamins Publishing Company

Amsterdam / Philadelphia



The paper used in this publication meets the minimum requirements of the American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

DOI 10.1075/bpa.13

**Cataloging-in-Publication Data available from Library of Congress:  
LCCN 2022012178 (PRINT) / 2022012179 (E-BOOK)**

ISBN 978 90 272 1117 0 (HB)

ISBN 978 90 272 5769 7 (E-BOOK)

© 2022 – John Benjamins B.V.

This e-book is Open Access under a CC BY-NC-ND 4.0 license.

<https://creativecommons.org/licenses/by-nc-nd/4.0>

This license permits reuse, distribution and reproduction in any medium for non-commercial purposes, provided that the original author(s) and source are credited. Derivative works may not be distributed without prior permission.

This work may contain content reproduced under license from third parties. Permission to reproduce this third-party content must be obtained from these third parties directly.

Permission for any reuse beyond the scope of this license must be obtained from John Benjamins Publishing Company, [rights@benjamins.nl](mailto:rights@benjamins.nl)

John Benjamins Publishing Company · <https://benjamins.com>

# Table of contents

Acknowledgements	VII
Notes on Authors	IX
CHAPTER 1	
Introduction	1
<i>Jan Hulstijn</i>	
CHAPTER 2	
L2 reading comprehension: Theory and research	5
<i>Junko Yamashita</i>	
CHAPTER 3	
L2 reading comprehension and its correlates: An updated meta-analysis	29
<i>Eun Hee Jeon and Junko Yamashita</i>	
CHAPTER 4	
L2 writing: Theory and research	87
<i>Rob Schoonen</i>	
CHAPTER 5	
L2 writing and its internal correlates: A meta-analysis	109
<i>Masumi Kojima and Taku Kaneta</i>	
CHAPTER 6	
L2 writing and its external correlates: A meta-analysis	159
<i>Masumi Kojima, Yo In'nami and Taku Kaneta</i>	
CHAPTER 7	
L2 listening comprehension: Theory and research	213
<i>Elvis Wagner</i>	
CHAPTER 8	
L2 listening and its correlates: A meta-analysis	235
<i>Yo In'nami, Rie Koizumi, Eun Hee Jeon and Yuya Arai</i>	

CHAPTER 9	
L2 speaking: Theory and research	285
<i>Jie Gao and April Ginther</i>	
CHAPTER 10	
L2 speaking and its internal correlates: A meta-analysis	307
<i>Rie Koizumi, Yo In'nami and Eun Hee Jeon</i>	
CHAPTER 11	
L2 speaking and its external correlates: A meta-analysis	339
<i>Eun Hee Jeon, Yo In'nami and Rie Koizumi</i>	
CHAPTER 12	
Discussion, limitations, and future research	369
<i>Eun Hee Jeon, Yo In'nami and Rie Koizumi</i>	
Index	387

## Acknowledgements

The editors of this volume sincerely thank the authors for their invaluable contributions. We would also like to thank the chapter reviewers, Jon Clenton, Sara Cushing, Nivja H. de Jong, Jason Fan, Zhi Li, Ryan Miller, Shangchao Min, Matthew Wallace, and Xun Yan among others, for their insights and expertise which greatly improved the quality of the earlier versions of this volume. We also thank the library staff at Chuo University, University of North Carolina at Pembroke, Waseda University, Kaizhou Luo, Jiayu Wang, and Andrew J. Burns for helping us as we conducted literature search for our meta-analyses. We also deeply appreciate Jan Hulstijn for his encouragement and support from the beginning to the end of this project. Lastly, we send our sincere thanks to the Bilingual Processing and Acquisition Series Editor, John W. Schwieter and Kees Vaes from John Benjamins Publishing Company for their support and incredible patience.

Additionally, we would like to note that an earlier version of Chapter 3 (L2 reading comprehension and its correlates: An updated meta-analysis) appeared as Jeon and Yamashita (2017, 2021). An earlier version of Chapter 5 (L2 writing and its internal correlates: A meta-analysis) appeared as Kojima (2017, 2020) and Kojima and Kaneta (2020). An earlier version of Chapter 6 (L2 writing and its external correlates: A meta-analysis) appeared as Kojima, In'nami, and Kaneta (2021). An earlier version of Chapter 10 (L2 speaking and its internal correlates: A meta-analysis) appeared as In'nami, Koizumi, and Jeon (2017). An earlier version of Chapter 11 (L2 speaking and its external correlates: A meta-analysis) appeared as Jeon, In'nami, and Koizumi (2016a), Jeon, In'nami, and Koizumi (2016b), and Jeon, In'nami, and Koizumi (2017).

This book was published open access funded by the Japan Society for the Promotion of Science (JSPS) KAKENHI, Grant-in-Aid for Scientific Research (C) Grant Numbers 19K00781, 20K00894, and 20K00781. Supplementary material is available at <https://osf.io/edzqa/>.

Yorktown, USA / Tokyo, Japan, May 2022  
Eun Hee Jeon & Yo In'nami



## References

- In'nami, Y., Koizumi, R., & Jeon, E. (March, 2017). L2 speaking proficiency and its features: A meta- analysis. *American Association for Applied Linguistics*, 2017, Portland, OR.
- Jeon, E. (June, 2017). Colloquium "Understanding L2 proficiency through meta-analysis: L2 reading, speaking, writing and their correlates." Presenters: Jan H. Hulstijn, Eun Hee Jeon, Junko Yamashita, Yo In'nami, Rie Koizumi, Masumi Kojima, International Symposium of Bilingualism, 2017, Limerick, Ireland (Organizer).
- Jeon, E., In'nami, Y., & Koizumi, R. (2016a). L2 speaking proficiency and its correlates: A meta-analysis of correlation-coefficients. *American Association for Applied Linguistics*, 2016, Orlando, FL.
- Jeon, E., In'nami, Y., & Koizumi, R. (2016b). L2 speaking proficiency and its correlates: A meta-analysis of correlation-coefficients. *Second Language Research Forum*, 2016, New York City, NY.
- Jeon, E., In'nami, Y., & Koizumi, R. (June, 2017). L2 speaking proficiency and its correlates: A meta- analysis. Paper presented as a part of a colloquium titled "Understanding L2 proficiency through meta-analysis: L2 reading, speaking, writing and their correlates." organized by E. Jeon. International Symposium of Bilingualism, 2017, Limerick, Ireland.
- Jeon, E., & Yamashita, J. (June, 2017). L2 reading comprehension and its correlates: A meta-analysis. Paper presented as a part of a colloquium titled "Understanding L2 proficiency through meta- analysis: L2 reading, speaking, writing and their correlates." organized by E. Jeon. International Symposium of Bilingualism, 2017, Limerick, Ireland.
- Jeon, E., & Yamashita, J. (March, 2021). L2 reading comprehension and its correlates: An updated meta- analysis. *American Association for Applied Linguistics*, 2021, Houston, Texas.
- Kojima, M. (2017). Correlations between ESL/EFL writing performance and its analytic features: A meta-analysis. Paper presented as a part of a colloquium titled "Understanding L2 proficiency through meta-analysis: L2 reading, speaking, writing and their correlates." organized by E. Jeon. International Symposium of Bilingualism, 2017, Limerick, Ireland.
- Kojima, M. (2020). A study synthesis on the relationship between second language writing performance and text features: Focusing on text-based measures and study features. *Learner Corpus Studies in Asia and the World*, 5, 1–24. <https://doi.org/10.24546/81012486>
- Kojima, M., & Kaneta, T. (2020). Raitingu hyoka to gengoteki shihyo no kankei: Meta bunseki ni yoru kenkyu seika no togo [The relationship between writing performance and linguistic indices: A meta-analysis]. In Y. Ishii & Y. Kondo (Eds.), *Eigo Kyoiku ni okeru jido saiten: Genjo to kadai* [Automated scoring in English language education: Its current situation and issues] (pp. 33–72). Tokyo, Japan: Hituzi Shobo.
- Kojima, M., In'nami, Y., & Kaneta, T. (2021, August). Writing ryoku to component skills no soukan kankei [A meta-analysis of writing and its components]. Paper presented at the 46th Japan Society of English Language Education, Nagano, Japan.

## Notes on Authors

**Yuya Arai** holds a degree of Master of Arts in Education from Waseda University, Japan. He currently teaches English as a Foreign Language (EFL) courses at a senior high school. He is interested in L2 extensive reading and language assessment. His most recent publication is about pre-service EFL teachers' perceptions of extensive reading, published in *Journal of Extensive Reading*. He pursues his PhD degree starting April 2022.

**Jie Gao** is a Lecturer of English at Fudan University, China. She is interested in investigating linguistic profiles demonstrated by L2 English learners at different language proficiency levels. Her research also focuses on the design of academic English courses and English writing center administration in L2 contexts. Her most recent publication in *Applied Corpus Linguistics* examines the possible relationship between novice L2 English writers' citation practices and the pedagogical materials used in writing classrooms.

**April Ginther** is an applied linguist and a professor at Purdue University where she directs The Oral English Proficiency Program (OEPP) and The Purdue Language and Cultural Exchange (PLaCE). She is responsible for the local exams used to evaluate the English language skills of approximately 1000 incoming international students each academic year. Her research is primarily concerned with locally developed, maintained, and administered language tests. From 2012 to 2017, she was the co-editor of *Language Testing*.

**Jan Hulstijn** is a Professor Emeritus of Second Language Acquisition at the University of Amsterdam. His interests concern language proficiency in native and non-native speakers, usage-based and neural-network accounts of first and second language acquisition, and philosophy of science. In 2018 he received the 2018 distinguished scholar award from the European Second Language Association (EuroSLA).

**Yo In'nami** is a Professor of English at Chuo University, Japan. He is interested in meta-analytic inquiry into the variability of effects and the longitudinal measurement of change in language proficiency. He has taught various courses in second language acquisition and language testing at undergraduate and postgraduate levels. His most recent publication reported on a meta-analysis of the relationship between working memory capacity and L2 reading, published in *Studies in Second Language Acquisition*.

**Eun Hee Jeon** is an Associate Professor of English at the University of North Carolina at Pembroke. She teaches various courses in TESOL, applied linguistics, and linguistics at the undergraduate and graduate level. Her research interests include second language (L2) reading comprehension, L2 proficiency, and research methods.

**Taku Kaneta** is a lecturer at the Department of School Education, Teikyo University of Science, Tokyo, Japan. His research interests include lexicographical user research, corpus linguistics, vocabulary acquisition, and playful language education. He has been teaching English in Japan and has dedicated his career to sparking motivation in learners. His most recent publication appeared in *Gamevironments*, and explored classroom practices that use board games.

**Rie Koizumi** is a Professor of English at Seisen University, Japan. Her research interests include assessing and modeling second language ability, performance, and development. She has published her work in *Language Testing*, *Language Assessment Quarterly*, *System*, *Assessment in Education: Principles, Policy & Practice*, and other journals. She is editing a book that brings together testing and assessment practices in secondary schools in Japan.

**Masumi Kojima** is an Associate Professor of Applied Linguistics at the Department of English in Gifu City Women's College, Japan. Her research interests include vocabulary acquisition, writing assessment, and meta-analytic inquiry into writing proficiency and individual variables. Her recent work appeared in K. Koda and J. Yamashita (Eds.), *Reading to learn in a foreign language: An integrated approach to foreign language instruction and assessment*.

**Rob Schoonen** is Chair Professor of Applied Linguistics at the Department of Language and Communication and the Centre for Language Studies of the Radboud University Nijmegen. His research concerns psycholinguistic aspects of language proficiency, language comprehension and language production, and the assessment of language proficiency. He holds a PhD in Psychology from the University of Amsterdam.

**Elvis Wagner** is an Associate Professor of TESOL at Temple University. His research focuses on second language assessment, especially the assessment of L2 listening ability and oral communicative competence. His primary research focus examines how L2 listeners process and comprehend unscripted, spontaneous spoken language. He is the co-author (with Gary Ockey) of *Assessing L2 listening: Moving towards authenticity* published in 2018.

**Junko Yamashita** is a Professor at the Graduate School of Humanities, Nagoya University, Japan. Her main research areas are second language reading, lexical processing, and meta-analysis. Her work has appeared in many journal articles, book chapters, and co-edited books. She has supervised MA and PhD theses and taught postgraduate courses in second language acquisition and foreign language teaching as well as various undergraduate EFL courses.



## Introduction

Jan Hulstijn

University of Amsterdam

Every person who has attended high school is familiar with the term ‘language proficiency’. It means the ability or skill to comprehend, speak and write a language well and is usually associated with education and career: instruction, learning, taking exams, and obtaining certificates. For foreign/second language (L2) professionals, language proficiency is foremost the business of assessment. Assessment of L2 proficiency has become an industry, where commercial companies and public institutions invest (and sometimes earn) large amounts of money. Before constructing and administering a proficiency test, specialists, working in the assessment field, must answer the question “What is language proficiency?” It is therefore no surprise that language proficiency has become one of the objects of study of ‘applied’ linguists, in particular language-testing specialists. Over the last 50 years, many books, conference presentations, and papers in international academic journals were devoted to the ‘construct’ of language proficiency: Should language proficiency be seen as a unitary construct or does it consist of components? How loosely or tightly do components hang together? To what extent is language proficiency related to, or even dependent on, other mental abilities?

In cognitive psychology, similar questions arose concerning the componential structure of intelligence and memory. The scientific study of these questions benefitted fruitfully from the Cognitive Revolution in psychology and linguistics, which allowed researchers to study what goes on in the ‘black box’ of the human mind, as behaviourists had earlier called it. The empirical study of the components of intelligence, memory, and language proficiency (associated with giants such as John B. Carroll, J. Paul Guilford, and Alan Baddeley) also benefitted from developments in psychological measurement (based on seminal work of Louis Leon Thurstone, Charles Spearman, and others), with increasingly more sophisticated statistical analyses, such as (confirmatory) factor analysis and structural equation modelling.

In this rich tradition, in the last three decades of the last century, various so-called ‘models’ of L2 proficiency were proposed, associated with, among others, John Oller Jr., Michael Canale together with Merrill Swain, and Lyle Bachman. With the number of empirical studies growing, and facilitated by new statistical

techniques, called meta-analyses, the need arose to compare the findings of dozens of studies, and to test the empirical robustness of claims made in the theoretical literature. The number of studies on testing proficiency in English as a second language (listening, speaking, reading, and writing) had become so large that the time had come for such meta-analyses.

Around five years ago, this need was recognized by Eun Hee Jeon and Yo In'nam, the editors of this volume, and their associates Yuya Arai, Taku Kaneta, Rie Koizumi, Masumi Kojima, and Junko Yamashita. I first heard about Jeon and Yamashita's meta-analyses when they presented their ongoing work at the 2011 *Second Language Research Forum*, held at Iowa State University. When their first big study (on L2 reading) appeared in *Language Learning* (2014), I just managed to include a reference to it in my language-proficiency book (2015) before it went in press. Shortly after the publication of the 2014 study, Eun Hee Jeon and her associates started planning a number of big meta-analyses and bringing these together in a volume. Preliminary findings of some of these meta-analyses were presented in a colloquium at the International Symposium of Bilingualism in 2017 (Limerick, Ireland) in which I participated as a discussant. As anyone browsing Chapters 3, 5, 6, 8, 10, and 11, will immediately see, conducting such work requires the highest expertise, and is tremendously time consuming. But here they are, sisterly together: six meta-analytical studies, forming a gold mine for researchers interested in the componential structure of language proficiency and its associations with other cognitive abilities. The volume also includes four chapters on theory and research in L2 listening (Elvis Wagner), speaking (Jie Gao & April Ginther), reading (Junko Yamashita), and writing (Rob Schoonen), bringing readers up to date with recent models and insights. A comparison with a similar update, published in *Annual Review of Applied Linguistics* (Vol. 18, 1998), shows that the field has developed substantially.

This volume allows readers to harvest and reflect on which associations between which factors can be said to be empirically robust, in which L1-L2 combinations, and in which contexts of learning and using English as a second language. Some conclusions can be safely drawn (see Ch. 12 for the discussion of all meta-analysis chapters). However, in scientific inquiry there is never certainty. Empirical observations have to be interpreted and explained, and explanations may change when scientific paradigms change. The best known models and theories of language proficiency were proposed between 1970 and 2000, during the heyday of the first wave of the Cognitive Revolution. That was the time when cognitive abilities in the human mind (e.g., visual perception, reasoning, learning and memory) were conceptualized in a 'box-and-arrow' fashion, with the boxes consisting of 'modules'. During the same period, generative linguistics was dominated by the question of whether the relation between sound and meaning was mediated by syntax and, if so, how the 'interfaces' between linguistic modules should be conceived. In cognitive

psychology of that time, models of information processing distinguished between dynamic ‘information processes’ and static ‘modules’ (representations of information), such as sensory perception and short-term memory. Processes (often visualized as circles or ovals) formed the connections between the modules (visualized as square boxes). Carroll’s (1993) *Three-Stratum Theory* of cognitive abilities formed the culmination of decades of correlational work.<sup>1</sup>

But the first wave of the Cognitive Revolution was followed by a second wave, manifested by usage-based linguistics and neural-network psychology. Under these views, first-language acquisition (before the acquisition of literacy skills) is a matter of implicit, bottom-up statistical learning, including self-organization in a multi-layered neural network, meaning that higher-order linguistic patterns probabilistically emerge from lower-order elements and patterns. Thus, under this view, there is, for example, no longer a clear border line between syntax and lexis. The brain/mind does not consist of neatly isolated modules (albeit that some areas are more typically involved in processing certain information than others). Furthermore, language (language in the individual speaker as well as language in a community of its speakers) is seen as a complex system, characterized by unequal distributions of its elements and variability in language productions. This raises the question of whether it is possible to assess, in a valid and reliable manner, a person’s linguistic repertoire by observing the person’s language production, elicited with ‘open’ speaking or writing tasks (in contrast to tests of vocabulary or grammar of the ‘closed’, discrete-point type). Can such an assessment only be successful with respect to the most frequent elements and patterns that typically occur in a certain discourse genre?

The studies whose findings are being ‘meta-analyzed’ in this volume, are invariantly of the correlational type. Regression analyses, factor analyses and structural equation models can only be meaningful, if participants’ test scores differ sufficiently. However, as De Jong and Verhoeven (1992, p. 10) remarked,<sup>2</sup> “because

---

1. I had the privilege of listening to a fascinating exchange of views between John Oller Jr., who had proposed a unitary model of language proficiency, and John B. Carroll, at the 1981 *LSA/TESOL Summer Institute*, held at the University of New Mexico. Carroll successfully convinced Oller that Oller had overinterpreted the outcomes of principal component analyses. Two years later, Oller published an edited volume (*Issues in Language Testing Research*, Newbury House, 1983), including a chapter written by Carroll. In the Introduction, Oller clearly stated (p. xiv) that “the strongest form of the unitary factor hypothesis is untenable”. Scholars who publicly acknowledge that they have been wrong, deserve our respect and should be the models of every earnest researcher.

2. De Jong, J. H. A. L. & Verhoeven, L. (1992). Modeling and assessing language proficiency. In L. Verhoeven & J. H. A. L. de Jong (Eds.), *The construct of language proficiency* (pp. 3–19). John Benjamins.



the factorial approach is based on individual differences, it will always fall short in revealing those basic components of language behavior that are likely to be mastered by all individuals in a relevant population.” This remark might stimulate us to construct a theory that accounts not only for individual differences but for commonalities as well. Person attributes, such as hearing ability, working-memory capacity, non-verbal intelligence, level of education, motivation to learn a language, and several other factors, have never been found to account for high amounts of variance (above 50%) in measures of language proficiency. Is this observation related, or not, to the ubiquitous phenomenon of typicality, variability and unequal distributions of linguistic elements in language production? These are challenging issues for future theoretical and empirical work.

In scientific inquiry, researchers always stand upon the shoulders of others. Innovative, new insights can only emerge in scholars who have made themselves thoroughly familiar with extant empirical findings. With the substantial accumulation of empirical research on L2 proficiency, in particular over the last 30 years, researchers of language proficiency will surely be extremely grateful to Eun Hee Jeon, Yo In’ami, Yuya Arai, Taku Kaneta, Rie Koizumi, Masumi Kojima, and Junko Yamashita for bringing all these studies together and analyzing their findings with state-of-the-art meta-analytical tools. Thank you for this big service to the field!

## References

- Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511571312>

# L2 reading comprehension

## Theory and research

Junko Yamashita

Nagoya University

Multicomponent views of reading constitute a contemporary standard platform toward a more comprehensive understanding of reading and its development. Based on component approaches to L2 reading comprehension, this chapter first outlines linguistic and general cognitive processing necessary for understanding text meaning, and then synthesizes current research insights into components of L2 reading comprehension covering differences between L1 and L2 reading and the cross-linguistic nature of L2 reading. In so doing, the chapter discusses influential theories in the relevant research areas. The chapter closes with discussions on theories and methodologies of L2 reading research.

### 1. Introduction

Reading comprehension is an interactive and constructive process where readers create their meaning representation by integrating the meaning built up from the text and their background knowledge stored in memory. It is an enormously complex skill that is accomplished by harmonious orchestration of the diverse mental processes working in tandem according to the purpose of reading. Because of its complexity, there are numerous theories and models of reading comprehension explaining its elements and how they work in the process of reading (Grabe, 2009). L2 reading comprehension is even more complex due to the involvement of dual languages (first language [L1] and second language [L2]) and a diversity of readers ranging from adult foreign language learners to language minority children.

Multicomponent views of reading comprehension constitute a contemporary standard platform toward a more comprehensive understanding of reading and its development. This chapter draws on component approaches (Carr & Levy, 1990) and synthesizes current insights into components of L2 reading comprehension covering differences between L1 and L2 reading and cross-linguistic nature of L2 reading.

## 2. Linguistic processing in reading

To integrate readers' background knowledge in the meaning-construction process of reading, accurate and facile text-meaning building is indispensable. This process is primarily supported by various linguistic processes.

### 2.1 Phonological processing

Phonological processing refers to the ability to use phonological information to process oral and written languages. Since writing systems encode the spoken language, phonological processing is universally involved; however, the way in which orthography maps onto phonology and morphology are different across languages, and, therefore, the reader needs to acquire phonological processing skills that accommodate a specific sound-symbol mapping system in the language they read (Perfetti, 2003).

Phonological processing involves various interrelated skills such as phonological awareness, decoding, encoding, and memory (Nassaji, 2014). Phonological awareness is the ability to reflect upon and manipulate phonological constituents in oral language. Even if it does not involve processing written symbols, phonological awareness is an indispensable pre-literate foundational skill since the first step of reading requires mapping written symbols to sounds. Phonological decoding (or simply decoding) means the ability to transform written symbols to their phonological forms utilizing grapheme-phoneme correspondence rules. Phonological encoding is the capacity to access lexical meaning through phonological information rather than visual-orthographic information (Coltheart, Curtis, Atkins, & Haller, 1993). Phonological memory is the capacity to maintain information phonologically for temporary storage in working memory. Because phonologically encoded information is durable in working memory, efficient use of phonological information is expected to facilitate the integration of word meaning into text meaning building (Koda, 2019). Overall, a consensus view is that efficient phonological processing is causally related to many aspects of reading including word reading, word learning, and text comprehension (Koda, 2019).

### 2.2 Orthographic processing

Orthographic processing refers to the use of visual-orthographic information in recognizing written words. Two types of orthographic knowledge are distinguished in the literature (Hagiliassis et al., 2006): word-specific knowledge and general orthographic knowledge. The former relates to information about the spelling of

a specific word including knowledge of letters and their shapes, and the latter is pertinent to conventional patterns of letter combinations in a given language. The general orthographic knowledge is more abstract in nature, entailing a range of orthographic information such as sequential dependencies, structural redundancies, and letter position frequencies (Conrad, Harris, & Williams, 2013), and thus can be applied to learning new word spellings.

Orthographic processing is useful when reading irregularly spelt words (e.g., yacht). Since grapheme-phoneme correspondence rules are not applicable, we typically access the representation of such words in a mental lexicon via visual orthographic information (Cook & Bassetti, 2005). Orthographic processing is also critical in reading words by sight from memory. Reading words within one second of seeing them indicates sight word reading (Ehri, 2007). Such efficient word recognition is achieved based on the mappings between orthographic patterns and their pronunciations along with other features (e.g., meanings) that are firmly established in memory. Orthographic mapping is achieved for both word-specific knowledge (typically in monosyllabic, high frequency words) and general orthographic knowledge (typically in recurring spelling patterns in multisyllabic words) (Ehri, 2007). Thus, together with phonological processing, orthographic processing contributes to reading comprehension via efficient word recognition.

### 2.3 Morphological processing

A morpheme is the smallest unit of meaning and serves as the basis of word formation. Morphological knowledge is regarded as central to all skills involved in understanding text meaning (Koda, 2019). This may be because it is integrally related to other aspects of language as morphemes have semantic, syntactic, and phonological properties (Kuo & Anderson, 2006). The importance of morphological processing increases at more advanced levels of reading development (middle to upper primary school) when texts contain increasing proportions of morphologically complex words (Anglin, 1993). Many alphabetic languages encode morphemes in their writing systems, and orthographic representations tend to give more priority to morphemes than phonology (a principle called the isomorphism); thus the same morpheme tends to be spelled identically despite different pronunciations (*sign-signature*, *heal-health*) (Kuo & Anderson, 2006). This feature suggests benefits of morphological knowledge in reading or inferring the meaning of morphologically complex words by identifying familiar elements in unfamiliar words.

Morphological knowledge may not always be useful in the inference of word meanings, because grasping derivational relationships through morphological decomposition and recombining morphological parts into the meaning of a whole

word is not always straightforward. However, the literature suggests advantages. In an oft-cited analysis of printed school English by Nagy and Anderson (1984), about 86%, 84%, and 73% of suffixed, prefixed, and compound words are semantically transparent (meanings are inferable from the base word), and semantically transparent derivatives are more numerous than basic and semantically opaque words among low frequency words with U values less than 1, which is, according to White, Power and White (1989), the region of most active vocabulary growth by typical grade 4 students. There is an increasing body of research supporting the positive relationship between morphological awareness and various reading-related skills from word to text levels (Zhang, 2017).

## 2.4 Lexical processing

The primary step of lexical processing is retrieving word meanings from a written word. The basis of this process is vocabulary knowledge. The strong association between vocabulary knowledge and reading comprehension is a known fact. Perfetti and Hart (2002) claims “[r]eading is about words... skill in reading comprehension rests to a considerable extent on knowledge of words” (p. 189). Knowledge of word meanings not only contributes to building text meaning but also activates readers’ knowledge of the world. Moreover, word meanings serve as “passcodes to experience-based knowledge in memory” (Koda, 2019, p. 36) because word meanings are connected to personal experiences as well as more general world knowledge. Among many theories, Lexical Quality Hypothesis (Perfetti & Hart, 2002) most clearly articulates the centrality of vocabulary knowledge in text comprehension. This hypothesis notes that not only vocabulary size but also the quality of knowledge is indispensable, which encompasses refined representations in form, meaning, and use of words. High lexical quality leads to efficient word recognition, a requisite for reading comprehension.

Vocabulary knowledge is multifaceted (Kieffer & Lesaux, 2012; Nation, 2013), and the quality dimension relates to vocabulary depth. Although consensus has yet to be reached on the construct of vocabulary depth, reading studies on the role of vocabulary depth supported its contribution to reading comprehension (Binder, Cote, Lee, Bessette, & Vu, 2017; Leider, Proctor, Silverman & Harring, 2013; Proctor et al., 2012; Qian, 2002; Qian & Schedl, 2004).

Lexical processing also involves word-to-text integration; namely, readers integrate the emerging meaning of a text to retrieve a context-appropriate word meaning (Perfetti & Stafura, 2014). Thus, efficient lexical processing depends on reciprocal interaction of readers’ vocabulary knowledge and their text comprehension. In this sense, word meaning comprehension is already part of text comprehension.

## 2.5 Syntactic processing

The meaning of each word needs to be integrated into larger linguistic units such as phrases, clauses and sentences to be finally integrated into text meaning. Syntactic knowledge plays a critical role in this process. The recognition of the importance of syntactic knowledge along with vocabulary is reflected in the fact that many readability formulae utilize various indices of lexical and syntactic complexity to predict the readability of texts. However, syntactic knowledge has received much less attention than other component skills in L1 reading research (Koda, 2005). This is because L1 children with normal language development acquire basic syntactic knowledge prior to schooling, and little individual difference is expected. Although weaker students may be exceptional (Marx et al., 2015), syntactic knowledge was either an insignificant or only a weak predictor of reading comprehension in L1 children in the middle to upper grade levels in primary school (Geva & Farnia, 2012; Goff et al., 2005).

Contrary to the research done on L1 readers, there is strong support for the importance of syntactic knowledge for explaining variance in L2 reading comprehension both in children (Geva & Farnia, 2012) and adults (Shiotsu, 2010).

## 2.6 Reading fluency

The ability to read connected texts fluently is one of the essential skills for successful reading comprehension (National Reading Panel, 2000). Reading fluency is a joint construct of accuracy and speed. Its importance is explained by automaticity-based theories that presume the limited capacity of human information processing (LaBerge & Samuels, 1974; Perfetti, 1985). When basic linguistic processing such as word recognition and syntactic parsing is carried out accurately, rapidly, and effortlessly, readers allocate residual cognitive resources to higher-order comprehension operations, which ultimately facilitate reading comprehension.

Reading fluency is manifested in two modes: oral and silent reading. Although it is ultimately the fluency in silent reading that matters beyond the elementary level, research has concentrated on oral reading fluency for several reasons, including its educational value (a useful tool for progress monitoring in a classroom) and higher reliability in assessment (based on an observable behavior). However, research in silent reading fluency is growing and has indicated that reading fluency in the two modes are highly related and share a similar set of subcomponents, despite being still distinct constructs (Bar-Kochva, 2013; Kim, Wagner, & Foster, 2011).

Reading fluency emerges in different linguistic units: sublexical, lexical, sentential, and textual. Clearly, underlying subcomponents are different. Word reading

fluency is built upon automaticity in lexical and sublexical processes; sentence reading fluency additionally includes syntactic parsing; text reading fluency further involves discourse level integration. Due to this difference, reading fluency that explains reading achievement may be different at different developmental stages; for instance, text reading fluency may be more important than word reading fluency in older readers (Geva & Farnia, 2012).

### 3. Cognitive processing in reading

Although linguistic resources are essential for text meaning building, general cognitive processes are also critical for reading and its development. General cognitive processes refer to a broad range of mental processes and abilities (e.g., associative learning, cognitive flexibility, processing speed, inhibitory control, self-regulation, and long-term/working memory) that underlie virtually all types of cognitive skills that human beings acquire, and “most can be viewed as parts of humans’ more primal survival mechanisms” (Grabe & Stoller, 2020, p. 14). Some of these operations are not yet commonly discussed in reading research, but some are studied under the notion of reading strategies.

Numerous reading strategies are recognized, such as goal-setting, questioning, predicting, summarizing, and comprehension monitoring (Hudson, 2007). Efficient use of these strategies could facilitate comprehension or resolve problems encountered during reading, and what ultimately matters is the flexible use of multiple strategies according to the purpose and context of reading. As such, researchers have long recognized the significance of mechanisms that orchestrate various reading strategies.

Metacognition is one of the overarching concepts tapping into the higher-order executive function and defined as “knowledge and cognition about cognitive phenomena” by Flavell (1979 p. 906), who is recognized as a founder of this concept. Below is a summary of the conception of metacognition in the current literature (Baker & Beall, 2009; Schmitt, 2005). It has two components: metacognitive knowledge/awareness and metacognitive control/regulation. In the former, there are three types of knowledge in three broad categories: declarative (knowing *what/that*), procedural (knowing *how*), and conditional (knowing *when/why*) knowledge about person/self, task, and strategies. Likewise, in metacognitive control, three types of processing are proposed: planning, monitoring, and revising, each of which is actualized through various strategies such as previewing, predicting, confirming/disconfirming predictions, re-reading, and problem-solving.

There are developmental and individual differences in metacognition in reading, with older and more skilled readers demonstrating more appropriate metacognitive

awareness (e.g., reading is a process of decoding vs. meaning-getting) and better comprehension monitoring from early childhood through later adulthood (Baker & Beall, 2009). Despite theoretical importance, empirical findings are mixed regarding the effect of metacognitive knowledge on L2 reading comprehension, with the magnitude ranging from strong (e.g., van Gelderen et al., 2007; van Steensel et al., 2016) to modest or even negligible (Jeon, 2011; Shiotsu, 2010).

#### 4. Simple View of Reading

The Simple View of Reading (SVR) (Gough & Tunmer, 1986) is one of the most influential models guiding component approaches (most typically L1 reading and L2 reading in bilingual contexts). Because many studies discussed in this chapter are based on this model, this section covers this model to lay the groundwork for the following discussions of recent studies.

The SVR postulates that reading comprehension is explained by decoding and linguistic comprehension. Decoding is a reading skill and defined as efficient context-free word recognition, while linguistic comprehension is in the domain of oral language and defined as the ability to use lexical information and construct sentence and discourse level interpretations (Hoover & Gough, 1990). The SVR has received substantial empirical support and been established in the field (e.g., Kirby & Savage, 2008).

An important developmental phenomenon has been identified based on the SVR; decoding tends to be more important in lower grade levels (*learning to read* stage), but linguistic comprehension gains its importance in more advanced levels (*reading to learn* stage) (Gough, Hoover, & Peterson, 1996; Hoover & Gough 1990). This is because beginning readers vary in their decoding skills and texts they read are simple in language and contents; therefore, once decoding is completed, comprehension is not very challenging for normally developing children. At this stage, variance in decoding skills is a major factor influencing reading comprehension. However, in more advanced levels when children have largely mastered decoding skills, reading texts become more and more demanding both linguistically and conceptually; thus, linguistic comprehension ability takes over decoding and becomes more responsible for individual differences in reading comprehension.

Although the model appears too simplistic to capture the complexity of reading, the SVR's strength lies in its simplicity (scientific reductionism [Kirby & Savage, 2008]). Since the SVR includes two central components, and both are multifaceted and supported by various subcomponents, subsequent research can use the SVR as a starting point and expand the scope of investigation by adding or modifying the original model. Initially, decoding and linguistic comprehension were



operationalized by pseudoword reading and listening comprehension (Hoover & Gough, 1990). Later, decoding has been operationalized by other measures as well, such as real word reading, letter knowledge, phonological awareness, and reading fluency, and linguistic comprehension by oral vocabulary and syntactic knowledge in addition to or in place of listening comprehension.

Different motivations are observed in the expansion of measurement. Some researchers conceptualize their addition as subcomponents of the SVR's components (e.g., measuring both vocabulary and listening comprehension for linguistic comprehension). Thus, they increase methodological rigor by measuring the components more comprehensively and support the model more strongly. Others see their new measures as a new addition to the model (e.g., adding reading fluency to the model [Joshi & Arron, 2000]). In this case, researchers propose modifications to the SVR. The SVR was also used to examine interactions between linguistic processes (those in the SVR) and cognitive processes (Cirino et al., 2019). Thus, the SVR has been serving as a platform from which researchers communicate their findings.

## **5. Differences between L1 and L2 reading in component skills**

There are similarities and differences in L1 and L2 reading. The basic reading processes and components supporting each process are primarily common; for instance, foundational skills such as word awareness, letter knowledge, and phonemic awareness predict word recognition and reading fluency both in L1 and L2 reading (Lesaux, Koda, Siegel, & Shanahan, 2006). However, there are important differences as well.

It is widely known that learners reading in L2 on average lag behind their monolingual counterparts. However, this does not mean L2 readers are weak in all component skills. Generally, they develop word level skills on a par with their L1 peers in early grade levels but stay weaker in oral language proficiency even after several years of schooling (Lesaux et al., 2006).

A meta-analysis by Melby-Lervåg and Lervåg (2014) supports this general trend. They compared L1 and L2 readers in four areas: reading comprehension, language comprehension (oral language skills operationalized by oral vocabulary or listening comprehension), decoding, and phonological awareness. The difference reached statistical significance in reading comprehension ( $d = -0.62$ ), language comprehension ( $d = -1.12$ ), and decoding ( $d = -0.12$ ) in favor of L1 readers but was not significant in phonological awareness ( $d = -0.08$ ). The researchers attributed this pattern to the relative simplicity of phonological awareness compared to other skills.

Findings from empirical studies converge on this trend. Differences were not significant in phonological awareness (Geva & Farnia, 2012; Kieffer & Vukovic, 2013; Marx et al., 2015) and decoding/word reading skills (Babayiğit, 2014, 2015; Droop & Verhoeven, 2003; Geva & Farina, 2012; van Gelderen et al., 2003; van Steensel et al., 2016), but significant in one or more skills in listening comprehension, vocabulary, and syntactic knowledge (Babayiğit 2014, 2015; Droop & Verhoeven 2003; Geva & Farnia, 2012; Kieffer & Vukovic, 2013; Kim 2012; Marx et al., 2015; van Gelderen et al., 2003). Findings are less consistent regarding working memory (Kieffer & Vukovic, 2013; Geva & Farnia 2012; Marx et al. 2015 vs. Babayiğit, 2015), text reading fluency (Babayiğit, 2015; Geva & Farnina, 2012 vs. Kim, 2012; Marx et al., 2015), and metacognitive knowledge (Mokhtari & Reichard, 2004; van Gelderen et al., 2003 vs. Taki, 2016; van Steensel et al., 2016).

Overall, the field agrees that L2 readers' development of phonological awareness and word reading skills more readily reach the levels comparable to L1 counterparts, but more complex linguistic skills such as vocabulary and listening comprehension continue to lag behind after several years of instruction. This difference in the level of achievement among components of reading leads to the hypothesis that difficulties in L2 reading comprehension after the initial stage are more attributable to language comprehension skills than phonological and word reading skills.

This hypothesis is most directly tested by comparing L1 and L2 reader groups in the strength of contribution of different components to L2 reading comprehension. Although some studies did not find major group differences (Marx et al., 2015; van Steensel et al., 2016; van Gelderen et al., 2003), available evidence seems to be more in support of the hypothesis. For instance, Droop and Verhoeven (2003) found a stronger contribution of vocabulary in L2 than L1 readers; Babayiğit (2014) showed that vocabulary and syntactic skills explained the difference in reading comprehension levels between L1 and L2 readers; Babayiğit (2015) reported a more pronounced contribution of language comprehension skills in L2 readers; Geva & Farnia (2012) showed that, in addition to commonly significant vocabulary, syntactic knowledge and listening comprehension predicted only L2 reading comprehension; finally, Trapman et al. (2014) found that L2 reading comprehension was predicted by vocabulary, grammar, and metacognitive knowledge, but L1 reading comprehension was predicted by word and sentence level fluency. Although different findings encourage more studies, it seems that problems in L2 reading comprehension are more likely "a language problem" (Alderson, 1984), including vocabulary, syntax, and listening comprehension, rather than a reading problem.

## 6. Cross-linguistic transfer in L2 reading

A unique aspect of L2 reading acquisition is the existence of literacy skills acquired through L1. Therefore, it is imperative for theories of L2 reading to explain the relationship between L1 and L2 reading. Because some literacy skills or basic insights into reading operate universally across languages (Koda, 2007), once acquired in L1, L2 readers do not have to learn the same thing twice as long as they can benefit from the transfer of their L1 skills. However, cross-linguistic transfer in reading is complex and affected by various factors. Theoretical and empirical efforts have been made to understand this unique aspect of L2 reading.

Among theories of cross-linguistic transfer in L2 literacy acquisition, one of the most influential models is the interdependence hypothesis and threshold hypothesis proposed by Cummins (e.g., 1979, 1991). Central to this theory is the notion of *common underlying proficiency* that functions both in L1 and L2, which is therefore sharable and transferable between the languages. The threshold is a notion constraining the transfer. It is necessary for children to acquire an adequate level of L2 linguistic proficiency to take advantage of the transfer of L1 academic skills to L2 literacy acquisition. Cummins' theory is developed in the bilingual context where children are developing L1 and L2 simultaneously. Therefore, the notion of threshold applies to both languages, and transfer in both directions are accommodated depending on the context of literacy acquisition.

A similar yet different conceptual framework has been put forward to explain the transfer of reading abilities in the case of L2 learners who possess an adequate basis of L1 literacy skills. This framework is known as the short circuit hypothesis (Clarke, 1980) or by a question "foreign language reading: a reading problem or a language problem?" (Alderson, 1984). The fundamental idea is that L2 learners can transfer their L1 higher-level reading comprehension abilities to their L2 reading when they have reached the threshold level of L2 proficiency. Different from Cummins' theory, the postulated transfer is unidirectional from L1 to L2, and the notion of linguistic threshold applies to only L2 proficiency.

These theoretical frameworks inspired much research in L2 reading (one for bilingual children and another for adult L2 learners), but they are also criticized for the lack of specificity in explaining what is actually transferred (Chung et al., 2019; Koda, 2005). More recently, researchers distinguish language-independent (universal) and language-dependent (specific) aspects in the transfer. This distinction reflects an insight that some skills, but not others, are transferable and the intention to identify exactly what is transferred across languages (e.g., the transfer facilitation model by Koda, 2008).

## 7. Cross-linguistic transfer in common skills

There is a large pool of cross-linguistic correlations (primary evidence of transfer) in the same reading-related skills, which enables meta-analysis. In three informative meta-analyses, the synthesized correlations are: .16 (oral language [oral vocabulary or listening comprehension]), .54 (decoding), and .60 (phonological awareness) in Melby-Lervåg and Lervåg (2011); .10 (vocabulary), .44 (decoding), .42 (phonological awareness), and .37 (morphological awareness) in Yang, Cooc, and Sheng (2017); .47 (reading comprehension) in Jeon and Yamashita (2014).

Results from the first two meta-analyses are similar: within the SVR framework, correlations in the oral language domain is much lower than those in the decoding domain. This convergence is noteworthy because the L1-L2 pairing was predominantly alphabetic in Melby-Lervåg and Lervåg (2011) but exclusively Chinese and English in Yang et al. (2017). The commonality bears theoretical significance; namely, language distance has little influence on what is more likely to be transferred (word level skills) than others (oral language skills) between languages. Melby-Lervåg and Lervåg (2011) attributed the lower correlation of oral language to its complexity. Indeed, the construct of oral language is quite complex and represented by a diverse range of processes, which leads to difficulties in consistent measurement as well; on the other hand, the construct of decoding is much simpler and easy to quantify (Proctor et al., 2006). Relatively high correlations in the domain of metalinguistic awareness support the contention that metalinguistic awareness is a cross-linguistically sharable property because it represents fundamental, abstract insights that are functional to reading acquisition regardless of the surface structure of languages (Koda, 2008).

Results from moderator analyses bear theoretical importance as well. The correlation of decoding was stronger when bilinguals were taught in L1 and L2 than in L2 only and when L1 and L2 were both alphabetic than ideographic/alphabetic pairs (Melby-Lervåg & Lervåg, 2011); the correlation of reading comprehension was stronger when L1 and L2 were both Indo-European languages than when they were non-Indo-European/Indo-European languages (Jeon & Yamashita, 2014). The first result implies the effects of language proficiency and the last two are pertinent to the language distance.

Reflecting the increasing body of research, evidence of cross-linguistic transfer in oral reading fluency is emerging. Significant cross-language associations were reported in primary school Spanish-English bilinguals in word reading fluency (Pasquarella et al., 2014), word/text reading fluency (Baker, Park, & Baker, 2012), and text reading fluency (Dominguez de Ramirez & Shapiro, 2007). Pasquarella et al. (2014), who also examined Chinese-English bilinguals, provided some further

insights: the transfer of oral reading fluency may be bidirectional and may not be bound by the script distance.

Taki (2016) compared metacognitive, online-reading strategy use by college students cross-linguistically across three conditions (reading in L1-English by Canadians, L2-English by Iranians, and L1-Farsi by the same Iranians). Overall, reported strategy use was much more similar between L1 and L2 in the Iranian readers than between L1-English by the Canadian and L2-English by the Iranian readers. This result suggests the transfer in perceived metacognitive strategy use.

## **8. Effects of L1 components on L2 reading comprehension**

Cross-linguistic transfer applies beyond corresponding skills. We can also hypothesize causal relationships of transferred competencies to the outcome measure; for instance, L1 decoding may contribute to L2 reading comprehension. The investigation into crossover effects (how L1 components explain or predict L2 reading) becomes much more complex. First, various causal paths can be hypothesized (direct and indirect). Second, decisions must be made about control variables. A critical issue is whether to include L2 components. Since many L1 and L2 components correlate, it is likely that when no L2 skills are present in the analysis, L1 skills are more likely to appear as significant. Research into crossover effects is growing, although consensus has yet to be reached in many areas of investigation. Similar to reading research in general, research in this field is more advanced in phonological skills where outcome measures are mostly L2 word level skills (decoding, word recognition). Although they are valuable, we focus on reading comprehension as the focal outcome construct.

Before discussing crossover effects, it is worth revisiting Melby-Lervåg and Lervåg (2011), because their results numerically illustrate some discussions below. The researchers synthesized correlations between L2 reading comprehension and two L1 and two L2 components: .24 (L1 decoding), .04 (L1 oral language, this was insignificant), .54 (L2 decoding), and .46 (L2 oral language). From these findings we could infer that L1 components may be weaker predictors than L2 components. Indeed, this is one of the trends emerging from recent studies, though there are controversial findings as well. Below, we will look at crossover effects in different skill domains.

## 8.1 Phonological skills

Due to a relatively large amount of research in phonological processing, a consensus has been reached on the crossover effect of phonological skills regardless of the typological distance of languages (Genesee & Geva, 2006). However, evidence of the contribution of phonological skills to outcomes beyond word-level skills is limited (Chung et al., 2019). In this sense, Manis et al., (2004), Kim (2012), and Kahn-Horwits and Saba (2018) are informative.

In none of these studies did L1 phonological skills have direct impact on L2 reading comprehension, but their effects were mediated by other L1 or L2 skills. Manis et al. (2004) examined contributions of parallel predictors of L1-Spanish and L2-English, one of which was phonological awareness. Without English predictors, most Spanish predictors were significant. However, once English predictors were in the analysis, all Spanish predictors became insignificant. Based on the common variance between English and Spanish components, the authors argued that the English predictors “nearly completely mediated the relationship of Spanish variables to English passage comprehension” (p. 222). Kim (2012) also examined Spanish-English bilinguals. Predictors were L1-Spanish phonological skills and three L2-English components (oral language, word reading, reading fluency). Only English predictors were significant. Since the Spanish phonological skills moderately correlated with English variables, Kim also argued that L1 phonological skills were “completely mediated by L2 oral language and literacy skills” (p. 698). Kahn-Horwits and Saba (2018) did not include L2 predictors but analyzed contributions of L1-Arabic literacy skills (phonological awareness, orthographic knowledge, and morphological skills) to word reading and reading comprehension in L2-English. A significant contribution of L1 phonological awareness was found only for English word reading (mediated by morphological awareness). Collectively, it seems that L1 phonological skills do not have direct crossover effects on L2 reading comprehension, and the impacts are more likely mediated. However, identifying indirect effects is valuable as it reveals the complex pathway that L1 phonological skills contribute to L2 reading comprehension.

## 8.2 Orthographic skills

Since orthographic skills are built on orthographic representation, it would be reasonable to expect little transfer across different writing systems due to the lack of overlapping orthographic symbols and units. Indeed, Chung et al., (2019) maintained that past bilingual studies converged on the conclusion that orthographic processing is language specific; cross-linguistic transfer of orthographic skills have

been identified only in language pairs that share the same Roman alphabet. Thus, orthographic processing, among others, seems susceptible to differences in L1 and L2 scripts. In support of this, crossover effects of orthographic skills were not observed even in word level skills either in Spanish-English (Sun-Alperin & Wang, 2011) or Chinese-English (Wang et al., 2009b) pairings. Aforementioned study of Kahn-Horwits and Saba (2018), on the other hand, found the crossover effect of L1-Arabic orthographic skills on L2-English word reading (without controlling for L2 orthographic skills). However, orthographic processing was not significant for L2 reading comprehension, indicating the less direct contribution of orthographic skills to reading comprehension than to word reading.

### 8.3 Morphological skills

Morphological skills may more strongly affect reading comprehension compared to phonological or orthographic skills because a morpheme is a unit of meaning which embodies various linguistic properties (phonology, orthography, meaning, and syntax). Indeed, among a range of linguistic skills, morphological awareness has much more consistently shown crossover effects on L2 reading comprehension. A growing body of research has been examining morphological skills' transfer using language pairs of distant scripts. This design is methodologically more rigorous as it controls for the potentially spurious influence of orthographic overlaps.

For instance, the following studies using L1-Chinese/L2-English pairing all supported transfer in a wide range of age groups in different contexts. Pasquarella, Chen, Lam, Luo, and Ramirez (2011) measured compound awareness in Chinese and English (a common feature) and English derivational awareness (largely English specific). English compound awareness predicted both English and Chinese reading comprehension, but English derivational awareness predicted only English reading comprehension. Zhang, Koda, and Sun (2014) involved compound awareness in both languages and Chinese radical awareness (Chinese specific). Chinese and English compound awareness predicted English reading comprehension, but only Chinese compound awareness predicted Chinese reading comprehension. Xue and Jiang (2017) included compound awareness in both languages, English derivational awareness, and Chinese homographic morpheme identification (Chinese specific). Chinese compound awareness accounted for English reading comprehension, but only Chinese homographic identification explained Chinese reading comprehension. To summarize, these studies similarly showed that transfer occurs in cross-linguistically common morphological features. However, the studies are different in the directionality of transfer: L1 to L2 (Xue & Jiang, 2017; Zhang et al., 2014, EFL contexts) or L2 to L1 (Pasquarella et al., 2011, a bilingual context). Apart



from many methodological disparities, this difference may be pertinent to readers' language proficiency. Chinese EFL learners had stronger literacy skills in L1 than L2. Contrarily, Chinese children in the bilingual context were likely either more balanced in the two languages or even stronger in L2 than in L1.

## 8.4 Reading fluency

Based on the implication that reading fluency is a transferable language-general capacity (Pasquarella et al., 2014), together with the recognition of increasing interest in this skill, despite the scarcity of empirical studies, reading fluency is briefly discussed in this section. Baker et al. (2012) examined bidirectional cross-over effects of oral reading fluency on reading comprehension in Spanish-English bilinguals; only within-language predictors were significant. However, this does not immediately deny crossover effects. Due to high correlations between Spanish and English reading fluency, the cross-linguistic transfer may have been mediated by within-language fluency.

## 8.5 Linguistic comprehension skills

This section deals with crossover effects of vocabulary, syntactic knowledge, listening comprehension, and reading comprehension in L1 on L2 reading comprehension.

As discussed above, studies based on the SVR operationalize linguistic comprehension skills through listening comprehension, vocabulary, and syntactic knowledge, and treat them in amalgamation under the umbrella notion of oral language. Very little crossover effects of L1 oral language was found (Geva & Geese, 2006). The following studies of Spanish-English bilinguals in elementary school years are illustrative. Gottardo and Mueller (2009) examined how L1 oral language and phonological awareness contribute to L2-English reading comprehension together with parallel L2 components and L2 word reading. No L1 components made significant contributions either directly or indirectly. Lesaux, Crosson, Kieffer, and Pierce (2010) tested an L2 model (predicting L2 reading comprehension by L2 oral language and L2 word reading) and a cross-language model (predicting L2 reading comprehension by parallel L1 skills). In the L2 model, L2 oral language was significant, but in the cross-language model, neither L1 oral language nor L1 word reading was significant. Thus, neither study supported contributions of L1 oral language even across typologically closer languages.

Studies in each component in the linguistic comprehension domain are also available. Among them, the effect of L1 vocabulary is inconclusive. In Manis et al. (2004) and Lindsey, Manis, and Bailey (2003), both of which examined Spanish-English



bilingual children, L1-Spanish vocabulary was not a significant predictor of English reading comprehension after controlling for English and Spanish predictors. Similar results were reported with English-French bilingual children (Jared, Levy, Cormier, & Wade-Woolley, 2011). On the other hand, Proctor et al. (2006), whose participants were Spanish-English children, identified a unique contribution of L1-Spanish vocabulary, albeit only minimally, controlling for English predictors and language of instruction. Li, McBride-Chang, Wong, and Shu (2012) examined Chinese-English bilinguals by predicting English reading comprehension by a set of Chinese and English predictors. In addition to English predictors, Chinese vocabulary was a consistently significant predictor.

There is a dearth of studies on the transfer effect of L1 syntactic processing. Therefore, although recent studies support its contributions, more research is needed. In Jared et al. (2011), L1-English grammar was the strongest predictor of L2-French reading comprehension in English-French bilingual children. Siu and Ho (2015) examined Chinese children studying English in grades 1 and 3. Two aspects of their Chinese and English syntactic skills were measured: word order (a common feature in the languages) and morphosyntax (a distant feature). Overall, results indicated the contribution of L1-Chinese syntactic skills to L2-English reading comprehension, and the effect appeared more strongly and earlier in grade 1 in the common feature than the distant feature. Together, these studies demonstrated crossover effects of L1 syntactic skills, and Siu and Ho (2015) also supported the language distance effect.

Research into the contribution of L1 listening comprehension is also limited, and the results are controversial. In Proctor et al. (2006), L1-Spanish listening comprehension was not a significant predictor among Spanish-English bilingual children. In contrast, Edele and Stanat (2016) showed significant contributions of L1 listening comprehension to L2-German reading comprehension. The students' L1s were Russian (linguistically closer to German) and Turkish (linguistically more distant from German). Commonly in both groups, the contribution of L1 listening comprehension was not only significant but also, surprisingly, even stronger than L2 vocabulary and reading fluency. Based on Cummins' linguistic interdependence hypothesis, the authors tested whether higher L1 listening comprehension would affect the degree of the cross-linguistic relationship. This L1 proficiency effect was observed only in the Turkish group, which was interpreted as the support of the language distance effect. Namely, the transfer occurs regardless of L1 proficiency between closer languages, whereas the transfer is affected by the level of L1 proficiency between distant languages.

Transfer of L1 reading comprehension is only relevant to L2 readers who possess adequate levels of L1 literacy skills. Despite this limitation, there is a substantial body of research leading to the consensus supporting the transfer. To illustrate,

van Gelderen et al., (2004, 2007) demonstrated strong cross-linguistic contribution from L1-Dutch to L2-English reading comprehension. Li et al. (2012) showed bidirectional transfer by demonstrating a strong contribution of L1-Chinese to L2-English reading comprehension, and vice versa. Yamashita and Shiotsu (2017) identified a significant contribution of L1-Japanese to L2-English academic reading comprehension. There are at least two widely known factors that affect the transfer of L1 reading comprehension ability: language distance (Jeon & Yamashita, 2014) and L2 proficiency (Yamashita & Shiotsu, 2017).

In summary, researching into crossover effects of linguistic comprehension skills is especially challenging due to the complexity of constructs and measurement as well as the decision to minimize spurious third factor influences. Except for solid positive evidence in the transfer in the domain of L1 reading comprehension, available empirical findings are either too limited or inconclusive to make a reliable generalization. We need more research before reaching firm conclusions.

## 8.6 Factors affecting transfer

One of the issues that make cross-linguistic transfer a complex phenomenon is a myriad of factors affecting transfer on its own or in combination. A full discussion is beyond the scope of this chapter. Potentially intervening factors are listed below for future examination: L1-L2 distance (writing system and typological structure), L1/L2 linguistic complexity (e.g., transparency of sound-symbol correspondences), L1/L2 proficiency, L1/L2 input and use, educational elements (e.g., language of instruction, teaching method, reading program), motivation and attitude, and social factors (e.g., social status of L1/L2, socioeconomic status, language policy).

## 9. Conclusion

In this chapter, we first discussed linguistic and cognitive processes underlying reading comprehension, then synthesized current research insights into differences between L1 and L2 reading and cross-linguistic transfer. We close this chapter with brief comments on theories and methodologies in L2 reading research.

Reading research has been guided by flexible theoretical frameworks more than by the testing of precise theories (Perfetti & Stafura, 2014). This is partly because “reading has too many components for a single theory” (Perfetti & Stafura, 2014, p. 22) and also because research is often primarily motivated for applied, most typically educational, purposes. Although not as precise as narrowly focused theories (e.g., theories of word recognition), broad frameworks have the value as a heuristic tool by flexibly accommodating additions and modifications to the original

framework and thus are broadly applicable. We have witnessed this advantage in the popularity of the SVR, which is one such global framework (Kirby & Savage, 2008). As a reflection of complexity of reading and the flexibility in the use of theories, researchers occasionally conclude that multiple theories are equally useful, or seemingly contradictory theories turn out to be complementary, in explaining their research findings (Bernhardt & Kamil, 1995; Geva & Siegel, 2000; Kuo, Uchikoshi, Kim & Yang, 2016). To understand reading research, we need to be equipped with different sets of theories or theoretical frameworks and stay flexible in finding the most efficient explanations for the phenomena in focus.

To reiterate, research findings are often inconclusive in many areas. This is partly because of a broad range of factors that influence research outcomes such as the ones mentioned above. However, research methodology itself is also a source of disparity (Chung et al., 2019). Indeed, the choice of reading comprehension tests can produce different results (Keenan, Betjemann, & Olson, 2008) as is the case with virtually all components. The research design, whether concurrent, cross-sectional, or longitudinal, is also influential. Another important source of disparity is, as already discussed, the choice of control variables. In the research into crossover effects, potential control variables logically double because of the consideration of predictors in dual languages. This simple fact increases variability in researchers' choices. Reaching conclusions from methodologically diverse studies is difficult. One of the promising avenues seems the accumulation of studies with reasonable rigor, which enables meta-analyses that go back to the basics, such as correlations, and powerfully synthesize researchers' endeavors in the course of research advancement.

## References

- Alderson, J. C. (1984). Reading in a foreign language: A reading problem or a language problem. In J. C. Alderson & A. H. Urquhart (Eds.), *Reading in a foreign language*. Longman.
- Anglin, J. M. (1993). Vocabulary development: A morphological analysis. *Monographs of the Society for Research in Child Development*, 58(10)[238], v–165. <https://doi.org/10.2307/1166112>
- Babayigit, S. (2014). The role of oral language skills in reading and listening comprehension of text: A comparison of monolingual (L1) and bilingual (L2) speakers of English language. *Journal of Research in Reading*, 37(S1), S22–S47. <https://doi.org/10.1111/j.1467-9817.2012.01538.x>
- Babayigit, S. (2015). The relations between word reading, oral language, and reading comprehension in children who speak English as a first (L1) and second language (L2): A multigroup structural analysis. *Reading and Writing*, 28(4), 527–544. <https://doi.org/10.1007/s11145-014-9536-x>
- Baker, L., & Beall, L. C. (2009). Metacognitive processes and reading comprehension. In S. E. Israel & D. D. Duffy (Eds.), *Handbook of research on reading comprehension* (pp. 373–388). Routledge.

- Baker, D. L., Park, Y., & Baker, S. K. (2012). The reading performance of English learners in grades 1–3: The role of initial status and growth on reading fluency in Spanish and English. *Reading and Writing*, 25(1), 251–281. <https://doi.org/10.1007/s11145-010-9261-z>
- Bar-Kochva, I. (2013). What are the underlying skills of silent reading acquisition? A developmental study from kindergarten to the 2nd grade. *Reading and Writing*, 26(9), 1417–1436. <https://doi.org/10.1007/s11145-012-9414-3>
- Bernhardt, E. B., & Kamil, M. L. (1995). Interpreting relationships between L1 and L2 reading: Consolidating the linguistic threshold and the linguistic interdependence hypotheses. *Applied Linguistics*, 16(1), 15–31. <https://doi.org/10.1093/applin/16.1.15>
- Binder, K. S., Cote, N. G., Lee, C., Bessette, E., & Vu, H. (2017). Beyond breadth: The contributions of vocabulary depth to reading comprehension among skilled readers. *Journal of Research in Reading*, 40(3), 333–343. <https://doi.org/10.1111/1467-9817.12069>
- Carr, T., & Levy, B. (1990). *Reading and its development*. Academic Press.
- Chung, S. C., Chen, X., & Geva, E. (2019). Deconstructing and reconstructing cross-language transfer in bilingual reading development: An interactive framework. *Journal of Neurolinguistics*, 50, 149–161. <https://doi.org/10.1016/j.jneuroling.2018.01.003>
- Cirino, P. T., Miciak, J., Ahmed, Y., Barnes, M. A., Taylor, W. P., & Gerst, E. H. (2019). Executive function: Association with multiple reading skills. *Reading and Writing*, 32(7), 1819–1846. <https://doi.org/10.1007/s11145-018-9923-9>
- Clarke, M. A. (1980). The short circuit hypothesis of ESL reading – or when language competence interferes with reading performance. *The Modern Language Journal*, 64(2), 203–209. <https://doi.org/10.1111/j.1540-4781.1980.tb05186.x>
- Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review*, 100(4), 589–608. <https://doi.org/10.1037/0033-295X.100.4.589>
- Conrad, N. J., Harris, N., & Williams, J. (2013). Individual differences in children’s literacy development: The contribution of orthographic knowledge. *Reading and Writing*, 26(8), 1223–1239. <https://doi.org/10.1007/s11145-012-9415-2>
- Cook, V., & Bassetti, B. (2005). An introduction to researching second language writing systems. In V. Cook & B. Bassetti (Eds.), *Second language writing systems* (pp. 1–67). Multilingual Matters. <https://doi.org/10.21832/9781853597954-003>
- Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. *Review of Educational Research*, 49(2), 222–251. <https://doi.org/10.3102/00346543049002222>
- Cummins, J. (1991). Interdependence of first- and second-language proficiency in bilingual children. In E. Byalystok (Ed.), *Language processing in bilingual children* (pp. 70–89). Cambridge University Press. <https://doi.org/10.1017/CBO9780511620652.006>
- Dominguez De Ramírez, R., & Shapiro, E. S. (2007). Cross-language relationship between Spanish and English oral reading fluency among Spanish-speaking English language learners in bilingual education classrooms. *Psychology in the Schools*, 44(8), 795–806. <https://doi.org/10.1002/pits.20266>
- Droop, M., & Verhoeven, L. (2003). Language proficiency and reading ability in first- and second-language learners. *Reading Research Quarterly*, 38(1), 78–103. <https://doi.org/10.1598/RRQ.38.1.4>
- Edele, A., & Stanat, P. (2016). The role of first-language listening comprehension in second-language reading comprehension. *Journal of Educational Psychology*, 108(2), 163–180. <https://doi.org/10.1037/edu0000060>

- Ehri, I. C. (2007). Development of sight word reading: Phases and findings. In M. J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 135–154). Blackwell.
- Flavell, J. H. (1979). Metacognition and cognitive monitoring. *American Psychologist*, 34(10), 906–911. <https://doi.org/10.1037/0003-066X.34.10.906>
- Genesee, F., & Geva, E. (2006). Cross-linguistic relationships in working memory, phonological processes, and oral language. In D. August & T. Shanahan (Eds.), *Developing literacy in second-language learners: Report of the National Literacy Panel on language-minority children and youth* (pp. 175–183). Lawrence Erlbaum Associates.
- Geva, E., & Farnia, F. (2012). Developmental changes in the nature of language proficiency and reading fluency paint a more complex view of reading comprehension in ELL and EL1. *Reading and Writing*, 25(8), 1819–1845. <https://doi.org/10.1007/s11145-011-9333-8>
- Geva, E., & Genesee, F. (2006). First-language oral proficiency and second language literacy. In D. August & T. Shanahan (Eds.), *Developing literacy in second-language learners: Report of the National Literacy Panel on language-minority children and youth* (pp. 185–195). Lawrence Erlbaum Associates.
- Geva, E., & Siegel, L. S. (2000). Orthographic and cognitive factors in the concurrent development of basic reading skills in two languages. *Reading and Writing: An Interdisciplinary Journal*, 12(1/2), 1–30. <https://doi.org/10.1023/A:1008017710115>
- Goff, D. A., Pratt, C., & Ong, B. (2005). The relations between children's reading comprehension, working memory, language skills and components of reading decoding in a normal sample. *Reading and Writing: An Interdisciplinary Journal*, 18(7–9), 583–616. <https://doi.org/10.1007/s11145-004-7109-0>
- Gottardo, A., & Mueller, J. (2009). Are first- and second-language factors related in predicting second-language reading comprehension? A study of Spanish-speaking children acquiring English as a second language from first to second grade. *Journal of Educational Psychology*, 101(2), 330–344. <https://doi.org/10.1037/a0014320>
- Gough, P. B., Hoover, W. A., & Peterson, C. L. (1996). Some observations on a simple view of reading. In C. Cornoldi. & J. Oakhill (Eds.), *Reading comprehension difficulties: Processes and intervention* (pp. 1–13). Lawrence Erlbaum Associates.
- Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education*, 7(1), 6–10. <https://doi.org/10.1177/074193258600700104>
- Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. Cambridge University Press.
- Grabe, W., & Stoller, F. L. (2020). *Teaching and researching reading* (3rd ed.). Routledge.
- Hagiliassiss, N., Pratt, C., & Johnston, M. (2006). Orthographic and phonological processes in reading. *Reading and Writing*, 19(3), 235–263. <https://doi.org/10.1007/s11145-005-4123-9>
- Hoover, W., & Gough, P. (1990). The simple view of reading. *Reading and Writing*, 2(2), 127–160. <https://doi.org/10.1007/BF00401799>
- Hudson, T. (2007). *Teaching second language reading*. Oxford University Press.
- Jared, D., Cormier, P., Levy, B. A., & Wade-Woolley, L. (2011). Early predictors of biliteracy development in children in French immersion: A 4-year longitudinal study. *Journal of Educational Psychology*, 103(1), 119–139. <https://doi.org/10.1037/a0021284>
- Jeon, E. H. (2011). Contribution of morphological awareness to L2 reading comprehension. *The Modern Language Journal*, 95(2), 217–235. <https://doi.org/10.1111/j.1540-4781.2011.01179.x>
- Jeon, E. H., & Yamashita, J. (2014). L2 reading comprehension and its correlates: A meta-analysis. *Language Learning*, 64(1), 160–212. <https://doi.org/10.1111/lang.12034>

- Joshi, R. M., & Aaron, P. G. (2000). The component model of reading: Simple view of reading made a little more complex. *Reading Psychology*, 21(2), 85–97.  
<https://doi.org/10.1080/02702710050084428>
- Kahn-Horwitz, J., & Saba, M. (2018). Weak English foreign language readers: The cross-linguistic impact of morphological awareness. *Reading and Writing*, 31(8), 1843–1868.  
<https://doi.org/10.1007/s11145-017-9810-9>
- Keenan, J. M., Betjemann, R. S., & Olson, R. K. (2008). Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading*, 12(3), 281–300. <https://doi.org/10.1080/10888430802132279>
- Kieffer, J. M., & Lesaux, N. K. (2012). Knowledge of words, knowledge about words: Dimensions of vocabulary in first and second language learners in sixth grade. *Reading and Writing*, 25(2), 347–373. <https://doi.org/10.1007/s11145-010-9272-9>
- Kieffer, M. J., & Vukovic, R. K. (2013). Growth in reading-related skills of language minority learners and their classmates: More evidence for early identification and intervention. *Reading and Writing*, 26(7), 1159–1194. <https://doi.org/10.1007/s11145-012-9410-7>
- Kim, Y.-S. (2012). The relations among L1 (Spanish) literacy skills, L2 (English) language, L2 text reading fluency, and L2 reading comprehension for Spanish-speaking ELL first grade students. *Learning and Individual Differences*, 22(6), 690–700.  
<https://doi.org/10.1016/j.lindif.2012.06.009>
- Kim, Y. S., Wagner, R. K., & Foster, E. (2011). Relations among oral reading fluency, silent reading fluency, and reading comprehension: A latent variable study of first-grade readers. *Scientific Studies of Reading*, 15(4), 338–362. <https://doi.org/10.1080/10888438.2010.493964>
- Kirby, J. R., & Savage, R. S. (2008). Can the simple view deal with the complexities of reading? *Literacy*, 42(2), 75–82. <https://doi.org/10.1111/j.1741-4369.2008.00487.x>
- Koda, K. (2005). *Insights into second language reading*. Cambridge University Press.  
<https://doi.org/10.1017/CBO9781139524841>
- Koda, K. (2007). Reading and language learning: Crosslinguistic constraints on second language reading development. *Language Learning*, 57(Supplement1), 1–44.  
<https://doi.org/10.1111/0023-8333.101997010-i1>
- Koda, K. (2008). Impacts of prior literacy experience on second language learning to read. In K. Koda & A. Zehler (Eds.), *Learning to read across languages: Cross-linguistic relationships in first- and second-language literacy development* (pp. 68–96). Routledge.  
<https://doi.org/10.4324/9780203935668-11>
- Koda, K. (2019). Integrated communication skills approach: Reading to learn as a basis for language and content integration. In K. Koda & J. Yamashita (Eds.), *Reading to learn in a foreign language: An integrated approach to foreign language instruction and assessment* (pp. 30–54). Routledge.
- Kuo, L., & Anderson, R. C. (2006). Morphological awareness and learning to read: A cross-language perspective. *Educational Psychologist*, 41(3), 161–180.  
[https://doi.org/10.1207/s15326985ep4103\\_3](https://doi.org/10.1207/s15326985ep4103_3)
- Kuo, L. J., Uchikoshi, Y., Kim, T. J., & Yang, X. (2016). Bilingualism and phonological awareness: Re-examining theories of cross-language transfer and structural sensitivity. *Contemporary Educational Psychology*, 46, 1–9. <https://doi.org/10.1016/j.cedpsych.2016.03.002>
- LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, 6(2), 293–323. [https://doi.org/10.1016/0010-0285\(74\)90015-2](https://doi.org/10.1016/0010-0285(74)90015-2)



- Leider, C. M., Proctor, C. P., Silverman, R. D., & Harring, J. R. (2013). Examining the role of vocabulary depth, cross-linguistic transfer, and types of reading measures on the reading comprehension of Latino bilinguals in elementary school. *Reading and Writing*, 26(9), 1459–1485. <https://doi.org/10.1007/s11145-013-9427-6>
- Lesaux, N. K., Crosson, A. C., Kieffer, M. J., & Pierce, J. (2010). Uneven profiles: Language minority learners' word reading, vocabulary, and reading comprehension skills. *Journal of Applied Developmental psychology*, 31(6), 475–483. <https://doi.org/10.1016/j.appdev.2010.09.004>
- Lesaux, N., Koda, K., Siegel, L., & Shanahan, T. (2006). Development of literacy. In D. August & T. Shanahan (Eds.), *Developing literacy in second-language learners* (pp.75–122). Lawrence Erlbaum Associates.
- Li, T., McBride-Chang, C., Wong, A., & Shu, H. (2012). Longitudinal predictors of spelling and reading comprehension in Chinese as an L1 and English as an L2 in Hong Kong Chinese children. *Journal of Educational Psychology*, 104(2), 286–301. <https://doi.org/10.1037/a0026445>
- Lindsey, K. A., Manis, F. R., & Bailey, C. E. (2003). Prediction of first-grade reading in Spanish-speaking English-language learners. *Journal of Educational Psychology*, 95(3), 482–494. <https://doi.org/10.1037/0022-0663.95.3.482>
- Manis, F. R., Lindsey, K. A., & Bailey, C. E. (2004). Development of reading in grades K-2 in Spanish-speaking English-language learners. *Learning Disabilities Research and Practice*, 19(4), 214–224. <https://doi.org/10.1111/j.1540-5826.2004.00107.x>
- Marx, A., Stanat, P., Roick, T., Segerer, R., Marx, P., & Schneider, W. (2015). Components of reading comprehension in adolescent first-language and second-language students from low-track schools. *Reading and Writing*, 28(6), 891–914. <https://doi.org/10.1007/s11145-015-9554-3>
- Melby-Lervåg, M., & Lervåg, A. (2011). Cross-linguistic transfer of oral language, decoding, phonological awareness and reading comprehension: A meta-analysis of the correlational evidence. *Journal of Research in Reading*, 34(1), 114–135. <https://doi.org/10.1111/j.1467-9817.2010.01477.x>
- Melby-Lervåg, M., & Lervåg, A. (2014). Reading comprehension and its underlying components in second-language learners: A meta-analysis of studies comparing first- and second-language learners. *Psychological Bulletin*, 140(2), 409–433. <https://doi.org/10.1037/a0033890>
- Mokhtari, K., & Reichard, C. (2004). Investigating the strategic reading processes of first and second language readers in two different cultural contexts. *System*, 32(3), 379–394. <https://doi.org/10.1016/j.system.2004.04.005>
- Nagy, B., & Anderson, R. C. (1984). How many words are there in printed school English? *Reading Research Quarterly*, 19(3), 304–330. <https://doi.org/10.2307/747823>
- Nassaji, H. (2014). The role and importance of lower-level processes in second language reading. *Language Teaching*, 47(01), 1–37. <https://doi.org/10.1017/S0261444813000396>
- Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press. <https://doi.org/10.1017/CBO9781139858656>
- National Reading Panel. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* (No.00-4769). National Institute of Child Health and Human Development.
- Pasquarella, A., Chen, X., Gottardo, A., & Geva, E. (2014). Cross-language transfer of word reading accuracy and word reading fluency in Spanish-English and Chinese-English bilinguals: Script-universal and script-specific processes. *Journal of Educational Psychology*, 107(1), 96–110. <https://doi.org/10.1037/a0036966>

- Pasquarella, A., Chen, X., Lam, K., Luo, Y. C., & Ramirez, G. (2011). Cross-language transfer of morphological awareness in Chinese-English bilinguals. *Journal of Research in Reading*, 34(1), 23–42. <https://doi.org/10.1111/j.1467-9817.2010.01484.x>
- Perfetti, C. A. (1985). *Reading ability*. Oxford University Press.
- Perfetti, C. A. (2003). The universal grammar of reading. *Scientific Studies of Reading*, 7(1), 3–24. [https://doi.org/10.1207/S1532799XSSR0701\\_02](https://doi.org/10.1207/S1532799XSSR0701_02)
- Perfetti, C. A., & Hart, L. (2002). The lexical quality hypothesis. In L. Verhoeven, C. Elbro, & P. Reitsma (Eds.), *Precursors of functional literacy* (pp. 189–213). John Benjamins. <https://doi.org/10.1075/swll.11.14per>
- Perfetti, C. A., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading*, 18(1), 22–37. <https://doi.org/10.1080/10888438.2013.827687>
- Proctor, C. P., August, D., Carlo, M. S., & Snow, C. (2006). The intriguing role of Spanish language vocabulary knowledge in predicting English reading comprehension. *Journal of Educational Psychology*, 98(1), 159–169. <https://doi.org/10.1037/0022-0663.98.1.159>
- Proctor, C. P., Silverman, R. D., Harring, J. R., & Montecillo, C. (2012). The role of vocabulary depth in predicting reading comprehension among English monolingual and Spanish–English bilingual children in elementary school. *Reading and Writing*, 25(7), 1635–1664. <https://doi.org/10.1007/s11145-011-9336-5>
- Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning*, 52(3), 513–536. <https://doi.org/10.1111/1467-9922.00193>
- Qian, D. D., & Schedl, M. (2004). Evaluation of an in-depth vocabulary knowledge measure for assessing reading performance. *Language Testing*, 21(1), 28–52. <https://doi.org/10.1191/0265532204lt2730a>
- Schmitt, M. C. (2005). Measuring students' awareness and control of strategic processes. In S. E. Israel, C. C. Block, K. L. Bauserman, & K. Kinnucan-Welsch (Eds.), *Metacognition in literacy learning* (pp. 101–119). Lawrence Erlbaum Associates.
- Shiotsu, T. (2010). *Components of L2 reading: Linguistic and processing factors in the reading test performances of Japanese EFL learners*. Cambridge University Press.
- Siu, C. T.-S., & Ho, C. S.-H. (2015). Cross-language transfer of syntactic skills and reading comprehension among young Cantonese-English bilingual students. *Reading Research Quarterly*, 50(3), 313–336. <https://doi.org/10.1002/rq.101>
- Sun-Alperin, M. K., & Wang, M. (2011). Cross-language transfer of phonological and orthographic processing skills from Spanish L1 to English L2. *Reading and Writing*, 24(5), 591–614. <https://doi.org/10.1007/s11145-009-9221-7>
- Taki, S. (2016). Metacognitive online reading strategy use: Readers' perceptions in L1 and L2. *Journal of Research in Reading*, 39(4), 409–427. <https://doi.org/10.1111/1467-9817.12048>
- Trapman, M., van Gelderen, A., van Steensel, R., van Schooten, E., & Hulstijn, J. (2014). Linguistic knowledge, fluency and meta-cognitive knowledge as components of reading comprehension in adolescent low achievers: differences between monolinguals and bilinguals. *Journal of Research in Reading*, 37(S1), S3–S21. <https://doi.org/10.1111/j.1467-9817.2012.01539.x>
- van Gelderen, A., Schoonen, R., de Glopper, K., Hulstijn, J., Simis, A., Snellings, P., & Stevenson, M. (2004). Linguistic knowledge, processing speed, and metacognitive knowledge in first- and second-language reading comprehension: A componential analysis. *Journal of Educational Psychology*, 96(1), 19–30. <https://doi.org/10.1037/0022-0663.96.1.19>



- van Gelderen, A., Schoonen, R., de Glopper, K., Hulstijn, J., Snellings, P., Simis, A., & Stevenson, M. (2003). Roles of linguistic knowledge, metacognitive knowledge and processing speed in L3, L2, and L1 reading comprehension: A structural equation modeling approach. *International Journal of Bilingualism*, 7(1), 7–25. <https://doi.org/10.1177/13670069030070010201>
- van Gelderen, A., Schoonen, R., Stoel, R. D., De Glopper, K., & Hulstijn, J. H. (2007). Development of adolescent reading comprehension in language 1 and language 2: A longitudinal analysis of constituent components. *Journal of Educational Psychology*, 99(3), 477–491. <https://doi.org/10.1037/0022-0663.99.3.477>
- van Steensel, R., Oostdam, R., van Gelderen, A., & van Schooten, E. (2016). The role of word decoding, vocabulary knowledge and meta-cognitive knowledge in monolingual and bilingual low-achieving adolescents' reading comprehension. *Journal of Research in Reading*, 39(3), 312–329. <https://doi.org/10.1111/1467-9817.12042>
- Wang, M., Yang, C., & Cheng, C. (2009b). The contributions of phonology, orthography, and morphology in Chinese–English biliteracy acquisition. *Applied Psycholinguistics*, 30(2), 291–314. <https://doi.org/10.1017/S0142716409090122>
- White, T. G., Power, M. A., & White, S. (1989). Morphological analysis: Implications for teaching and understanding vocabulary growth. *Reading Research Quarterly*, 24(3), 283–304. <https://doi.org/10.2307/747771>
- Xue, J., & Jiang, X. (2017). The developmental relationship between bilingual morphological awareness and reading for Chinese EFL adult learners: a longitudinal study. *Reading and Writing*, 30(2), 417–438. <https://doi.org/10.1007/s11145-016-9683-3>
- Yamashita, J., & Shiotsu, T. (2017). Comprehension and knowledge components that predict L2 reading: A latent-trait approach. *Applied Linguistics*, 38(1), 43–67. <https://doi.org/10.1093/applin/amu079>
- Yang, M., Cooc, N., & Sheng, L. (2017). An investigation of cross-linguistic transfer between Chinese and English: A meta-analysis. *Asian-Pacific Journal of Second and Foreign Language Education*, 2(1). <https://doi.org/10.1186/s40862-017-0036-9>
- Zhang, D. (2017). Derivational morphology in reading comprehension of Chinese-speaking learners of English: A longitudinal structural equation modeling study. *Applied Linguistics*, 38(6), 871–895. <https://doi.org/10.1093/applin/amv072>
- Zhang, D., Koda, K., & Sun, X. (2014). Morphological awareness in biliteracy acquisition: A study of young Chinese EFL readers. *International Journal of Bilingualism*, 18(6), 570–585. <https://doi.org/10.1177/1367006912450953>

## L2 reading comprehension and its correlates

### An updated meta-analysis

Eun Hee Jeon and Junko Yamashita

University of North Carolina at Pembroke / Nagoya University

The present study updates Jeon and Yamashita's (2014) meta-analysis by adding a total of 40 independent samples from 30 additional studies published between 2011 and 2017. Using the method of a quantitative meta-analysis of correlation coefficients, the study synthesizes weighted and, when possible, corrected (for attenuation due to measurement error) correlations between passage-level, L2 reading comprehension and each of the following correlates: decoding, orthographic knowledge, phonological awareness, morphological knowledge, vocabulary knowledge, grammar knowledge, L1 reading comprehension, L2 listening comprehension, working memory, metacognition, and oral reading fluency. The results showed that L2 knowledge variables were invariably strong correlates of L2 reading comprehension. Language-general variables such as working memory and metacognition were, on the other hand, only weakly correlated with L2 reading comprehension. Of the two companion proficiency variables, namely, L2 listening comprehension and L1 reading comprehension, the former showed a much stronger correlation with L2 reading comprehension. In sum, these results indicated that L2 reading comprehension is much more strongly associated with L2 knowledge rather than language-general, cognitive or metacognitive variables or the long-assumed general reading abilities indicated by L1 reading comprehension. Lastly, the study introduced oral reading fluency as a promising correlate of L2 reading comprehension.

#### 1. Introduction: Purpose of the present study

The present study is an updated version of the article, “*L2 Reading Comprehension and its Correlates: A Meta-Analysis*”, published in *Language Learning* in 2014 (henceforth, “the 2014 study”). Given that most other meta-analyses published before and after the 2014 study tended to focus on a single or a small set of related variables, the 2014 study made a unique contribution to our understanding of L2 reading comprehension by including multiple correlates ranging from

lower- to higher-order processes involved in reading. In addition, the 2014 study also included two other proficiency variables, namely, L1 reading comprehension and L2 listening comprehension, both of which are theorized to have an important relationship with L2 reading comprehension. The unique contribution of the 2014 study has been well-received as indicated by the citation index of 254 at the time of writing this chapter in 2020 (<https://scholar.google.com/citations?user=aM3eHTUAAAJ&hl=en>). The present study (henceforth, the 2020 study) updates the 2014 study by including studies published between 2011 (the search end date of the 2014 study) and July 2017, and by including oral reading fluency, a correlate which was not included in the 2014 study.

Some key changes noted between the 2014 study and the 2020 study are summarized in Table 1. The first notable change is the substantial increase in the body of relevant research which led to the increase of high-evidence correlates. In the 2014 study, correlates with the number of effect sizes bigger than fifteen were considered to be high-evidence correlates. Correlates with effect sizes ranging between five and fourteen, on the other hand, were considered as low-evidence correlates due to the concern of less than trustworthy and generalizable findings. Based on this criterion, only four of the ten included correlates had qualified as high-evidence correlates in the 2014 study. In contrast, in the 2020 study, seven of the eleven included correlates qualified as high-evidence correlates with effect sizes ranging from nineteen to 51. The second point of note is the varying levels of research attention different correlates have been receiving. Always a popular variable in reading research, vocabulary knowledge, which had the highest number of effect sizes in the 2014 study ( $k = 31$ ) was again the most popularly investigated correlate in the 2020 study ( $k = 51$ ) with an increase of 20 effect sizes. Interestingly, and possibly reflecting a strong interest in language-universal reading abilities as expressed in Linguistic Interdependence Hypothesis (Cummins, 1978, 1981) and the Simple View of Reading (Hoover & Gough, 1990), L1 reading comprehension also saw a notable increase of twelve effect sizes from the 2014 study, making it the second most popularly investigated variable of the 11 correlates included in the 2020 study. Decoding ( $k$  size increase of 9, percent increase of 45%) and phonological awareness ( $k$  size increase of 9, percent increase of 82%), morphological knowledge ( $k$  size increase of 8, percent increase of 133%) and grammar knowledge ( $k$  size increase of 8, percent increase of 44%), working memory ( $k$  size increase of 9, percent increase of 90%), and L2 listening comprehension ( $k$  size increase of 6, percent increase of 90%) showed moderate but sustained interest in these variables. On the other hand, orthographic knowledge and metacognition only showed a minimal increase in their effect size ( $k$  size increase of 1, percent increase of 20% and 10%, respectively). It is especially intriguing that metacognition, which had seen a moderate level of research interest as documented in the 2014 study had an increase by one study only. Although it

is hard to tell why metacognition as a reading correlate has lost its popularity in recent years, a few conjectures can be made with caution; (1) as reported in the 2014 study, metacognition was found to be the weakest reading correlate of 10 included correlates ( $r = .32$ ). With the exception of Van Gelderen et al. (2004), which yielded an effect size of .87, most of the effect sizes were in the low to medium range. The relative underperformance of metacognition may have led to the less than enthusiastic response among researchers investigating relative contributions of reading components.

**Table 1.** Summary of changes between the 2014 and 2020 study

Variables included (lower to higher processes)	2014 Study (search end date: 2011)	2020 Study (search end date: 2017)	Increase in $k^a$ size from 2014 (increase in %)
<b>Decoding</b>	<b>20</b>	<b>29</b>	9 (45%)
Orthographic knowledge	5	6	1 (20%)
<b>Phonological awareness</b>	<b>11</b>	<b>20</b>	9 (82%)
Morphological knowledge	6	14	8 (133%)
<b>Vocabulary knowledge</b>	<b>31</b>	<b>51</b>	20 (56%)
<b>Grammar knowledge</b>	<b>18</b>	<b>26</b>	8 (44%)
<b>L1 reading comprehension</b>	<b>22</b>	<b>34</b>	12 (55%)
<b>L2 listening comprehension</b>	<b>14</b>	<b>20</b>	6 (43%)
<b>Working memory</b>	<b>10</b>	<b>19</b>	9 (90%)
Metacognition	10	11	1 (10%)
Oral reading fluency	Not included	8	NA (NA)

*Note.* High-evidence correlates (i.e., correlates with 15 or more effect sizes) are indicated in bold.

a. The number of effect sizes.

## 2. A review of reading theories and models that influenced the primary studies of L2 reading comprehension between 1979–2017

Before reporting on the quantitative research synthesis results of the included studies, we present in this section an updated account of the reading theories, models, and trends in research practices which have influenced L2 reading comprehension studies published between 1979 and 2017.

The first noticeable trend was the use of the multicomponent approach to reading at the macro- and micro-level. That is, researchers not only conceptualize L2 reading comprehension as a complex construct composed of various cognitive, linguistic, conative, and affective variables, but they also assume that each of those components is made up of their own subcomponents. This trend is most pronounced among the studies investigating the contributions of morphological

awareness, vocabulary knowledge, and working memory to L2 reading comprehension. For example, morphological awareness was at times further separated into different subcomponents such as derivational or compound awareness so that their relative contribution to reading comprehension could be examined (e.g., Lam et al., 2012; Pasquarella et al., 2011; Xue & Jiang, 2017). Similarly, vocabulary knowledge was also divided into its own subcomponents such as vocabulary depth or breadth for a closer study of their relative contributions to reading comprehension (e.g., Horiba, 2012; Guo & Roehrig, 2011; Zhang, 2012). Lastly, in line with the research tradition which acknowledges various dimensions and assessment methods of working memory (e.g., processing vs. storage component, simple vs. complex working memory, verbal vs. nonverbal memory), L2 reading studies included here have also separated working memory into subcomponents to examine their relative contributions to reading comprehension (e.g., Kieffer & Vukovic, 2013).

Whether the investigation is at the macro- or micro-level, studies taking the multicomponent approach to reading are often interested in examining the relative importance of different reading components in accounting for overall reading outcome. On a larger scale, researchers have investigated the relative importance of language-specific vs. language-general reading components in achieving reading comprehension. This line of research is perhaps best embodied by Alderson's (1984) historical question which asked whether reading in a foreign language is a reading problem or a language problem. Whether linguistic knowledge, which is typically operationalized as sublexical, lexical, and syntactic knowledge in the L2, plays a more important role than language-general reading processes and mechanisms such as inferencing, comprehension monitoring and strategy use has been consistently investigated by L2 reading researchers as clearly indicated in the primary studies included in the present study (e.g., Nassaji & Geva, 1999; Olmez, 2016; Rydland et al., 2012). Similarly, the studies utilizing the multicomponent approach on a smaller scale also aim to examine the relative contributions of subcomponents of a reading component to overall reading outcome (e.g., contributions of different dimensions of syntactic awareness to reading comprehension in Siu & Ho [2015], contributions of vocabulary depth vs. breadth to reading comprehension in Zhang [2012]) (Zhang & Koda, 2018).

Perhaps the latest trend in the multicomponent approach to reading is the move away from the heavy focus on cognitive variables in pursuit of a more holistic model of reading comprehension that includes affective, conative, and sociocultural variables. In contrast with other L2 skills such as listening, speaking, or writing where affective, conative, and sociocultural variables have received well-balanced attention from the research community, the investigation of L2 reading has been disproportionately focused on cognitive variables. A review of the primary studies included here seems to indicate, however, that reading researchers are now also

realizing the importance of non-cognitive variables in accounting for variability in reading outcome; studies like Jia et al. (2014), which argued that L2 literacy skills must be understood along with socio-cultural variables such as acculturation, Winke (2013), which pursued a more comprehensive reading model including both conative and cognitive variables, or Rai et al. (2015), which investigated the role of stress on reading through its effects on working memory, represent this trend.

Other themes of the research domain entail more fundamental questions about the nature of comprehension. One such line of research entails whether language comprehension is governed by a set of general cognitive resources and mechanisms regardless of input modes (e.g., aural vs. written) or even languages (e.g., Gernsbacher, 1991). One key influencer of this line of research is Hoover and Gough's (1990) Simple View of Reading. This parsimonious model proposes that reading comprehension is largely determined by two variables: decoding (defined as "efficient word recognition" [p. 130]) and linguistic comprehension (defined as "the ability to take lexical information and derive sentence and discourse interpretations" [p. 131]). The simplicity of this model, amid a proliferation of highly elaborate multicomponent approaches to reading, certainly gained traction among researchers. The Simple View of Reading has since been empirically and successfully tested for its utility (e.g., Gottardo & Mueller, 2009) and has inspired reading assessment research among young English Language Learners (ELLs) (e.g., Grant et al., 2012).

Not unrelated to the pursuit of the language-universal, general cognitive processes and mechanisms underlying reading comprehension, the investigation of crosslinguistic transfer of reading subskills has also been a key motivator of many studies included here. Cummins' (1979) Linguistic Interdependence Hypothesis is perhaps the most frequently referenced theoretical framework in this line of research (e.g., Baker et al., 2011; Bernhardt & Kamil, 1995; Koda, 1998) which investigated to what extent reading subskills in different languages are correlated with or predict later reading outcome. It is interesting to note that this line of research is also taking the componential approach at a more microscopic level; for example, some researchers have investigated whether reading subskills such as orthographic knowledge or morphological awareness are transferable (e.g., Deacon et al., 2007; Koda, 1998; Li et al., 2012; Pasquarella et al., 2011) and if so, to what extent.

One relatively new, yet noteworthy trend in L2 reading research is the influence of reading fluency and related constructs such as automaticity (Segalowitz, 2003) and processing efficiency of lower-order reading components (e.g., Baker et al., 2011; Ding et al., 2011). The fundamental premise of this line of research is that certain cognitive resources (e.g., working memory or attention) necessary for reading comprehension are limited in capacity. One way to optimize the use of such limited-capacity resources is to automatize all reading-related processes that are

subject to automatization, thereby minimizing the cognitive resources in use. This then, maximizes residual resources available for higher-order processes that are less likely to be subject to automatization (e.g., inferencing, comprehension monitoring, troubleshooting), and the outcome is improved reading comprehension. Reading fluency, a marker of automatized lower-level processes, has long been considered an important topic among L1 reading research for this reason (Jeon, 2018). Unlike L1 reading, however, the investigation of reading fluency among L2 readers comes with a set of challenges such as an interference of foreign accents in assessment or the possibility of comprehension without recoding, i.e., the process of mapping phonemes on graphemes and sounding them out (see Lems [2003] and Jeon [2018] for further discussion on this matter). The emerging interest in reading fluency among L2 reading researchers, however, indicates that despite these challenges reading fluency is increasingly recognized as a noteworthy variable in L2 reading.

### 3. Review of correlates of L2 reading comprehension

In this section, we provide a brief review of the eleven reading correlates included in this study in the following order: (1) high-evidence correlates: decoding, phonological awareness, vocabulary knowledge, grammar knowledge, L1 reading comprehension, L2 listening comprehension, and working memory; (2) low-evidence correlates: orthographic knowledge, morphological knowledge, metacognition, and oral reading fluency. As the more comprehensive theoretical review of these variables is provided in Chapter 2, we limit our discussion here to what we consider to be most necessary for understanding the present study. Lastly, we would like to note that within each of the high-evidence vs. low-evidence correlates category, the correlates were generally ordered from lower- to higher-order variables.

#### 3.1 High-evidence correlates

##### 3.1.1 *Decoding*

Following the 2014 study, we maintain that L2 decoding is “a process during which a reader converts letters (graphemes) to sounds (phonemes) and, essentially, to language,” (Jeon & Yamashita, 2014, p. 163). Similarly, to refer to Chapter 2 of this volume, decoding is “the ability to transform written symbols to their phonological forms utilizing grapheme-phoneme correspondence rules” (Yamashita, 2020, p. 12). As indicated in these definitions, decoding is a subskill unique to reading because it involves the processing of the written language. Subject to automatization through repeated practice, efficient decoding frees up the cognitive resources that are limited in capacity, leaving residual resources available for higher-order



comprehension processes or lower-order processes that have not been automatized yet. Considered as a key contributor of successful reading comprehension by L1 literacy researchers, biliteracy researchers, and L2 reading researchers alike, research interest in decoding does not seem to have waned (e.g., Edele & Stanat, 2015; Goodwin et al., 2015; Kieffer & Vukovic, 2013). In these studies, decoding ability is labelled in various ways; some commonly used labels of decoding include, but are not limited to, phonological decoding and word recognition. As for the operationalization of decoding, studies have used accuracy or efficiency (accuracy and speed) of word reading, pseudoword reading, letter identification/naming, and at times, reading rate and accuracy of word- or passage-level oral reading. As with the 2014 study, we take an inclusive approach and have considered all operationalizations of decoding that fall within the aforementioned range.

### 3.1.2 *Phonological awareness*

In the 2014 study, we defined phonological awareness as “the reader’s sensitivity concerning the segmentation, identification, and manipulation of L2 sound structures such as phoneme, syllable, onset, coda, or rhyme” (Jeon & Yamashita, 2014, p. 166). This definition is also mirrored in Grabe and Stoller (2019, p. 279): “general ability of learners to recognize phonemic sounds in a word, syllables in a word, or syllable parts within a syllable”. As noted in Chapter 2 of this volume, unlike decoding, phonological awareness is a pre-literate, metalinguistic resource that is used for both oral and written language processing; it has been reported that monolingual children become aware of which phonetic distinctions are phonemic in their native language as early as between the 6th and 12th month of age (Kuhl et al., 2006). Young children rely on phonological awareness to analyze a stream of speech sounds into meaningful units (e.g., phonemes, syllables). When they start learning to read, they use phonological awareness and its interrelated processes such as decoding and lexical processing to construct meaning from the written language. To read in L2, readers must learn a new set of phonological rules specific to the target language as well as the rules of phoneme-grapheme correspondence. At the forefront of written language processing, phonological awareness has been a popular variable of study among L2 reading researchers with a focused interest in younger L2 readers (e.g., Abu-Rabia & Sanitsky, 2010; Lam et al, 2012; Limbird et al., 2014) and older L2 readers with a beginning-level proficiency (Koda, 1998), a trend which indicates the importance of phonological awareness as a foundational skill among less experienced readers. Among these studies, phonological awareness is often operationalized as the ability to manipulate target language phonemes although other test tasks tapping into syllable-level awareness are also used (e.g., Grant et al., 2012). Examples of frequently used test tasks include phoneme deletion (e.g., Carlisle et al., 2012), substitution (e.g., Koda, 1998), blending (e.g., Gottardo



& Mueller, 2009), and segmentation (e.g., Swanson et al., 2011). Such tests are often administered individually so that the proctor can provide oral input (e.g., “Say *tree* without the sound *r*”) or, in the case of group testing, involve audio recorded input (Koda, 1998). The present study considered all measurements tapping into the aforementioned construct definition of phonological awareness acceptable.

### 3.1.3 *Vocabulary knowledge*

In this study, we maintain our definition of L2 vocabulary knowledge as a multi-dimensional construct as we earlier proposed in the 2014 study. That is, we view vocabulary knowledge as comprised of size and depth involving knowledge of word form, meaning and use in receptive/productive and written/spoken modes of language use (Nation, 2013). Along with grammar knowledge, vocabulary knowledge has long been considered as a key component of L2 knowledge (Yamashita & Shiotsu, 2017) and has been popularly studied among L2 reading researchers. Beyond the obvious explanation that the knowledge of word form, meaning, and use is essential for reading comprehension, Perfetti’s (2007; Perfetti & Hart, 2002) Lexical Quality Hypothesis comprehensively articulates the relationship between the quality of lexical knowledge representation and reading outcome. Rooted in usage-based language learning, the Lexical Quality Hypothesis considers the lexical quality of a word to be high if its representation is complete (i.e., the orthographic, phonological, grammatical, and semantic information of the word is present in the inactive memory) and stable (i.e., the high quality of lexical representation is sustained long enough for any necessary subsequent processes to complete). It follows then, that readers who have high-quality representations for more words as well as the ability to quickly suppress irrelevant activation (e.g., a homophone) from contextual cues can achieve better reading comprehension through their efficient lexical processing. In contrast, readers who have incomplete and unstable representations are likely to have a poorer reading outcome due to their inefficient lexical processing (e.g., activating the wrong homophone and getting confused by it, failure to locate the right meaning of a polysemous word). While high frequency words are likely to have higher quality representation than low frequency words even among skilled readers (usage-based language learning), they are likely to have overall higher quality representation for both high and low frequency words compared to poor readers (Perfetti & Hart, 2002), thereby leading to a better reading outcome.

Due to its multidimensional nature, vocabulary knowledge is operationalized and assessed in various ways that include, but are not limited to, the following examples: estimating the breadth (e.g., Vocabulary Levels Test by Nation [1983], Vocabulary Size Test by Nation & Beglar, [2007]) and depth (e.g., Word Associates Test by Read (1998), synonym/collocate selection), and in the case of young

learners, picture selection based on aural input (e.g., Peabody Picture Vocabulary Test). The present study considered all measures tapping into various dimensions of L2 vocabulary knowledge acceptable. Having said that, however, the focal construct of vocabulary knowledge examined in reading research is learners' knowledge of the meaning of words rather than the grammatical functions or knowledge of word parts (morphemes). This focus on lexical meaning operationally differentiates vocabulary knowledge from theoretically overlapping constructs such as grammar and morphological knowledge.

#### 3.1.4 *Grammar knowledge*

In this study, grammar knowledge is defined as all types of linguistic knowledge involved in the morpho-syntactic and syntactic processing of a written text (e.g., knowledge needed to recognize parts of speech of words, knowledge of inflectional morphemes, knowledge needed to recognize a phrase or a clause, knowledge of how words are ordered in L2). As a reader begins to read, she starts unpacking syntactic and semantic information of words. For example, the reader identifies groups of words that form a unit (e.g., a phrase or clause), identifies the clause element of a given phrase (e.g., subject, object, or adverbial), identifies the syntactic role of a given clause (e.g., adverbial clause vs. main clause) and consequently, builds propositions or idea units (see Grabe [2009] for a more detailed account of syntactic processing in reading). In other words, grammar knowledge is used to integrate semantic information encoded in individual words forming a text (De Jong & Van Ginkle, 1992). Efficient syntactic processing, therefore, is essential for successful reading comprehension.

Regarding the role of syntactic knowledge as a reading component, one important question concerns how different types of syntactic knowledge (e.g., explicit vs. implicit knowledge) relate to reading comprehension. Traditionally, grammar tests have often measured explicit knowledge. Such tests may ask the examinee to recognize syntactic or morphosyntactic errors in a sentence (e.g., the Structure and Written Expression subtest of TOEFL PBT, an untimed or unspeeeded grammaticality judgment test [GJT]) and correct them (e.g., Oh, 2015), to complete a sentence by providing a well-formed word (e.g., Khaldieh, 2001), or to rearrange jumbled up words to construct a well-formed sentence (e.g., Siu & Ho, 2015). Recently, however, more L2 reading studies have attempted to isolate implicit knowledge of grammar (e.g., Erçetin & Alptekin, 2013; Zhang, 2012) to examine its contribution to reading comprehension. A highly speeded GJT can tap into implicit grammar knowledge as it would not allow the test taker the time to access explicit knowledge. It must be noted, however, that if a GJT is timed, yet not sufficiently speeded for the test takers, they may still engage explicit knowledge, presenting a threat to the construct validity of the test (see Ellis, [2011], for a more detailed discussion

on the measurement of implicit knowledge of grammar). For the present study, all measures that were considered relevant to any form of grammar knowledge were considered acceptable.

### 3.1.5 *L1 reading comprehension*

Thoughts on the relationship between L1 and L2 reading have evolved extensively over the years, advancing theories of L2 reading. Earlier works have argued for a stronger, and rather straightforward relationship between the two. Rigg (1977) and Clarke (1979) argued that reading processes are similar across languages. As a result, L2 reading abilities were thought to entirely depend upon L1 reading abilities, and poor L2 reading was, rather simplistically, attributed to poor L1 reading abilities (Clarke, 1979). Cummins, (1978, 1979) in his Interdependence Hypothesis, similarly argued that the linguistic knowledge a child gains in L1 provides a firm basis for L2 linguistic knowledge, suggesting that there may be a common underlying proficiency that supports both L1 and L2 reading. This view, however, gradually shifted to another direction as researchers increasingly came to acknowledge the importance of L2 knowledge (e.g., Clarke's [1980] Short Circuit Hypothesis, Cummins' [2000] Linguistic Threshold Hypothesis) and as more studies unveiled the more complex relationship between L1 and L2 reading. In his well-known review, Alderson (1984) rejected the notion that the primary source of problems in L2 reading is poorly developed L1 reading abilities. Rather, he argued that poor L2 readers have poorly developed L2 knowledge and that this is the cause of their poor reading outcome in L2. More fine-grained studies examining what is and is not likely transferable between L1 and L2 reading abilities expanded our understanding of the relationship between L1 and L2 reading. Such studies suggest that if the languages involved are dissimilar in writing system and typology, then orthographic, vocabulary, and syntactic knowledge in L1 are unlikely to transfer to corresponding L2 skills (see Grabe [2009] and Jeon & Yamashita [2014] for a more detailed review of relevant studies). It seems, however, among bilingual children who are exposed to two languages early and consistently in a naturalistic setting, there is stronger evidence for crosslinguistic transfer of linguistic knowledge (e.g., Melby-Lervåg & Lervåg, 2011), suggesting that the observed association between L1 and L2 reading may indeed be attributable to the transferable linguistic knowledge. The present study also assumes that the nature of relationship between L1 and L2 reading is complex and nuanced, and that its investigation must go above and beyond simply capturing the size of the association between the two variables. To this end, the present study includes primary studies which assessed all forms of passage-level reading comprehension in L1 using various measures.

### 3.1.6 *L2 listening comprehension*

Qualitatively, L2 listening comprehension is unlike most of the reading correlates included in this study because it is not a component of reading, but rather a companion proficiency variable of L2 reading comprehension. Theoretical grounds for studying listening comprehension to better understand reading comprehension can be found in the influential research programs of the 1990s. Gernsbacher, Varner, and Faust (1990) first proposed the notion of General Comprehension Skill. This notion assumes that the same set of cognitive components and processes are involved in comprehending linguistic input of different modalities (i.e., written or aural) and even nonlinguistic input (e.g., textless movies or picture stories). In addition to the generally high correlations between reading and listening comprehension, the presence of General Comprehension Skill is further supported by ample empirical evidence showing that comprehenders process input, be it aural, written, linguistic, or non-linguistic, in a similar way. For example, comprehenders' construction of an episode structure, inferencing, and forgetting behavior were found to be similar across input materials of different linguistic modalities or language use (e.g., text, textless movie, textless cartoons) (Gernsbacher et al., 1990). In a similar vein, Simple View of Reading (Gough & Tunmer, 1986; Hoover & Gough, 1990) also views reading comprehension as comprised of decoding ability and linguistic comprehension ability, rather than comprehension abilities of written texts only. Again, the theoretical assumption here is that both listening and reading comprehension are controlled by common, linguistic comprehension abilities (Hoover & Gough, 1990). For this reason, many studies inspired by the Simple View of Reading include an oral language comprehension test (i.e., listening comprehension test) to measure linguistic comprehension (e.g., Goodwin et al., 2015; Proctor et al., 2005). In the present study, we considered all measures that tap into passage-level listening comprehension acceptable.

### 3.1.7 *Working memory*

Working memory is a cognitive system which temporarily maintains various types of information (e.g., verbal, non-verbal, visual, auditory) activated for processing (Grabe, 2009). Working memory is limited in capacity and is considered to have multiple components (Baddeley & Hitch, 1974). While different models of working memory may vary in detail, most researchers agree that it has two key functions: storage and processing (Daneman & Carpenter, 1980). In reading, relevant information such as decoded graphemes, the semantic and syntactic information of recognized words, or the proposition of the previously read sentences remain activated in working memory as more incoming information is supplied to form a coherent model of the text. Efficient storage and processing functions of working

memory, therefore, are essential for successful reading comprehension. For this reason, working memory has been popularly investigated as a potentially important contributor to L2 reading among reading researchers.

Empirical findings on the strength of the relationship between working memory and L2 reading comprehension, however, has not been straightforward. One possible reason for this is the effect of working memory test tasks (Shin, 2020); in short, complex tasks (tasks that tap both storage and processing functions) have shown to correlate higher with L2 reading comprehension than simpler tasks that only tap storage functions. In addition, testing language (e.g., L1 vs. L2) and the nature of the input material (e.g., visual, numerical, verbal) may also result in systematic variations between the relationship between working memory and L2 reading comprehension. The present study takes an inclusive approach and includes all measures of working memory.

### 3.2 Low-evidence correlates

#### 3.2.1 *Orthographic knowledge*

While orthographic knowledge is closely related to decoding and phonological processing, it uniquely involves the knowledge of language-specific, orthographic patterns or characteristics, which encompass both regularly and irregularly spelt words. Nassaji and Geva (1999) noted that readers use orthographic knowledge to evaluate orthographic patterns and to determine whether the given pattern complies with the orthographic conventions of the target language. As noted in Chapter 2 of this volume, the knowledge of language-specific orthographic conventions facilitates efficient word recognition by making contributions above and beyond that of decoding. For example, a Chinese L2 reader who knows that in a phono-semantic compound, the semantic component (i.e., radical) tends to appear on the left and the phonetic component on the right, will be able to efficiently co-activate both pronunciation and meaning from the written input, resembling the word recognition process of a proficient L1 reader (Perfetti & Liu, 2006). Similarly, an English L2 reader who understands that some derivations result in spelling changes to the stem (e.g., elude + -ive = elusive) will be able to easily infer the meaning of the newly derived word rather than assuming that it is an entirely different word. Reflecting this construct definition, orthographic knowledge tests also often tap into the ability to identify language-specific orthographic conventions. For example, on Treiman's (1993) orthographic constraints test, the test taker is asked to dictate multi-morphemic words that follow typical orthographic structures of L2. Other tests of orthographic knowledge or processing ask the test taker to determine whether the provided pseudoword conforms to the orthographic conventions of the

target language (e.g., Abu-Rabia & Sanitsky, 2010). In the present study, all measures tapping into the earlier mentioned orthographic knowledge of orthographic processing were considered acceptable.

### 3.2.2 *Morphological knowledge*

Inflection, derivation, and compounding exist across most languages (Lam et al., 2012). Inflectional morphemes encode grammatical information such as case (e.g., in Russian, the accusative marker *-a* is added to the end of a word to indicate its object status), gender (e.g., in French, the feminine gender marker *-se* in the word, *chanteuse* [*singer*], indicates that the word's grammatical gender is feminine), tense (e.g., in English, the tense marker, *-ed*, indicates that the verb is in the past tense), or number (e.g., in English, the number marker, *-s*, indicates that word is plural). Derivational morphemes encode both semantic (e.g., the prefix *un-* means “not”, “lacking”, or “the opposite of”) and grammatical information (e.g., the suffix *-tion* indicates that the word is a noun). Compounding is a process through which two or more words are combined to create a new word (e.g., *book + case = bookcase*). Given this, we can say that morphological knowledge is multidimensional in nature and overlaps with vocabulary as well as grammatical knowledge as defined in the present study.

Since most words of most languages are multimorphemic (Libben, 2008), the ability to recognize the morphemic structure of a word (i.e., morphological awareness), and knowing the meaning and function of recognized morphemes (i.e., morphological knowledge) are crucial to achieving good reading outcome. We also expect that the importance of morphological awareness and knowledge as a reading component will increase as the reader takes on texts that are informationally dense and involve higher proportions of morphologically complex (i.e., multimorphemic) words.

The operationalization of morphological knowledge within L2 reading research closely reflects the construct definition of morphological knowledge (Ke et al., 2021). Test tasks often involve decomposing multimorphemic words (e.g., base extraction task in Goodwin et al., [2015] and Leider et al. [2013]), or creating words using the knowledge of inflectional, derivational morphology or compounding (e.g., Jeon, 2011; Pasquarella et al., 2011). Although it is true that morphological knowledge overlaps with vocabulary and grammar knowledge as mentioned above, measurement practices in reading research operationally differentiate morphological knowledge from these types of linguistic knowledge, if not completely. Measures of morphological knowledge tap into knowledge of word parts (morphemes) rather than the whole words (vocabulary) or focus on, even not exclusively, derivational morphology and compounding rather than inflectional morphology (contrarily,

grammar tests tend to measure knowledge of inflectional morphemes rather than the other two). In this study, we consider all measures that tap into the awareness and knowledge needed to recognize and manipulate all types of morphemes acceptable.

### 3.2.3 *Metacognition*

Metacognition, “the knowledge of what we know” (Grabe & Stoller, 2019, p. 42), influences almost all processes of reading through its executive functions (e.g., planning, comprehension monitoring, troubleshooting). As the reader approaches a text, she must decide the purpose of reading (planning or goal setting), be it skimming for a main idea, scanning for targeted information, reading to learn, or reading to memorize. The purpose of reading will then determine the optimal reading rate which can range from 138 words per minute (for reading for memorizing) to 600 words per minute (for reading for scanning) in the case of college-level, L1 readers (Carver, 1997). While reading, the reader may also notice problems arising in comprehension due to various reasons (e.g., unknown key vocabulary, complex sentence structures, trouble with identifying the referent of a pronoun). An astute reader will then engage their monitoring strategies to assess the level of comprehension, to identify the source of the comprehension problem, and attempt to troubleshoot to achieve better comprehension (comprehension monitoring and troubleshooting).

Grabe and Stoller (2019) note that the role of metacognition may become even more important among L2 readers than among L1 readers because the majority of L2 readers learn to read in L2 after acquiring literacy skills in L1 and after having received (often explicit) language instruction in L2. These prior experiences naturally enhance their metalinguistic, and consequently, metacognitive awareness as they engage in reading in L2.

At present, operationalization of metacognition in L2 reading component research is largely limited to self-report surveys/assessments, and this trend is well reflected in the primary studies included in the present meta-analysis. We have included all such types of measures in this study.

### 3.2.4 *Oral reading fluency*

Fluent oral reading is characterized by a speedy, accurate, and when reading a connected text, an appropriately phrased oral rendition of written texts (e.g., word lists or reading passages) (Grabe, 2009; Jeon, 2012). As a trait of skilled reading across text types and reading purposes, fluent reading is a sign that the reader’s lower-level reading processes such as decoding, lexical-, and syntactic processing have been automatized and run in a well-coordinated manner while consuming



minimal amounts of capacity-limited resources such as attention or working memory. Oral reading fluency has often been used as a proxy for silent reading fluency, probably due to its ease of observation, assessment, and task fidelity (i.e., one cannot simply skip to the end of the passage and claim that he/she has finished reading).

While oral reading fluency has enjoyed much attention in L1 and bilingual reading research (Jenkins et al., 2003; Klauda & Guthrie, 2008), it has only started getting traction among L2 reading researchers as indicated by the relatively small group of studies included in the present study (6 studies, 8 independent samples). There may be a few reasons that explain the lack of popularity of oral reading fluency among L2 researchers. One such reason is the concern of “comprehending without recoding” (Lems, 2006, p. 240). That is, compared to L1 reading, in L2 reading there may be more instances of successful word recognition and text comprehension in a silent condition. Studies have suggested that this is especially likely with L2 readers in cultures where oral reading is not encouraged among older readers (Jeon, 2012; Taguchi, 1997). To our knowledge, however, it is unclear whether this concern has sufficient empirical evidence. For example, Sparks and Luebbbers (2018) reported that all poor L2 readers (American students studying Spanish in the US) in their study were either categorized as hyperlexic (good decoding, poor comprehension) or garden variety (poor decoding, poor comprehension), but never dyslexic (poor decoding, good comprehension), a result which casts doubt on how common comprehension without recoding really is in L2. In response, the present study aims to empirically examine the association between oral reading fluency and reading comprehension in L2.

In order to disambiguate decoding and oral reading fluency (as decoding is sometimes measured by oral reading of word or pseudoword lists), the present study only included studies that assessed oral reading fluency via passage-level oral reading.

The present study examined the following research questions:

1. What are the strengths of association between passage-level L2 reading comprehension and the following eleven correlates?: L2 decoding, L2 orthographic knowledge, phonological awareness, L2 morphological knowledge, L2 vocabulary knowledge, L2 grammar knowledge, L1 reading comprehension, L2 listening comprehension, working memory, metacognition, L2 oral reading fluency?
2. Do theoretically motivated moderating variables significantly moderate the relationship between passage level L2 reading comprehension and its correlates?



## 4. Method

### 4.1 Literature search and inclusion criteria

The search methods and inclusion criteria largely remained consistent with those of our 2014 study with a new search end date of July, 2017. Five major electronic databases (ERIC, LLBA, ProQuest, PsycINFO, PsycARTICLES) and 29 journals within applied linguistics, L1 and L2 literacy studies, and education were electronically or manually searched to locate eligible studies. Since the search end date of the 2014 study was May 2011, literature search for the present study covered studies published between May 2011 and July 2017. As with the 2014 study, various combinations of the following key terms were used for the electronic database search: L2, reading, abilit\* (truncation), comprehension, predictors, components, componential, analysis, decoding, word recognition, pseudoword, real word, phon\*, phonological awareness, phonemic awareness, orthograph\*, orthographic knowledge, orthographic awareness, morph\*, morphological awareness, morphological knowledge, grammar knowledge, grammatic\*, synta\*, sentence processing, parsing, vocabulary, word knowledge, lexical knowledge, lexical semantic, metacogniti\*, metacognitive awareness, metacognitive knowledge, working memory, oral reading, oral reading accuracy, and oral reading fluency.

After periodicals had been searched, full texts of book chapters and monographs were manually searched. Consistent with the 2014 study, for a quality-control purpose, only studies that were published in refereed journals or by a reputable academic publisher were included in the present study. In addition, eligible studies also had to include at least one measure that assessed passage-level L2 reading comprehension and report correlations between the reading comprehension measure and one of the eleven correlates included in the present study. In order to allow for a moderator analysis involving language types (L1-L2 language distance, script distance), eligible studies also had to report for each sample which L1 and L2 were involved. A literature search for a large-scale meta-analysis like the present study is conducted over a span of a few months and as electronic database search results tend to yield varying numbers of hits, it is nearly impossible to provide an exact number of studies identified at each step of the search period. For this reason, we simply report here that the final dataset included 107 independent samples from 87 primary studies. The included studies are marked with an asterisk in the references section of the chapter.

## 4.2 Acceptable measures of L2 reading comprehension and its correlates

Detailed information on what was considered as acceptable measures of L2 reading comprehension and its correlates is provided in Appendix A

## 4.3 Analytical procedures

In case multiple studies reported on the same data or largely overlapping data (duplicate reporting), only one study was included in the analysis. If the study adopted a longitudinal design and reported multiple correlations taken at different time points, in most cases, the first correlation was used. At times, however, it was deemed necessary to use an alternative approach and choose the correlation most comparable to the rest of the studies in the pool; for example, if most correlations were based on adolescent/adult sample and the longitudinal study reported correlations taken at Grade 7, 8, 9 and 10, the last correlation was included in the study. Studies also often reported correlations between reading comprehension and sub-constructs of a correlate (e.g., a correlation between reading comprehension and derivation, a correlation between reading comprehension and decomposition). In such a case, we used an average value of the reported correlations.

The first and second author of the study independently coded various study features to ensure an acceptable level of intercoder reliability. The percentage agreement of correlations and study features ranged from 96%–99%, indicating a high level of agreement. Any disagreements were resolved through discussion.

## 4.4 Meta-analytic procedures

Study names, sample sizes, correlations weighted for sample size and corrected for attenuation (when possible), and coding results for various moderators were entered in Comprehensive Meta-Analysis Version 2 (Borenstein et al., 2005). In order to compute corrected correlations, we followed Spearman's (1904) formula using the measurement reliability reported in the primary study. When the primary study did not report reliability information, we used the average of the reliability indices of the study pool under analysis.

In order to examine and address the potential file drawer problem, several measures were undertaken: the classic fail-safe  $N$ , Orwin's fail-safe  $N$ , the examination of funnel plots and the use of trim and fill method when appropriate.

Weighted average correlations between reading comprehension and the eleven correlates were then computed along with their 95% confidence interval (CI). A

correlation with its CI exceeding zero indicates that it is statistically significantly different from zero.

When there was a theoretical motivation, a statistical indication (i.e., a statistically significant  $Q$  test of homogeneity, Hedges & Olkin [1985] and a large  $I^2$  statistic), and a minimum of 3 effect sizes per group to be compared with each other, a moderator analysis was carried out in order to identify the source of variation among samples. All analyses, including main and moderator analyses were carried out using random effect model due to our assumption that there would be heterogeneity above and beyond the sampling error.

#### 4.4.1 *Coding of moderator variables*

**Age.** Consistent with the 2014 study, all participants at or younger than grade six (or 12 years old) were coded as Child. All participants older than grade 6 were coded as Adolescent/Adult.

**Language Setting.** The sample was coded as SL if L2 is the language predominantly used in the country where the study was conducted and as FL if L2 is learned only at school.

**L1-L2 Language Distance.** Combinations of Indo-European L1 and Indo-European L2 (or vice versa) were coded as II. Combinations of an Indo-European L1 and a non-Indo-European L2 (or vice versa) were coded as IN.

**L1-L2 Script Distance.** Combinations of an alphabetic L1 and an alphabetic L2 (or vice versa) were coded as AA; an alphabetic L1 and an logographic L2 were coded as AL; and an alphabetic L1 and a mixed-system L2 were coded as AM. When AL or AM categories had only one sample, these two were combined as AN (an alphabetic L1 and non-alphabetic L2).

**L2 Proficiency.** Participants who had less than 1 year of instruction or exposure to L2 were coded as basic (B) and the rest were coded as Beyond Basic (BB)

**Measurement Characteristics.** Different measurement characteristics were coded for different correlates as follows.

**Decoding.** Real word reading vs. pseudoword reading vs. pseudo and real word reading (3 levels)

**Vocabulary.** Researcher-made test vs. standardized (2 levels), selection vs. production (2 levels) and contextualized vs. isolated vs. mixed (3 levels).

**Grammatical Knowledge.** Researcher-made test vs. standardized test (2 levels), and completion test vs. grammaticality judgement test (GJT) vs. other types (3 levels)

**L1 Reading.** Researcher-made test vs. standardized test (2 levels)

**Phonological Awareness.** Researcher-made test vs. standardized test (2 levels)

**Working Memory.** L1 (L1 as the testing language) vs. L2 (L2 as the testing language)

## 5. Results

Table 2 summarizes average correlations between L2 reading comprehension and eleven correlates with 95% CIs (seven high-evidence variables with 19 or more effect sizes and four low-evidence variables with 14 or fewer effect sizes) together with measures to examine a potential file drawer problem and the homogeneity of effect sizes. Classical fail-safe  $N$  and Orwin's fail-safe  $N$  indicated little concern for the file drawer problem in this study. The  $Q$  test was significant and  $I^2$  statistics were reasonably large for all the variables, providing a partial rationale for moderator analyses. Following the principle of the 2014 study, we ran moderator analyses only for high-evidence variables (see Appendix B for all moderator analyses of significant and nonsignificant results). This section first reports on the high-evidence variables with the results of their moderator analyses and proceeds to the low-evidence variables.

**Table 2.** Mean correlations between L2 reading comprehension and correlates

Variable name	$k$	$r$ [95% CI]	Significant test of difference ( $Q$ test)	$I^2$	Classical fail-safe $N$	Orwin's fail-safe $N$	Adjusted effect estimate after trim and fill	Trimmed studies
Decoding	29	.586 [.453–.694]	957.581**	97.076	2398	386	.660 [.556–.742]	7
Phonological awareness	20	.611 [.515–.693]	159.682**	88.101	4135	278	.676 [.593–.745]	5
Vocabulary knowledge	51	.724 [.636–.794]	2806.174**	98.218	7145	872	.799 [.792–.806]	13
Grammar knowledge	26	.697 [.517–.818]	3057.617**	99.182	2075	565	.796 [.682–.873]	7
L1 reading comprehension	34	.483 [.361–.589]	701.173**	95.294	8634	350	.571 [.473–.656]	9
L2 listening comprehension	20	.812 [.638–.907]	2664.370**	99.287	9367	370	.895 [.792–.948]	6
Working memory	19	.334 [.234–.427]	63.025**	71.440	692	123	No adjustments	0
Orthographic knowledge	6	.590 [.341–.761]	94.040**	94.683	879	72	No adjustments	0
Morphological knowledge	14	.635 [.556–.703]	63.844**	79.638	2476	193	No adjustments	0
Metacognition	11	.330 [.083–.539]	337.468**	97.037	766	83	.451 [.237–.623]	3
Oral reading fluency	8	.640 [.521–.741]	61.387**	88.597	969	102	No adjustments	0

Note. Variables in boldface are high-evidence variables.

\*\*  $p < .01$

5.1 Results on high-evidence correlates and their moderators

5.1.1 Decoding

Twenty-nine weighted and corrected correlations from 25 studies involving 4,156 participants were included in the analysis (mean sample size = 143.310, *SD* = 121.28, range = 20–502). Correlations, their significance, 95% CI, and their graphic representations are presented in Figure 1. Participants’ age ranged from kindergarten to postgraduate levels and their L1s were diverse as follows: Spanish (4 samples), Chinese (5 samples), Russian (2 samples), Korean (2 samples), Japanese (2 samples), Farsi (1 sample), Dutch (1 sample), Portuguese (1 sample), Cantonese (1 sample). The most popularly studied L2 was English (27 samples) although Hebrew (1 sample) and German (1 sample) were also represented. The overall mean correlation was in the medium to large range,  $r = .586$ , 95% CI [.453–.694] (Cohen, 1988), and statistically significant ( $p = .00$ ). The variability across studies was significant and large,  $Q(28) = 957.581$ ,  $p = .00$ ,  $I^2 = 97.076$ . As a result, a series of moderator analyses on age, language setting, measurement type, language proficiency, and language difference were conducted to identify the source of this variability. Results showed that language setting was the only statistically significant moderator variable. The moderator analysis results are summarized in Table 3.

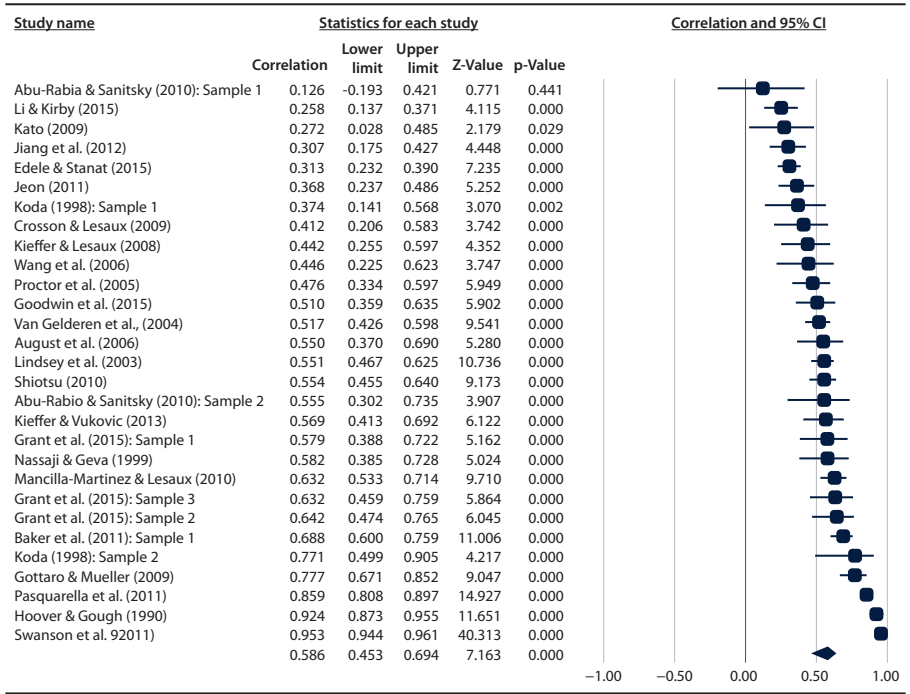


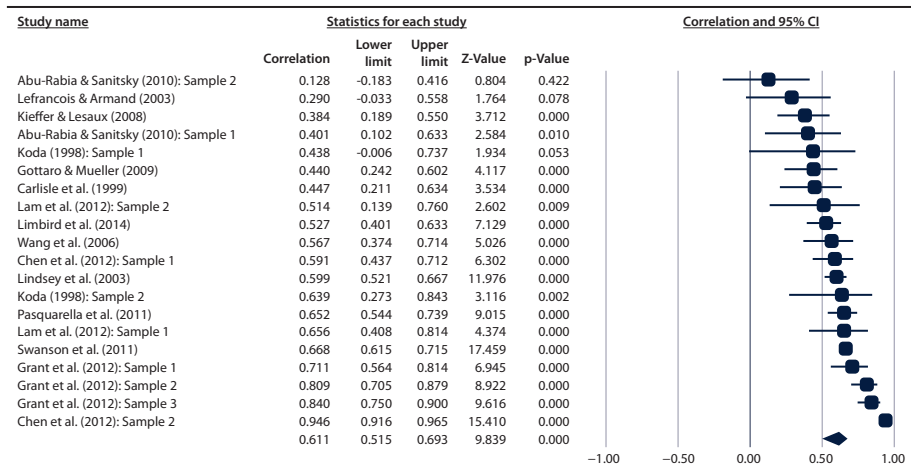
Figure 1. Average correlation (the diamond at the bottom) and correlation with confidence interval for each study correlating L2 decoding and L2 reading comprehension

**Table 3.** Results of moderator analysis for decoding

Moderator variable	<i>k</i>	<i>r</i> [95% CI]	Significant test of difference ( <i>Q</i> test)
<i>Language setting</i>			
FL	6	.424 [.310–.527]	4.174 ( <i>p</i> = .041)
SL	23	.622 [.462–.743]	

### 5.1.2 Phonological awareness

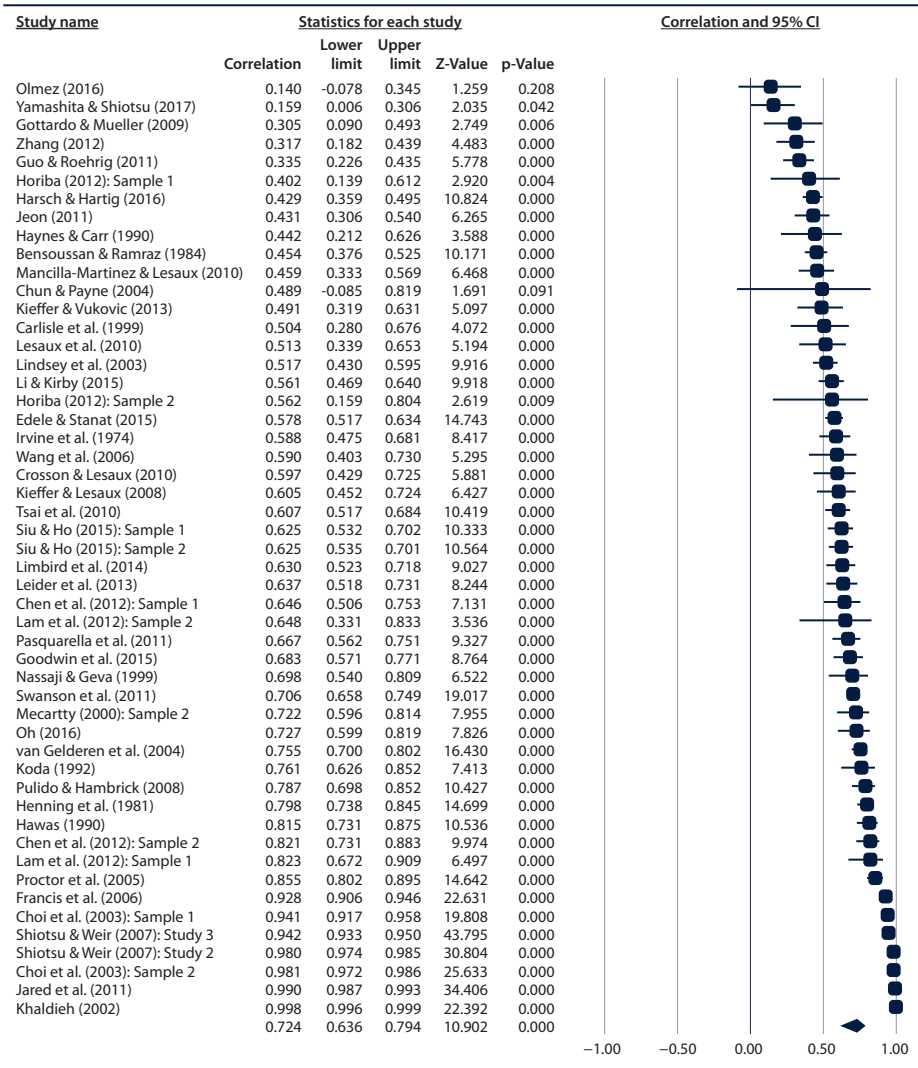
Twenty weighted and corrected correlations from 14 studies involving 1,928 participants were included in this analysis (mean sample size = 96.40, *SD* = 108.67, range = 20–471). Correlations, their significance, 95% CI, and their graphic representations are presented in Figure 2. Participants' age ranged from kindergarten to college levels and their L1s were diverse: Spanish (8 samples), Chinese (6 samples), Cantonese (1 sample), Korean (1 sample), Hebrew (1 sample), Russian (1 sample), Portuguese (1 sample), and Turkish (1 sample). As for the target language, English was again the most frequently studied L2 (17 samples), with Hebrew (1 sample), French (1 sample), and German (1 sample) in the minority. The overall mean correlation was in the medium to large range,  $r = .611$ , 95% CI [.515–.693] (Cohen, 1988), and statistically significant ( $p = .00$ ). The variability across studies was significant and large,  $Q(19) = 159.682$ ,  $p = .00$ ,  $I^2 = 88.101$ . As a result, a series of moderator analyses on age, L1-L2 language difference, L1-L2 script distance, and measurement type were conducted. Results showed, however, that none of the moderator analyses reached a statistically significant level.



**Figure 2.** Average correlation (the diamond at the bottom) and correlation with confidence interval for each study correlating *L2 phonological awareness* and *L2 reading comprehension*

### 5.1.3 *Vocabulary knowledge*

Fifty-one weighted and corrected correlations from 45 studies involving 8,354 participants were included in the analysis (mean sample size = 163.80,  $SD = 138.50$ , range = 13–624). Correlations, their significance, 95% CI, and their graphic representations are presented in Figure 3. Participants' age ranged from kindergarten to postgraduate levels and their L1s were diverse as follows: Spanish (14 samples), Chinese (11 samples), English (6 samples), Korean (5 samples), Japanese (3 samples), Farsi (2 samples), Cantonese (2 samples), Arabic (2 samples), Turkish (2 samples), Hebrew (1 sample), Russian (1 sample), Dutch (1 sample), German (1 sample). English (40 samples) was again the most frequently studied L2 with German (3 samples), Japanese (3 samples), Spanish (3 samples), Arabic (1 sample), and French (1 sample) in the minority. The overall mean correlation was leaning large,  $r = .724$ , 95% CI [.636–.794] (Cohen, 1988), and statistically significant ( $p = .00$ ). The variability across studies was significant and large  $Q(50) = 2806.174$ ,  $p = .00$ ,  $I^2 = 98.218$ . As a result, a series of moderator analyses on age, L1-L2 script distance, language setting, measurement characteristics, language proficiency, and language difference were conducted. Results showed that none of the moderator analyses reached a significant level although script distance was at the significance level of .05 (Table 4).



**Figure 3.** Average correlation (the diamond at the bottom) and correlation with confidence interval for each study correlating L2 vocabulary and L2 reading comprehension

**Table 4.** Results of moderator analysis for vocabulary

Moderator variable	<i>k</i>	<i>r</i> [95% CI]	Significant test of difference ( <i>Q</i> test)
Script distance			
AA	32	.748 [.640–.827]	5.985 ( <i>p</i> = .050)
AL	14	.599 [.512–.674]	
AM	5	.827 [.038–.961]	



### 5.1.4 Grammar knowledge

Twenty-six weighted and corrected correlations from 23 studies involving 6,383 participants were included in the analysis (mean sample size = 245.520  $SD$  = 397.43, range = 38–2,083). Correlations, their significance, 95% CI, and their graphic representations are presented in Figure 4. Participants' age ranged from elementary school to postgraduate levels and their L1s were diverse as follows: Japanese (3 samples), Korean (4 samples), Chinese (4 samples), Spanish (3 samples), Cantonese (2 samples), Farsi (2 samples), English (2 samples), Russian (1 sample), Hebrew (1 sample), Arabic (1 sample), Dutch (1 sample), Turkish (1 sample), and German (1 sample). As for the target language, English was again the most frequently studied (22 samples), with Hebrew (1 sample), Arabic (1 sample), French (1 sample), and Chinese (1 sample) in the minority. The overall mean correlation was in the medium to large range,  $r = .697$ , 95% CI [.517–.818] (Cohen, 1988) and statistically significant ( $p = .00$ ). The variability across studies was significant and large  $Q(25) = 3506.253$ ,  $p = .00$ ,  $I^2 = 99.287$ . As a result, a series of moderator analyses on age, L1-L2 script distance, L1-L2 language difference, language setting, two different types of measurement characteristics, and language proficiency were conducted to identify the source of this variability. Results showed that script distance and measurement type (sentence completion, GJT, other) were the only statistically significant moderator variable. The moderator analysis results are summarized in Table 5.

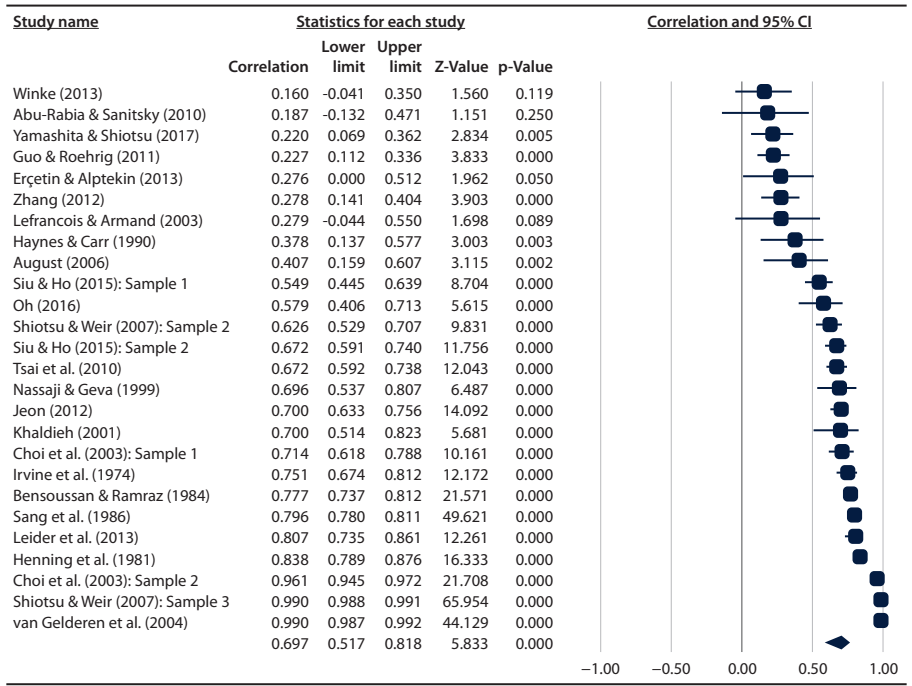


Figure 4. Average correlation (the diamond at the bottom) and correlation with confidence interval for each study correlating L2 grammar and L2 reading comprehension

**Table 5.** Results of moderator analysis for grammar

Moderator variable	<i>k</i>	<i>r</i> [95% CI]	Significant test of difference ( <i>Q</i> test)
Script distance			
AA	16	.750 [.603–.847]	6.085* ( <i>p</i> = .048)
AL	6	.485 [.295–.638]	
AM	4	.737 [–.436–.982]	
Measurement characteristic			
Completion	16	.793 [.604–.898]	9.400**( <i>p</i> = .009)
GJT	5	.367 [.150–.550]	
Other	5	.562 [.331–.730]	

### 5.1.5 *L1 reading comprehension*

Thirty-four weighted and corrected correlations from 25 studies involving 3,982 participants were included in the analysis (mean sample size = 117.12, *SD* = 88.87, range = 20–366). Correlations, their significance, 95% CI, and their graphic representations are presented in Figure 5. Participants' age ranged from kindergarten to college levels and their L1s were diverse: Chinese (8 samples), English (7 samples), Spanish (6 samples), Japanese (5 samples), Cantonese (3 samples), Korean (2 samples), Hebrew (1 sample), Russian (1 sample), and Dutch (1 sample). As for the target language, English was again the most frequently studied L2 (26 samples), with French (5 samples), Hebrew (2 samples), and Korean (1 sample) in the minority. The overall mean correlation was medium size,  $r = .483$ , 95% CI [.361–.589] (Cohen, 1988), and statistically significant ( $p = .00$ ). The variability across studies was significant and large,  $Q(33) = 701.173$ ,  $p = .00$ ,  $I^2 = 95.294$ . As a result, a series of moderator analyses on age, L1-L2 script distance, L1-L2 language difference, language setting, measurement type (researcher-made vs. standardized) and language proficiency were conducted to identify the source of this variability. Results showed that L1-L2 language difference was the only statistically significant moderator variable. The moderator analysis results are summarized in Table 6.

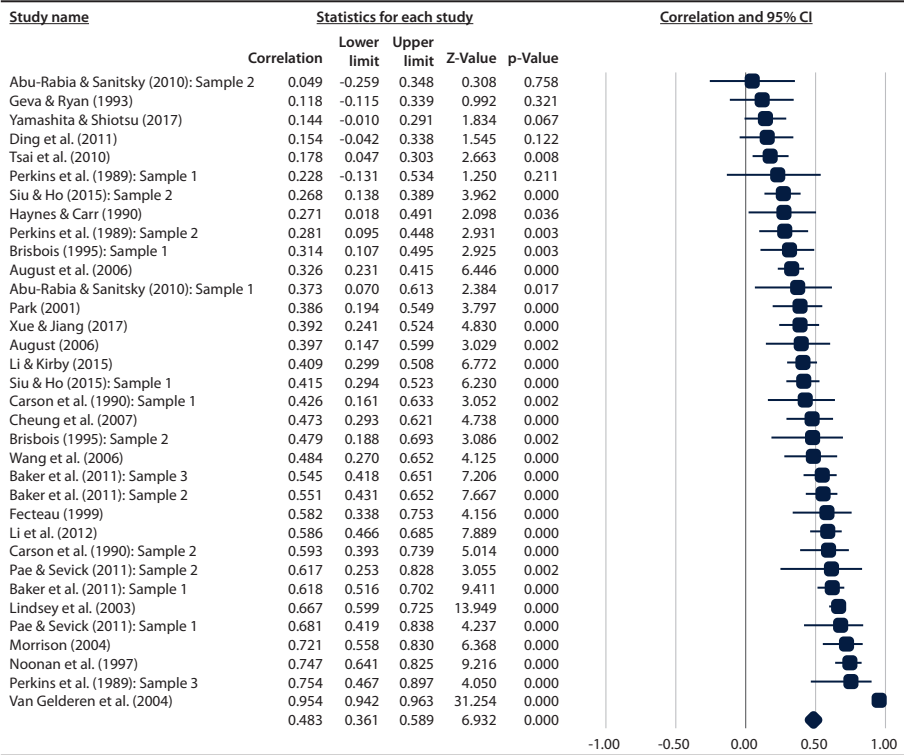


Figure 5. Average correlation (the diamond at the bottom) and correlation with confidence interval for each study correlating L1 reading comprehension and L2 reading comprehension

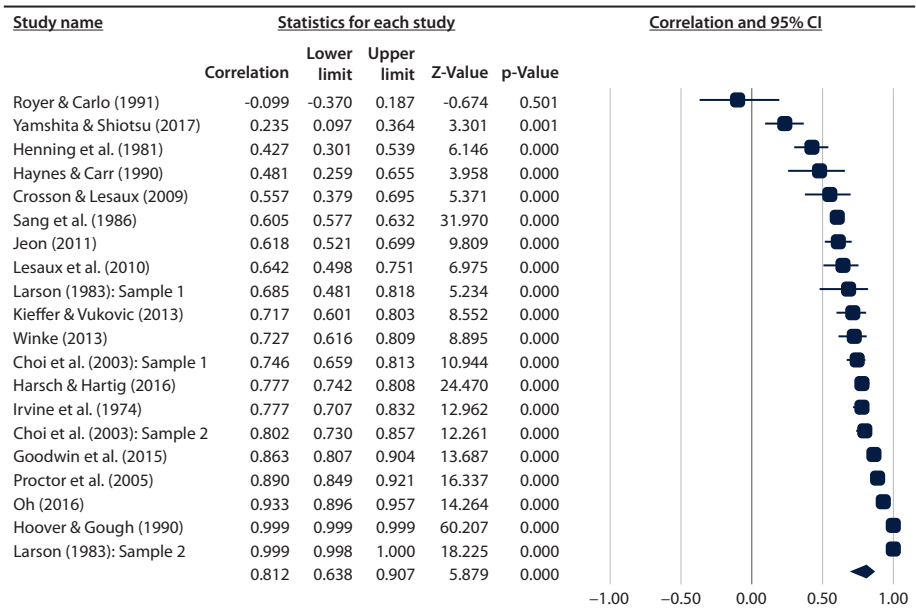
Table 6. Results of moderator analysis for L1 reading comprehension

Moderator variable	k	r [95% CI]	Significant test of difference (Q test)
Language difference			
II	12	.626 [.403-.778]	
IN	22	.369 [.296-.439]	4.571* (p = .033)

5.1.6 L2 listening comprehension

Twenty weighted and corrected correlations from 18 studies involving 4,700 participants were included in the analysis (mean sample size = 235, SD = 449.68, range = 26–2,083). Correlations, their significance, 95% CI, and their graphic representations are presented in Figure 6. Participants’ age ranged from elementary school to postgraduate levels and their L1s were diverse: Spanish (7 samples), Korean (4 samples), English (3 samples), German (2 samples), Chinese (1 sample),

Arabic (1 sample), Farsi (1 sample), Japanese (1 sample). As for the target language, 17 samples involved English and the remaining three involved French, German, and Chinese, respectively. The overall mean correlation was large,  $r = .812$ , 95% CI [.638–.907] (Cohen, 1988) and statistically significant ( $p = .00$ ). The variability across studies was significant and large,  $Q(19) = 2664.370$ ,  $p = .00$ ,  $I^2 = 99.287$ . As a result, a series of moderator analyses on age, L1-L2 script distance, L1-L2 language difference, and language setting were conducted in order to identify the source of this variability. Results showed that script distance was the only statistically significant moderator variable. The moderator analysis results are summarized in Table 7.



**Figure 6.** Average correlation (the diamond at the bottom) and correlation with confidence interval for each study correlating L2 listening comprehension and L2 reading comprehension

**Table 7.** Results of moderator analysis for listening comprehension

Moderator variable	<i>k</i>	<i>r</i> [95% CI]	Significant test of difference (Q test)
Script distance			
AA	17	.844 [.669–.930]	
AN (AM and AL)	3	.507 [.122–.759]	4.739* ( $p = .029$ )

5.1.7 Working memory

Nineteen weighted and corrected correlations from 17 studies involving 1,495 participants were included in the analysis (mean sample size = 78.68,  $SD = 99.76$ , range = 13–471). Correlations, their significance, 95% CI, and their graphic representations are presented in Figure 7. Participants' age ranged from elementary school to postgraduate levels and their L1s were diverse: English (6 samples), Chinese (2 samples), Spanish (2 samples), Turkish (3 samples), Cantonese (1 sample), Portuguese (1 sample), Japanese (1 sample), Farsi (1 sample), Korean (1 sample), and French (1 sample). As for the target languages, English was the most frequently studied L2, represented in 12 samples, with Spanish (2 samples), German (1 sample), Hebrew (1 sample), Korean (1 sample), and Chinese (1 sample) in the minority. The overall mean correlation was in the small to medium range,  $r = .334$ , 95% CI [.234–.427] (Cohen, 1988) and statistically significant ( $p = .00$ ). The variability across studies was significant and large,  $Q(18) = 63.025$ ,  $p = .00$ ,  $I^2 = 71.440$ . As a result, a series of moderator analyses on age, L1-L2 script distance, L1-L2 language difference, language setting, and testing language were conducted in order to identify the source of this variability. Since two independent samples involved a nonlinguistic working memory measure, they were excluded from the analysis, leaving 17 effect sizes. Results showed that language setting and testing language were statistically significant moderator variables. The moderator analysis results are summarized in Table 8.

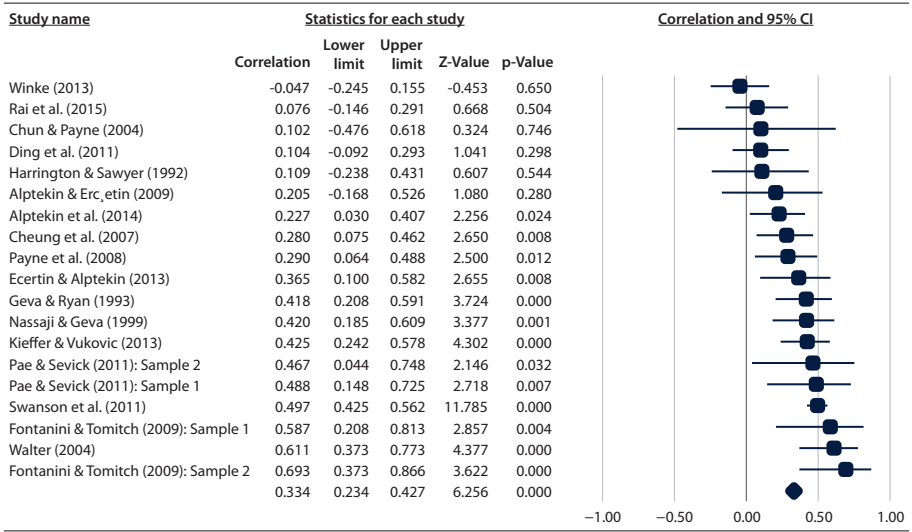


Figure 7. Average correlation (the diamond at the bottom) and correlation with confidence interval for each study correlating working memory and L2 reading comprehension

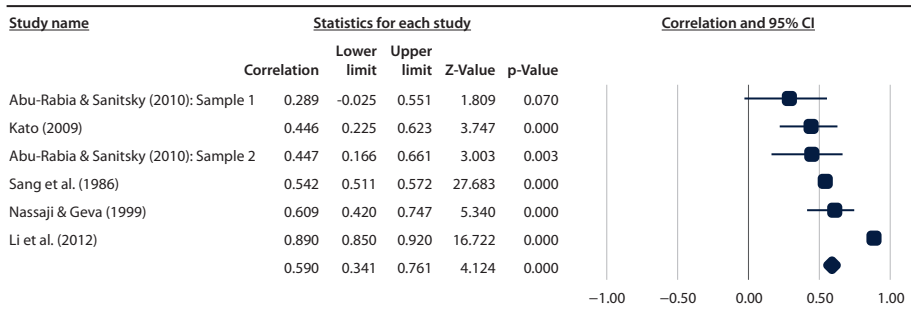
**Table 8.** Results of moderator analysis for working memory

Moderator variable	<i>k</i>	<i>r</i> [95% CI]	Significant test of difference ( <i>Q</i> test)
Language setting			
FL	15	.275 [.170–.373]	13.822** ( <i>p</i> = .000)
SL	4	.487 [.425–.544]	
Testing language			
L1	5	.203 [.041–.355]	6.208* ( <i>p</i> = .013)
L2	12	.434 [.332–.526]	

## 5.2 Results on low-evidence correlates

### 5.2.1 Orthographic knowledge

Six weighted and corrected correlations from five studies involving 2,430 participants were included in this analysis (mean sample size = 405., *SD* = 82.88, range = 40–2,083). Correlations, their significance, 95% CI, and their graphic representations are presented in Figure 8. Participants' age ranged from grade 6 to post-secondary levels and their L1s were diverse: Chinese (1 sample), Hebrew (1 sample), Russian (1 sample), Japanese (1 sample), Farsi (1 sample), and German (1 sample). As for the target language, English was again the most frequently studied L2 (5 samples), with Hebrew (1 sample) in the minority. The overall mean correlation fell in the medium to large range,  $r = .590$ , 95% CI [.341–.761] (Cohen, 1988) and was statistically significant ( $p = .00$ ).



**Figure 8.** Average correlation (the diamond at the bottom) and correlation with confidence interval for each study correlating orthographic knowledge and L2 reading comprehension

5.2.2 Morphological knowledge

Fourteen weighted and corrected correlations from 11 studies involving 1,403 participants were included in the analysis (mean sample size = 100.21, *SD* = 63.16, range = 24–246). Correlations, their significance, 95% CI, and their graphic representations are presented in Figure 9. Participants’ age ranged from elementary school to college levels and their L1s were diverse: Chinese (7 samples), Spanish (4 samples), Russian (1 sample), Hebrew (1 sample), and Korean (1 sample). As for the target language, 13 samples involved English and only one sample involved Hebrew. The overall mean correlation fell in the medium to large range,  $r = .635$ , 95% CI [.556–.703] (Cohen, 1988) and was statistically significant ( $p = .00$ ).

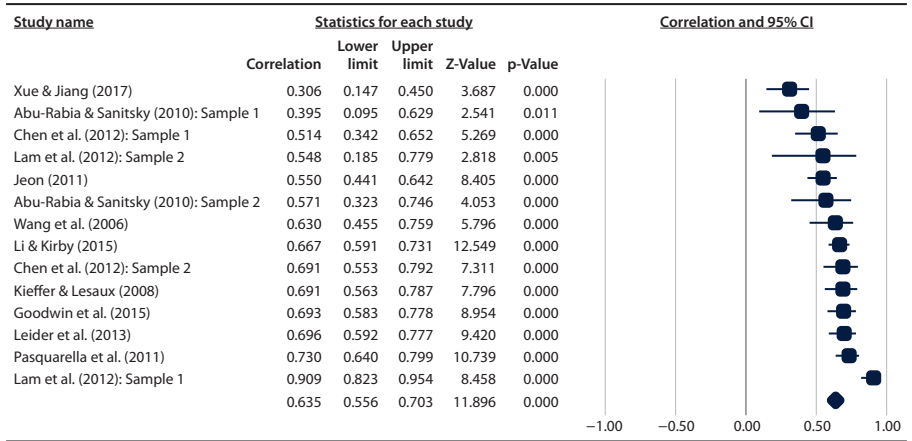
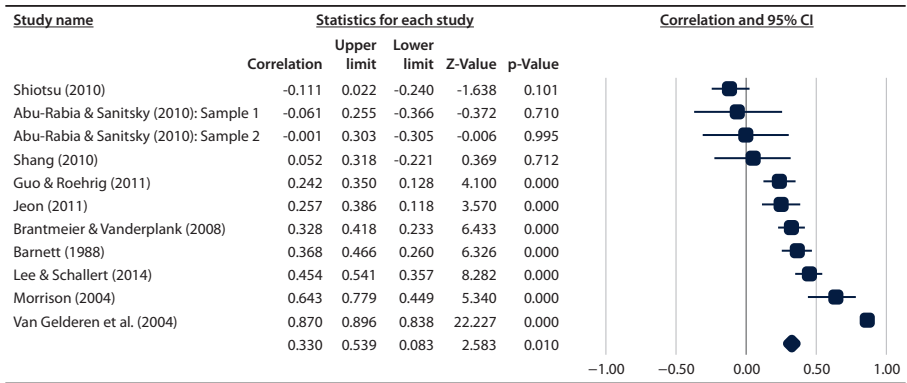


Figure 9. Average correlation (the diamond at the bottom) and correlation with confidence interval for each study correlating morphological knowledge and L2 reading comprehension

5.2.3 Metacognition

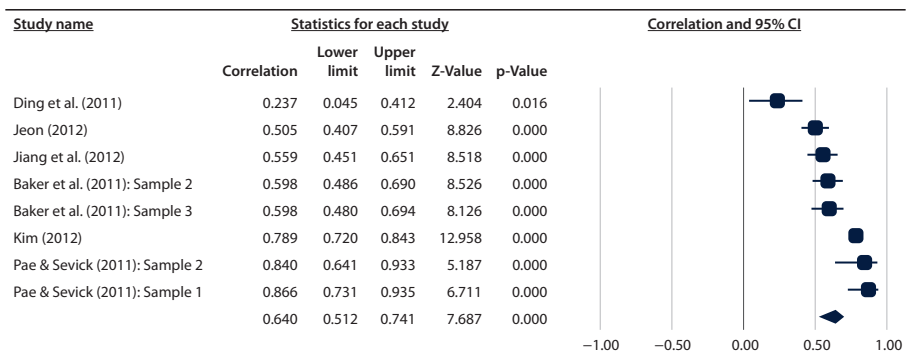
Eleven weighted and corrected correlations from 10 studies involving 2,073 participants were included in the analysis (mean sample size = 188.45, *SD* = 120.06, range = 40–359). Correlations, their significance, 95% CI, and their graphic representations are presented in Figure 10. Participants’ age ranged from grade 6 to college levels and their L1s were diverse: English (2 samples), Korean (2 samples), Chinese (2 samples), Russian (1 sample), Hebrew (1 sample), Japanese (1 sample), and Dutch (1 sample). English was the more frequently studied L2 (7 samples), with French (2 samples), Hebrew (1 sample), and Spanish (1 sample) in the minority. The overall mean correlation fell in the small to medium range,  $r = .330$ , 95% CI [.083–.539] (Cohen, 1988) and was statistically significant ( $p = .00$ ).



**Figure 10.** Average correlation (the diamond at the bottom) and correlation with confidence interval for each study correlating metacognition and L2 reading comprehension

#### 5.2.4 Oral reading fluency

Eight weighted and corrected correlations from 6 studies involving 2,430 participants were included in this analysis (mean sample size = 100,  $SD = 61.13$ , range = 21–156). Correlations, their significance, 95% CI, and their graphic representations are presented in Figure 11. Participants' age ranged from grade 6 to post-secondary levels and their L1s were diverse: Chinese (1 sample), Hebrew (1 sample), Russian (1 sample), Japanese (1 sample), Farsi (1 sample), and German (1 sample). As for the target language, English was again the most frequently studied L2 (5 samples), with Hebrew (1 sample) in the minority. The overall mean correlation fell in the medium to large range,  $r = .640$ , 95% CI [.512–.741] (Cohen, 1988) and was statistically significant ( $p = .00$ ).



**Figure 11.** Average correlation (the diamond at the bottom) and correlation with confidence interval for each study correlating oral reading fluency and L2 reading comprehension



6. Discussion

This study meta-analyzed the correlation coefficients between passage-level L2 reading comprehension and eleven linguistic, cognitive, and metacognitive variables related to L2 reading comprehension. In this section, we first discuss the overall results of these variables, comparing relative size of correlations found in the 2014 study and the present one. Then, we discuss each variable separately. For the seven high-evidence variables, we also discuss the moderator analysis results.

For ease of comparison, key results of the 2014 study and the present study are summarized in Table 9. Perhaps, the most notable, sustained finding is the role of L2 knowledge in L2 reading comprehension. With 8 and 20 studies added to the 2014 study pool, respectively, L2 grammar knowledge and L2 vocabulary knowledge once again emerged as the strongest correlates of L2 reading comprehension. Although the correlation for L2 grammar knowledge was slightly higher ( $r = .790$ ) than for L2 vocabulary knowledge ( $r = .724$ ), this result should not be overemphasized because their 95% CIs completely overlapped, a finding that resonates with the 2014 study. What remains undoubtedly clear is that these two core L2 knowledge variables are at the heart of abilities required for L2 reading comprehension. The overwhelming importance of L2 knowledge/skills variables in L2 reading is

Table 9. Results of 2014 and 2020 meta-analyses

2014 Study	<i>k</i>	<i>r</i> [95% CI]	2020 Study	<i>k</i>	<i>r</i> [95% CI]
<b>Decoding</b>	20	.56 [.46–.65]	<b>Decoding</b>	29	.586 [.453–.694]
Phonological awareness	11	.48 [.39–.57]	<b>Phonological awareness</b>	20	.611 [.515–.693]
<b>Vocabulary knowledge</b>	31	.79 [.69–.86]	<b>Vocabulary knowledge</b>	51	.724 [.636–.794]
<b>Grammar knowledge</b>	18	.85 [.58–.95]	<b>Grammar knowledge</b>	26	.697 [.517–.818]
<b>L1 reading comprehension</b>	22	.50 [.30–.66]	<b>L1 reading comprehension</b>	34	.483 [.361–.589]
L2 listening comprehension	14	.77 [.58–.88]	<b>L2 listening comprehension</b>	20	.812 [.638–.907]
Working memory	10	.42 [.29–.53]	<b>Working memory</b>	19	.334 [.234–.427]
Orthographic knowledge	5	.51 [.43–.58]	Orthographic knowledge	6	.590 [.341–.761]
Morphological knowledge	6	.61 [.52–.69]	Morphological knowledge	14	.635 [.556–.703]
Metacognition	10	.32 [.03–.55]	Metacognition	11	.330 [.083–.539]
			Oral reading fluency	8	.640 [.521–.741]

Note. Variables in boldface are high-evidence variables.

further supported by the correlations yielded by other L2 knowledge variables (ordered by the size of correlation): oral reading fluency ( $r = .640$ ), morphological knowledge ( $r = .635$ ), phonological awareness ( $r = .611$ ), orthographic knowledge ( $r = .590$ ), and decoding ( $r = .586$ ), all of which were in the medium to large range (Cohen, 1988). The two linguistic comprehension variables showed either a large or a small-to-medium size correlation: listening comprehension ( $r = .812$ ) and L1 reading comprehension ( $r = .483$ ). Finally, the language-general, cognitive and metacognitive variables showed medium correlations: working memory ( $r = .318$ ) and metacognition ( $r = .330$ ).

## 6.1 Decoding and L2 reading comprehension

Decoding demonstrated a strong association with L2 reading comprehension ( $r = .586$ ). From the moderator analyses conducted to find the source of variability among the samples, only language setting reached a statistically significant level (Table 3); studies in the SL settings produced a significantly higher correlation ( $r = .622$ ) than in the FL settings ( $r = .424$ ).

However, this result is not easy to interpret. Theoretically, it is assumed that L2 reading is more highly correlated with lower-level skills like decoding among younger readers compared to older readers. Given this, we had anticipated that age, rather than language setting, would be the significant moderator. Age itself, however, despite such a tendency (Child:  $r = .624$ , Adolescent/Adult:  $r = .451$ ), failed to reach statistical significance ( $p = .08$ ). In our samples, over 80% (19 out of 23) study participants of the SL group were younger readers in a minority language setting, whereas over 80% (five out of six) study participants of the FL group were older readers (college age or older). We thus conjecture that, with the caution and the endorsement of the need of further examination, it is possible that the effect of language setting was confounded with that of age. Had we used a different categorization for age and create a group of younger (than grade 6) children, the age effect may have shown to be more pronounced.

## 6.2 Phonological awareness and L2 reading comprehension

With the increase of nine effect sizes from the 2014 to 2020 study, phonological awareness kept its strong role in L2 reading comprehension. Despite the indication of heterogeneity, none of our moderator analyses reached statistical significance. While we do not know the reason for this for sure, some nonsignificant moderator analysis results are worth discussing.

In the pool of our studies, participants' age was severely skewed to the child population with only two samples examining adults and the rest 18 children. This focus on the younger readers may not be surprising because lower-level processes including phonological awareness are more likely to be significant predictors of individual differences in reading comprehension in children than adults. Generally, with the gradual mastery of lower-level processes, higher-level processes become stronger predictors of reading comprehension in adults or more advanced readers (Geva & Ryan, 2005; Gough & Tunmer, 1986). The large difference in the sample sizes of the age groups (18 vs. 2) being compared may have weakened the power of the moderator analysis on age ( $p = .571$ ), but at least the descriptive tendency supported the stronger association among children than adults ( $r = .616$  vs.  $r = .546$ ). Although the imbalance found in the study population may call for more studies with adult participants, future studies might benefit by allowing more nuanced approach than testing direct contributions of phonological awareness to reading comprehension (e.g., Yoshikawa & Yamashita, 2014). Such research would also enable an investigation of the contribution of phonological awareness in adult L2 readers of varying proficiency levels.

Considering that phonological awareness is a pre-literate foundational skill that does not involve any reading (either written symbols or syntax), nonsignificant effects of the language distance or the script distance might not be too surprising. Unlike age, there was not a potential threat of severe sample imbalance: language distance (II vs IN = 9 vs. 11), script distance (AA vs AN = 13 vs. 7). As such, we believe the nonsignificant results can be taken at face value without further explanation.

### 6.3 Vocabulary knowledge and L2 reading comprehension

As mentioned above, vocabulary maintained its strong association with L2 reading comprehension with the increase of 20 effect sizes. Of all moderators, only script distance was found to be marginally significant. The strength of associations was ordered in size from AM (alphabetic script – mixed script [i.e., Japanese] pairing), AA (alphabetic script – alphabetic script pairing), and AL (alphabetic script – logographic script pairing):  $r = .827$  [95% CI: .038–.961],  $r = .748$  [95% CI: .640–.827], and  $r = .599$  [95% CI: .512–.674]). While it is unclear as to why vocabulary knowledge showed the highest correlation with L2 reading comprehension in the AM subgroup and the lowest correlation in the AL subgroup, it might be useful to note that AM subgroup participants were exclusively Japanese L1 readers while AL subgroup participants were exclusively Chinese L1 readers. The AA subgroup, on the other hand, featured diverse L1 and L2 pairings. Given this, we conjecture that

some underlying factors above and beyond script distance between L1 and L2 might have been at play such as different degrees of individual differences in vocabulary knowledge as a function of their L1 membership.

Two moderator effects that were significant in the 2014 study turned out to be nonsignificant in the 2020 study. One is the effect of age: the correlation in the Adolescent/Adult group was higher than the Child group ( $r = .84$  vs.  $r = .66$ ) in the 2014 study, but this effect was no longer significant in the present analysis ( $r = .746$  vs.  $r = .698$ ). Previously, we argued that, with the Adolescent/Adult samples being mostly college-level or postgraduate students, vocabulary tests for this group were more likely to be heavily inclusive of lower-frequency, academic words than those of younger counterparts. As a result, individual differences in vocabulary knowledge may have been better captured among the Adolescent/Adult group (in contrast to the high-frequency, everyday vocabulary that are likely to be tested for younger children), which then may have led to the higher correlation with reading comprehension in the senior readers. In the current pool of enlarged samples that offered correlations of vocabulary, the qualitative nature of the Adolescent/Adult group was consistent (mostly college-level or postgraduate students), but newly added studies expanded the age range of the Child group to younger, lower elementary grade readers (e.g., Jared et al. 2011; Lam et al., 2012; Siu & Ho, 2015; Swanson et al., 2011). We conjecture that the individual differences in vocabulary knowledge in high-frequency, everyday words are better captured in the current sample than in the 2014 sample (i.e., individual differences in the knowledge of such vocabulary are likely to be larger among younger than older children). Therefore, the expansion of the age range in the 2020 study may have led to the current findings. The nonsignificant effect of age indicates the universal importance of vocabulary for reading comprehension regardless of age.

Another difference from the 2014 findings concerns the moderator effects of contextualized vs. decontextualized test item types. Previously, we found that the correlation between vocabulary knowledge and L2 reading comprehension was significantly higher for contextualized item tests (the test item presents the target word embedded in a phrase or a sentence) than for decontextualized item tests (the test item presents the target word without any context):  $r = .92$  vs.  $r = .71$ . We conjectured that this was because the ability to respond to contextualized vocabulary test items overlaps with L2 reading comprehension to a higher degree. In the present study, the general trend of the correlations was sustained (i.e., higher correlation for the contextualized test item subgroup ( $r = .867$ ) than the decontextualized test item subgroup ( $r = .689$ )). However, this difference was not statistically significant. The closing of the gap between contextualized vs. decontextualized vocabulary tests found in the present study may be attributed to the types of decontextualized

vocabulary tests used in the studies added to the pool; while most studies involving children almost invariably used the Peabody Picture Vocabulary Test or its abridged form, studies involving adults featured a range of different tests focusing not only on the size but also depth (through testing word association knowledge) of vocabulary knowledge. One study (Harsch & Hartig, 2016) also used a unique vocabulary test on which the test taker had to discern real words from nonwords. If the closer gap between contextualized and decontextualized vocabulary test does indicate improved potential for decontextualized vocabulary tests, this is good news because these tests could minimize the construct overlap between vocabulary knowledge and reading comprehension and be administered in shorter test times if other conditions are identical.

#### 6.4 Grammar knowledge and L2 reading comprehension

To reiterate, grammar knowledge demonstrated one of the strongest correlations with L2 reading comprehension ( $r = .697$ ) along with vocabulary. Script distance and one measurement characteristic (sentence completion test, GJT, vs. other test types) were found to be statistically significant moderator variables.

The significant effect of script distance is a new finding in this study compared to the 2014 study. The result showed an interesting pattern which partially resonates with the moderator analysis on vocabulary: the AL group (i.e., alphabetic- and logographic-script pairing, which was exclusively L1 Chinese) showed the lowest correlation between L2 grammatical knowledge and L2 reading comprehension. Although this time, the correlation for the AA group (alphabetic script-alphabetic script pairing) was slightly higher than that of the AM group (alphabetic script – mixed script pairing, which was exclusively Japanese), the correlations of the AA and AM groups were much higher than that of the AL group. Whether this result indicates a qualitative difference that sets apart the Chinese L1 group from the other L1 groups, and if so, what causes such a difference is unclear. Nevertheless, it is interesting to note that the L1 Chinese group sustained the same pattern with two key variables representing L2 knowledge (L2 vocabulary and grammar knowledge) and this finding warrants further investigation.

The results from the measurement characteristic (sentence completion test, GJT, and other test types) replicated the 2014 finding. The correlations were ordered in size from sentence completion test ( $r = .793$ ), other test types ( $r = .562$ ), and GJT ( $r = .367$ ). The sentence completion test, on which the test taker completes a sentence with a blank by selecting an option or by writing in the blank, was the most common test type (16 of 26 samples), while GJT and other test types (e.g., word order rearrangement, sentence combining) were used in five samples, respectively. This tendency of preferred test types was also similar between the 2014 and

the present study. In the 2014 study, based on the observation that the GJTs used in the L2 reading research were mostly untimed, we maintained that the weaker correlation of the GJT may be due to the limited scope of grammar knowledge (i.e., explicit knowledge) that the timed GJT taps into. However, as indicated in the present study, studies are increasingly adopting both timed and untimed GJTs, aiming to capture both explicit and implicit knowledge of grammar. Considering the expansion of the type of GJT and its long-term contribution in the SLA research, the continued underwhelming correlation of GJT is puzzling. We speculate that the higher likelihood of accidentally guessing the correct answer in the dichotomous judgment (grammatical or ungrammatical) may be a potential threat to GJT (Shiotsu, 2010). In contrast, the continued high correlation of the completion test underscores again its usefulness in assessing the grammar knowledge that underlies L2 reading comprehension.

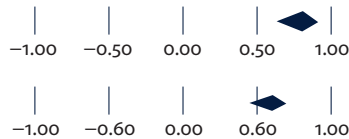
## 6.5 L1 reading comprehension and L2 reading comprehension

L1 reading comprehension resulted in medium-size overall correlations both in the 2014 study ( $r = .50$ ) and the present study ( $r = .483$ ). However, the significant moderating effect of the language distance between L1 and L2 (II: Indo-European and Indo-European pairing vs. IN: Indo-European vs. Non-Indo-European pairing) indicates that we should not take this overall correlation at face value. Both in the 2014 and 2020 study, the II group produced a significantly higher overall correlation than the IN group ( $r = .60$  vs.  $r = .36$  in 2014;  $r = .626$  vs.  $r = .369$  in 2020). The correlation of the II group is large, whereas that of the IN group is medium. Given the reasonably large numbers of samples in both categories (II vs. IN = 12 vs. 22), the consistent results across the 2014 and the present study strongly support that the transfer between L1 and L2 reading comprehension abilities are more likely when the involved L1 and L2 are linguistically more similar than different. These results are in line with key theories such as the Linguistic Interdependence Hypothesis (Cummins, 1979, 1991), Linguistic Threshold Hypothesis (Cummins, 1976) as well as the more current research on L1 and L2 reading transfer (e.g., Koda, 2008).

## 6.6 L2 Listening comprehension and L2 reading comprehension

In the 2014 study, L2 listening comprehension was highly correlated with L2 reading comprehension ( $r = .77$ ), a finding which provided support for the Simple View of Reading (Gough & Tunmer, 1986) as a viable model of L2 reading. With six additional effect sizes, this trend became even stronger in the present study ( $r = .812$ ). The strong association between L2 listening and reading comprehension is likely

due to the important role L2 knowledge plays in L2 reading comprehension as indicated by the consistently strong correlations between L2 knowledge variables and reading comprehension found in the present study. Although some subconstructs of L2 reading comprehension such as decoding or orthographic knowledge are unique to reading, L2 listening comprehension and reading comprehension share many central L2 knowledge variables such as morphology, vocabulary, and grammar. However, perhaps the more interesting finding of the present study regarding the Simple View of Reading is the relative importance of decoding and listening comprehension in L2 reading comprehension. The present study found that although both decoding and L2 listening comprehension were strong correlates of L2 reading comprehension ( $r = .586$  for decoding and  $r = .812$  for L2 listening comprehension), their 95% CIs showed a moderate overlap before adjustment (no overlap after trim and fill). Together, these results indicate that L2 listening comprehension was a significantly stronger correlate than decoding to reading comprehension. For a visual representation of this data see Figure 12 where the width of the diamond indicates the range of 95% CI of L2 listening comprehension and decoding before trim and fill adjustment.



**Figure 12.** Overall correlations with 95% CIs of listening comprehension (above) and decoding (below)

The present study's results resonate with the findings of Sparks and Luebbers (2018); in this study, the authors classified poor L2 readers (L2-English L1-Spanish students in the US) based on the framework of the Simple View of Reading: (1) dyslexic (poor decoding, good comprehension), (2) hyperlexic (good decoding, poor language comprehension), and (3) garden variety (poor decoding, poor comprehension). The authors reported that poor L2 readers fit either the hyperlexic or the garden variety profile. No dyslexic readers were identified in this study. Taken together, the results of the present study and of Sparks and Luebbers (2018) underscore the importance of addressing skills related to language comprehension in L2 instruction, including L2 linguistic knowledge. Although it is not our intention to downplay the importance of decoding, the relative contribution of decoding and listening comprehension changes developmentally. Generally, decoding is more critical in younger/novice readers than older/more skilled readers since decoding is a simpler skill than comprehension and is mastered in earlier stages



of development. As the samples included in this meta-analysis involved a wide range of L2 readers, the higher correlation of listening comprehension may point to the longer-lasting contribution of this skill to explaining individual differences in reading comprehension.

In the moderator analysis, only the script distance resulted in a significant difference between the AA (Alphabetic- and Alphabetic script pairing) and AN (Alphabetic- and Non-Alphabetic script pairing) subgroups ( $r = .844$  vs.  $r = .507$ ). However, this result should be interpreted with caution because there were only three samples in the AN group with not much diversity in L1 (2 samples with Chinese L1 and 1 sample with Japanese L1). Notwithstanding this caution, other theoretically more plausible moderators help interpreting this result. Although not statistically significant, language distance and language setting displayed the expected tendency. The correlation was stronger in the II group ( $k = 12$ ) than in the IN group ( $k = 8$ ) ( $r = .872$  vs  $r = .677$ ) ( $p = .118$ ), and in the SL group ( $k = 8$ ) than in the FL group ( $k = 12$ ) ( $r = .862$  vs  $r = .755$ ) ( $p = .541$ ). The AN group in the script distance analysis belonged to the IN group or FL group. Thus, we may have a better understanding of the script distance effect on the listening comprehension by considering interactions with other potential moderator variables.

## 6.7 Working memory and L2 reading comprehension

Working memory, which saw a substantial increase of nine effect sizes from the 2014 pool of ten, resulted in an overall correlation of .334. This result is comparable to other meta-analyses investigating the relationship between working memory and L1 or L2 performance. Daneman and Merikle (1996) meta-analyzed correlations between working memory and comprehension ability (operationalized as reading comprehension and vocabulary knowledge) of L1 readers. The resulting effect sizes ranged from .14 to .52, depending on which comprehension and working memory measures were used. Linck et al. (2014), another meta-analysis which synthesized correlations between working memory and various L2 performance outcomes (both comprehension and production) of adult L2 learners, reported an overall correlation of .255. Similarly, Shin's (2020) meta-analysis of correlations between working memory measured through reading span tasks and L2 reading comprehension reported an overall correlation of .30. In sum, meta-analytic findings show that working memory is an important correlate of L2 reading and language performance in general, but that its correlation is consistently in the small to medium range.

One possible explanation for this modest effect size may have to do with the inherent nature of working memory; working memory is a complex construct with its own sub-components. Accordingly, the working memory measures featured in



the primary studies included in the present meta-analysis are also diverse; some measures focus on storage only while others assess both storage and processing; some measures involve verbal memory while others are language-free (e.g., visual); some measures are administered in the study participants' L1 while others are given in L2. As expected, working memory scores yielded by different measures have shown to correlate with reading comprehension to different degrees. For instance, reading span tasks that tap into both storage and processing functions of working memory tended to show higher correlations than simple storage tasks (e.g., digit span tasks) (Daneman & Carpenter, 1980; Daneman & Merikle, 1996). Similarly, verbal tasks have produced higher correlations than with non-verbal tasks (Daneman & Merikle, 1996).

Motivated by relevant literature, effects of language of the working memory test (L1 vs. L2) were examined using a moderator analysis and was found to be statistically significant; the correlation between working memory and L2 reading comprehension was significantly higher for the L2 working memory test group ( $r = .434$ ) than that of the L1 working memory test group ( $r = .203$ ). This result, which replicates Linck et al.'s (2014) finding, indicates that although working memory is not a language-specific construct, the language in which the test is administered can indeed exert a significant effect on the test outcome as a function of the test taker's language proficiency.

Further, Alptekin and Ercetin (2009) have also posited how L2 proficiency can complicate the relationship between working memory and L2 reading comprehension. The researchers noted that L2 learners with limited L2 proficiency tend to overly rely on literal comprehension without engaging inferential comprehension. This tendency inhibits working memory from being deployed to achieve reading comprehension. If this were true, individual differences in working memory may not necessarily predict individual differences in reading comprehension. From the potential influences of L2 learners' reading processes or the effect of testing language, we conjecture that working memory, much like other higher-level comprehension processes, is also subject to the Linguistic Threshold Hypothesis (Cummins, 1976). Careful considerations of the construct validity of working memory measures as well as the role of L2 proficiency will bring about further insights into the relationship between working memory and L2 reading comprehension.

## 6.8 Orthographic knowledge and L2 reading comprehension

As one of key components of L2 knowledge necessary for reading, orthographic knowledge once again showed a medium to large correlation ( $r = .590$ ) with L2 reading comprehension in this study. Despite its theoretical and empirical importance,

orthographic knowledge has only seen an increase of one study from 2014. One possible reason may be that most of the samples (83%) included in the present study have English as the target language, for which the importance of phonological processing is far more emphasized than orthographic processing especially in the earlier stages of reading development. Had languages with a logographic writing system such as Chinese been dominant in the target language, on the other hand, orthographic processing may have received more attention. As L2 reading research gradually expands its target languages beyond English, we may see more interest in orthographic knowledge in the future. It would be interesting to examine whether different L2s moderate the relationship between orthographic knowledge and reading comprehension.

On another note, Chen et al., (2020) argued that the lack of examination of orthographic processing is partially due to the lack of comprehensive and accurate measures of orthographic knowledge. Indeed, despite the well-accepted theoretical distinction, researchers are aware of difficulties of clearly separating orthographic processing from phonological processing. However, efforts that are being made to develop more comprehensive and purer measures of orthographic knowledge (Chen et al., 2020) may contribute to the increase of research on this component.

## 6.9 Morphological knowledge and L2 reading comprehension

Although it failed to qualify as a high-evidence variable in this study by missing one sample, morphological knowledge is a rapidly growing area as can be seen in the number of samples which more than doubled from 2014 to 2020 (6 to 14). Morphological knowledge is expected to be critical for reading comprehension because it relates to many facets of the language (lexical semantics, syntax, phonology) and assists text processing by solving potential difficulties during reading. For example, morphological knowledge may mediate lexical inferencing, it may support syntactic parsing by signaling words' syntactic functions, or it may improve decoding efficiency of morphologically complex words. As expected, morphological knowledge continued to demonstrate a large effect size ( $r = .635$ ) in the present study.

Among the three systems of morphology, namely, inflection, derivation, and compounding, derivational morphology was most commonly investigated in the primary studies included in this study. This may be due to the popularity of English as the target language (13 out of 14 samples); English features a complex derivational system compared to its other morphological systems that are relatively simpler and acquired before derivations. Morphological knowledge is supposed to increase its importance in later stages of reading development (after middle

grades in primary school through adulthood) because of the increasing proportions of morphologically complex words in the texts that more advanced readers read. Probably resonating with this fact, most participants in our samples were grade three and above, although a few younger groups were also represented.

With the common distinction of the three main systems of morphology (inflection, derivation, and compounding), the assessment of morphological knowledge may seem straightforward. However, a closer inspection of the primary studies included in the present meta-analysis revealed that there were substantial variations in its measurement methods (cf. Ke et al., 2021). With the exception of some common tasks such as Carlisle's (2000) derivation and decomposition tests, measurement varied on many dimensions; target morphemes were provided with or without context, tests used real- or pseudo-word items, adopted productive vs. recognition/multiple-choice tasks, and assessed different aspects of morphological knowledge such as form vs. meaning vs. parts of speech, or explicit vs. implicit knowledge. Given the continued efforts to improve the measurements of morphological knowledge (Bernstein et al., 2020; Mizumoto et al., 2019) and the increase in the number of relevant studies, we anticipate more fine-grained findings on the contribution of morphological knowledge to L2 reading comprehension in the near future.

## 6.10 Metacognition and L2 reading comprehension

Another language-general variable, metacognition, only saw an increase of one effect size from the 2014 study's pool of ten, and as expected, the size of overall correlation did not change much:  $r = .32$  (the 2014 study) to  $r = .330$  (the 2020 study). However, we would like to note that the little increase in the study pool does not necessarily mean a lack of interest in this variable among L2 reading researchers. It simply indicates that among the many studies that may have involved metacognition within the context of L2 reading, there was only one study that also met our inclusion criteria.

As for the generally lower levels of overall correlations, we offer a few conjectures as follows. First, the majority of studies included in the present meta-analysis assessed reading-related metacognition using a type of self-report measure (e.g., reading strategy questionnaire, reading-related metacognitive awareness questionnaire). While such measures are clearly less invasive than online, think-aloud measures which can disrupt normal reading processes, there is a potential concern for construct validity. Much like how highly proficient and less proficient L2 readers qualitatively differ in their use of working memory, it is also possible that they differ in their use of metacognitive resources. For example, among highly proficient L2 readers, certain metacognitive processes such as comprehension monitoring or

use of strategies may be automatized and, therefore, not necessarily in the realm of consciousness. Less-proficient L2 readers, on the other hand, may have explicit knowledge of comprehension monitoring as well as various reading strategies, but they may not necessarily be able to deploy them as they read. Also, less proficient L2 readers, due to their usually lower metacognitive awareness, may report that they use appropriate strategies, when in fact, they do not. This is hardly an implausible scenario as less-proficient L2 readers are often inaccurate about their own reading performance (Morrison, 2004). Second, the problem of measurement also relates to the complexity of the construct of metacognition. What was targeted at under the name of metacognition substantially varied in the primary studies, including reading strategies (sometimes with writing and learning strategies as well), self-assessment, self-efficacy, cognitive attitudes, comprehension monitoring, and knowledge about text in general. Scoring methods were also different across studies. Many used a Likert scale, but some used a dichotomous scoring like a test. This scoring difference seems to reflect different conceptualizations of metacognition among researchers. The inconsistencies in the construct and the limitations in its operationalization limit our understanding of how declarative knowledge of metacognition such as reading strategies are used along with other reading skills that are subject to automatization. The fact that three out of 11 correlations between metacognition and reading comprehension were in the negative seem to be consistent with these problems. Since metacognition is no doubt useful for self-regulating reading processes, we should make efforts to answer fundamental questions pertinent to the definition and measurement of metacognition to improve our understanding of the relationship between metacognition and reading comprehension.

### 6.11 Oral reading fluency and L2 reading comprehension

As noted earlier, oral reading fluency is a relatively new variable under investigation in L2 reading research, and as a result, relevant empirical evidence is only just starting to emerge from a limited number of recent studies (6 studies yielding 8 independent samples). Upon reviewing the primary studies included here, a few trends were noted. First, possibly reflecting the emphasis on oral reading fluency among younger, novice readers, most samples involved children (Grades 1 through 4) while only two samples involved adolescents (Jeon, 2012) and adults (Jiang et al., 2012). For younger readers, standardized oral reading fluency tests normed on monolingual children (e.g., DIBELS Oral Reading Fluency subtest, the fluency subtest of Gray Oral Reading Test) were often used while studies involving older readers tended to use researcher-made tests. The results showed a strong overall correlation ( $r = .640$ ) between passage-level oral reading fluency and L2

reading comprehension. Most notably, the two independent correlations yielded by the studies involving older L2 readers were also in the large range:  $r = .505$  for Jeon (2012) and  $r = .559$  for Jiang et al. (2012). Taken together with Sparks and Luebber's (2018) findings that showed comprehension without recoding is an unlikely phenomenon among high school level L2 readers, the present study's findings indicate that passage-level oral reading fluency is indeed an important correlate and even a proxy of L2 reading comprehension among both young and older readers. One explanation for the strong relationship between passage-level oral reading fluency and L2 reading comprehension is found in one of the primary studies included in the present analysis; Jeon (2012) reported from her exploratory factor analysis which included a range of reading subconstructs and correlates (i.e., word-level oral reading fluency using pseudowords and real words, passage-level oral reading fluency, morphological awareness, word knowledge, grammar knowledge, reading-related metacognitive awareness, listening comprehension), passage-level oral reading fluency cross-loaded with both the oral reading fluency factor and L2 comprehension factor (whose indicators were L2 knowledge variables and L2 reading and listening comprehension). Word-level oral reading fluency, on the other hand, did not cross-load with the comprehension factor. In other words, unlike word-level oral reading fluency, passage-level oral reading fluency is strongly tied with constructs related to L2 knowledge and L2 comprehension in general. The strong overall correlation found between passage-level oral reading fluency and L2 reading comprehension across young and older L2 readers is likely a manifestation of this underlying relationship.

## 7. Conclusion

The present study updated our 2014 meta-analysis of the correlation between L2 reading comprehension and its correlates. In six years between 2011 (the literature search end in the 2014 study) and 2017 (that in the 2020 study), with the same inclusion criteria, 40 new independent samples were added (8 were from a new variable, oral reading fluency). Although the number of added samples varies greatly across correlates, the sizable addition of eligible studies overall signals a robust interest in component approaches in L2 reading research. Both consistent and new findings were noted across the 2014 and 2020 study. Of them, perhaps the most important, sustained finding is the strong relationship of L2 linguistic knowledge (especially grammar and vocabulary) and L2 reading comprehension. Together with the strong correlations from other L2 knowledge/skills components, the results continually endorse the importance of L2 language ability to achieve higher levels of L2 reading comprehension. Another consistent finding is the moderating

effect of L1-L2 language distance on the correlation between L1 and L2 reading comprehension. In line with theoretical expectations, closer language pairings have shown to facilitate transfer of reading comprehension ability. In comparison with the strong and sustained effects of L2 knowledge variables, the modest correlations of language-general correlates (working memory and metacognition) also replicated the 2014 findings. As discussed above, both of these variables are complex and multi-faceted in their constructs, and there were substantial variations and potential limitations in the measurement methods across different studies. Therefore, we must take heed that the relatively lower correlations may partially be the result of inconsistencies in the constructs and operationalizations adopted by different primary studies. Lastly, a new variable, oral reading fluency, showed an overall correlation that fell in the medium to large range, and suggested a promising future of passage-level oral reading fluency as a component of, or even as a proxy of L2 reading comprehension.

## References

*Note.* Studies that were included in the meta-analysis are marked with an asterisk (\*).

- \*Abu-Rabia, S., & Sanitsky, E. (2010). Advantages of bilinguals over monolinguals in learning a third language. *Bilingual Research Journal*, 33(2), 173–199. <https://doi.org/10.1080/15235882.2010.502797>
- Alderson, J. C. (1984). Reading in a foreign language: A reading problem or a language problem? In J. C. Alderson & A. H. Urquhart (Eds.), *Reading in a foreign language* (pp. 1–24). Longman.
- \*Alptekin, C., & Ercetin, G. (2009). Assessing the relationship of working memory to L2 reading: Does the nature of comprehension process and reading span task make a difference? *System: An International Journal of Educational Technology and Applied Linguistics*, 37(4), 627–639. <https://doi.org/10.1016/j.system.2009.09.007>
- \*Alptekin, C., Özemir, O., & Erçetin, G. (2014). Effects of variations in reading span task design on the relationship between working memory capacity and second language reading. *The Modern Language Journal*, 98(2), 536–552. <https://doi.org/10.1111/modl.12089>
- \*August, G. C. (2006). So, what's behind adult English second language reading? *Bilingual Research Journal*, 30(2), 245–264. <https://doi.org/10.1080/15235882.2006.10162876>
- \*August, D., Francis, D. J., Hsu, H., & Snow, C. E. (2006). Assessing reading comprehension in bilinguals. *Elementary School Journal*, 107(2), 221–238. <https://doi.org/10.1086/510656>
- Baddeley, A. D., & Hitch, G. (1974). Working memory. *Psychology of Learning and Motivation*, 8, 47–89. [https://doi.org/10.1016/S0079-7421\(08\)60452-1](https://doi.org/10.1016/S0079-7421(08)60452-1)
- \*Baker, D. L., Park, Y., & Baker, S. K. (2011). The reading performance of English learners in grades 1–3: the role of initial status and growth on reading fluency in Spanish and English. *Reading and Writing*, 25(1), 251–281. <https://doi.org/10.1007/s11145-010-9261-z>
- \*Barnett, M. A. (1988). Reading through context: How real and perceived strategy use affects L2 comprehension. *The Modern Language Journal*, 72(2), 150–162. <https://doi.org/10.1111/j.1540-4781.1988.tb04177.x>

- \*Bensoussan, M., & Ramraz, R. (1984). Testing EFL reading comprehension using a multiple-choice rational cloze. *The Modern Language Journal*, 68(3), 230–239. <https://doi.org/10.1111/j.1540-4781.1984.tb01569.x>
- Bernhardt, E. B., & Kamil, M. L. (1995). Interpreting relationships between first language and second language reading: Consolidating the linguistic threshold and the linguistic interdependence hypotheses. *Applied Linguistics*, 16(1), 15–34. <https://doi.org/10.1093/applin/16.1.15>
- Bernstein, S. E., Flipse, J. L., Jin, Y., & Odegard, T. N. (2020). Word and sentence level tests of morphological awareness in reading. *Reading and Writing*, 33(6), 1591–1616. <https://doi.org/10.1007/s11145-020-10024-6>
- Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2005). *Comprehensive meta-analysis version 2* [software]. Biostat.
- \*Brantmeier, C., & Vanderplank, R. (2008). Descriptive and criterion-referenced self-assessment with L2 readers. *System: An International Journal of Educational Technology and Applied Linguistics*, 36(3), 456–477. <https://doi.org/10.1016/j.system.2008.03.001>
- \*Brisbois, J. E. (1995). Connections between first- and second-language reading. *Journal of Reading Behavior*, 27(4), 565–584. <https://doi.org/10.1080/10862969509547899>
- Carlisle, J. F. (2000). Awareness of the structure and meaning of morphologically complex words: Impact on reading. *Reading and Writing: An Interdisciplinary Journal*, 12(3–4), 169–190. <https://doi.org/10.1023/A:1008131926604>
- \*Carlisle, J. F., Beeman, M. M., Davis, L. H., & Spharim, G. (1999). Relationship of metalinguistic capabilities and reading achievement for children who are becoming bilingual. *Applied Psycholinguistics*, 20(4), 459–478. <https://doi.org/10.1017/S0142176499004014>
- \*Carson, J., Carrell, P. L., Silberstein, S., Kroll, B., & Kuehn, P. A. (1990). Reading-writing relationships in first and second language. *TESOL Quarterly*, 24(2), 245–266. <https://doi.org/10.2307/3586901>
- Carver, R. (1997). Reading for one second, one minute, or one year from the perspective of rauding theory. *Scientific Studies of Reading*, 1(1), 3–43. [https://doi.org/10.1207/s1532799xssr0101\\_2](https://doi.org/10.1207/s1532799xssr0101_2)
- \*Chen, X., Ramirez, G., Luo, Y. C., Geva, E., & Ku, Y.-M. (2012). Comparing vocabulary development in Spanish and Chinese-speaking ELLs: The effects of metalinguistic and sociocultural factors. *Reading and Writing*, 25(8), 1991–2020. <https://doi.org/10.1007/s11145-011-9318-7>
- Chen, Y.-J. I., Wilson, M., Irey, R. C., & Requa, M. K. (2020). An innovative measure of orthographic processing: Development and initial validation. *Language Testing*, 37(3), 435–452. <https://doi.org/10.1177/0265532220909310>
- \*Cheung, H., Chan, M., & Chong, K. (2007). Use of orthographic knowledge in reading by Chinese-English bi-scriptal children. *Language Learning*, 57(3), 469–505. <https://doi.org/10.1111/j.1467-9922.2007.00423.x>
- \*Choi, I., Kim, K., & Boo, J. (2003). Comparability of a paper-based language test and a computer-based language test. *Language Testing*, 20(3), 295–320. <https://doi.org/10.1191/0265532203lt2580a>
- \*Chun, D. M., & Payne, J. (2004). What makes students click: Working memory and look-up behavior. *System: An International Journal of Educational Technology and Applied Linguistics*, 32(4), 481–503. <https://doi.org/10.1016/j.system.2004.09.008>
- Clarke, M. (1979). Reading in Spanish and English: Evidence from adult ESL students. *Language Learning*, 29(1), 121–150. <https://doi.org/10.1111/j.1467-1770.1979.tb01055.x>
- Clarke, M. A. (1980). The short circuit hypothesis of ESL reading – or when language competence interferes with reading performance. *The Modern Language Journal*, 64(1), 203–209. <https://doi.org/10.1111/j.1540-4781.1980.tb05186.x>



- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates.
- \*Crosson, A. C., & Lesaux, N. K. (2010). Revisiting assumptions about the relationship of fluent reading to comprehension: Spanish-speakers' text-reading fluency in English. *Reading and Writing: An Interdisciplinary Journal*, 23(5), 475–494. <https://doi.org/10.1007/s11145-009-9168-8>
- Cummins, J. (1976). The influence of bilingualism on cognitive growth: A synthesis of research findings and explanatory hypotheses. *Working Papers on Bilingualism*, 9, 1–43.
- Cummins, J. (1978). Bilingualism and the development of metalinguistic awareness. *Journal of Cross-Cultural Psychology*, 9(2), 131–149. <https://doi.org/10.1177/002202217892001>
- Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. *Review of Educational Research*, 49(2), 222–251. <https://doi.org/10.3102/0034654304900222>
- Cummins, J. (1981). The role of primary language development in promoting educational success for language minority students. In California State Department of Education (Ed.), *Schooling and language minority students: A theoretical framework* (pp. 3–49). National Dissemination and Assessment Center.
- Cummins, J. (1991). Interdependent of first- and second-language proficiency in bilingual children. In E. Byalystok (Ed.), *Language processing in bilingual children* (pp.70–89). Cambridge University Press. <https://doi.org/10.1017/CBO9780511620652.006>
- Cummins, J. (2000). *Language, power and pedagogy: Bilingual children in the crossfire*. Multilingual Matters. <https://doi.org/10.21832/9781853596773>
- Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, 19(4), 450–466. <https://doi.org/10.1037/0278-7393.9.4.561>
- Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin and Review*, 3(4), 422–433. <https://doi.org/10.3758/BF03214546>
- Deacon, S., Wade-Woolley, H., & Kirby, J. (2007). Crossover: The role of morphological awareness in French immersion children's reading. *Developmental Psychology*, 43(3), 732–746. <https://doi.org/10.1037/0012-1649.43.3.732>
- De Jong, J. H. A. L., & van Ginkel, L. W. (1992). Dimensions in oral foreign language proficiency. In J. H. A. L. De Jong (Ed.), *The construct of language proficiency* (pp. 112–140). John Benjamins. <https://doi.org/10.1075/Z.62.19jon>
- \*Ding, Y., Guo, J-P., Yang, L-Y., Zhang, D., Ning, H., & Richman, L. C. (2011). Rapid Automated Naming and immediate memory functions in Chinese children who read English as a second language. *Journal of Learning Disabilities*, 46(4), 347–362. <https://doi.org/10.1177/0022219411424209>
- \*Edele, A., & Stanat, P. (2016). The role of first-language listening comprehension in second-language reading comprehension. *Journal of Educational Psychology*, 108(2), 163–180. <https://doi.org/10.1037/edu0000060>
- Ellis, N. (2011). Implicit and explicit SLA and their interface. In C. Sanz & R. P. Leow (Eds.), *Implicit and explicit language learning: Conditions, processes, and knowledge in SLA and bilingualism* (pp. 35–47). Georgetown University Press.
- \*Erçetin, G., & Alptekin, C. (2013). The explicit/implicit knowledge distinction and working memory: Implications for second-language reading comprehension. *Applied Psycholinguistics*, 34(4), 727–753. <https://doi.org/10.1017/S0142716411000932>



- \*Fecteau, M. L. (1999). First- and second-language reading comprehension of literary texts. *The Modern Language Journal*, 83(4), 475–493. <https://doi.org/10.1111/0026-7902.00036>
- \*Fontanini, I., & Tomitch, L. (2009). Working memory capacity and L2 university students' comprehension of linear texts and hypertexts. *International Journal of English Studies*, 9, 1–18.
- \*Francis, D. J., Snow, C. E., August, D., Carlson, C. D., Miller, J., & Iglesias, A. (2006). Measures of reading comprehension: A latent variable analysis of the diagnostic assessment of reading comprehension. *Scientific Studies of Reading*, 10(3), 301–322. [https://doi.org/10.1207/s1532799XSSR1003\\_6](https://doi.org/10.1207/s1532799XSSR1003_6)
- Gernsbacher, M. A. (1991). Cognitive processes and mechanisms in language comprehension: The structure building framework. In G. H. Bower (Ed.), *The psychology of learning and motivation* (pp. 217–263). Academic Press.
- Gernsbacher, M. A., Varner, K. R., & Faust, M. (1990). Investigating differences in general comprehension skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(3), 430–445. <https://doi.org/10.1037/0278-7393.16.3.430>
- \*Geva, E., & Ryan, E. B. (1993). Linguistic and cognitive correlates of academic skills in first and second languages. *Language Learning*, 43(1), 5–42. <https://doi.org/10.1111/j.1467-1770.1993.tb00171.x>
- \*Goodwin, A. P., August, D., & Calderon, M. (2015). Reading in multiple orthographies: Differences and similarities in reading in Spanish and English for English learners. *Language Learning*, 65(3), 596–630. <https://doi.org/10.1111/lang.12127>
- \*Gottardo, A., & Mueller, J. (2009). Are first and second language factors related in predicting L2 reading comprehension? A study of Spanish-speaking children acquiring English as a second language from first to second grade. *Journal of Educational Psychology*, 101(2), 330–344. <https://doi.org/10.1037/a0014320>
- Gough, P., & Tunmer, W. (1986). Decoding, reading, and reading disability. *Remedial and Special Education*, 7(1), 6–10. <https://doi.org/10.1177/074193258600700104>
- Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. Cambridge University Press.
- Grabe, W., & Stoller, F. (2019). *Teaching and researching reading* (3rd ed.). Routledge. <https://doi.org/10.4324/9781315726274>
- \*Grant, A., Gottardo, A., & Geva, E. (2012). Measures of reading comprehension: Do they measure different skills for children learning English as a second language? *Reading and Writing*, 25(1), 1889–1928. <https://doi.org/10.1007/s11145-012-9370-y>
- \*Guo, Y., & Roehrig, A. D. (2011). Roles of general versus second language (L2) knowledge in L2 reading comprehension. *Reading in a Foreign Language*, 23, 42–64.
- \*Harrington, M., & Sawyer, M. (1992). L2 working memory capacity and L2 reading skill. *Studies in Second Language Acquisition*, 14(1), 25–38. <https://doi.org/10.1017/S0272263100010457>
- \*Harsch, C., & Hartig, J. (2016). Comparing C-tests and Yes/No vocabulary size tests as predictors of receptive language skills. *Language Testing*, 33(4), 555–575. <https://doi.org/10.1177/0265532215594642>
- \*Hawas, H. M. (1990). Vocabulary and reading comprehension: An experimental study. *International Review of Applied Linguistics*, 87–88, 45–63. <https://doi.org/10.1075/itl.87-88.03haw>
- \*Haynes, M., & Carr, T. H. (1990). Writing system background and second language reading: A component skills analysis of English reading by native speakers of Chinese. In T. H. Carr & B. A. Levy (Eds.), *Reading and its development: Component skills approaches* (pp. 375–418). Academic Press.
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press.

- \*Henning, G. H., Ghawaby, S. M., Saadalla, W. Z., El-Rifai, M. A., Hannallah, R. K., & Mattar, M. S. (1981). Comprehensive assessment of language proficiency and achievement among learners of English as a foreign language. *TESOL Quarterly*, 15(4), 457–466. <https://doi.org/10.2307/3586486>
- \*Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing An Interdisciplinary Journal*, 2(2), 127–160. <https://doi.org/10.1007/BF00401799>
- \*Horiba, Y. (2012). Word knowledge and its relation to text comprehension: A comparative study of Chinese- and Korean-speaking L2 learners and L1 speakers of Japanese. *The Modern Language Journal*, 96(1), 108–121. <https://doi.org/10.1111/j.1540-4781.2012.01280.x>
- \*Irvine, P., Atai, P., & Oller, J. W. (1974). Cloze, dictation and the Test of English as a Foreign Language. *Language Learning*, 24(2), 245–252. <https://doi.org/10.1111/j.1467-1770.1974.tb00506.x>
- \*Jared, D., Cormier, P., Levy, B. A., & Wade-Woolley, L. (2011). Early predictors of biliteracy development in children in French immersion: A 4-year longitudinal study. *Journal of Educational Psychology*, 103(1), 119–139. <https://doi.org/10.1037/a0021284>
- Jenkins, J. R., Fuchs, L. S., van den Broek, P., Espin, C., & Deno, S. L. (2003). Sources of individual differences in reading comprehension and reading fluency. *Journal of Educational Psychology*, 95(4), 719–729. <https://doi.org/10.1037/0022-0663.95.4.719>
- \*Jeon, E. (2011). Contribution of morphological awareness to second-language reading comprehension. *The Modern Language Journal*, 95(2), 217–235. <https://doi.org/10.1111/j.1540-4781.2011.01179.x>
- \*Jeon, E. (2012). Oral reading fluency in second language reading. *Reading in a Foreign Language*, 24(2), 186–208.
- Jeon, E. (2018). Oral reading fluency. Teaching reading during reading strategies. *Wiley Online Library*. <https://doi.org/10.1002/9781118784235.eelto463>
- Jeon, E., & Yamashita, J. (2014). L2 reading comprehension and its correlates: A meta-analysis. *Language Learning*, 64(1), 160–212. <https://doi.org/10.1111/lang.12034>
- \*Jia, F., Gottardo, A., Koh, P. W., Chen, X., & Pasquarella, A. (2014). The role of acculturation in reading a second language: Its relation to English literacy skills in immigrant Chinese adolescents. *Reading Research Quarterly*, 49(2), 251–261. <https://doi.org/10.1002/rq.69>
- \*Jiang, X., Sawaki, Y., & Sabatini, J. (2012). Word reading efficiency and oral reading fluency in ESL reading comprehension. *Reading Psychology*, 33(4), 323–349. <https://doi.org/10.1080/02702711.2010.526051>
- \*Kato, S. (2009). Suppressing inner speech in ESL reading: Implications for developmental changes in second language word recognition processes. *The Modern Language Journal*, 93(4), 471–488. <https://doi.org/10.1111/j.1540-4781.2009.00926.x>
- Ke, S., Miller, R. T., Zhang, D., & Koda, K. (2021). A synthesis and meta-analysis of relations between morphological awareness and second language reading development. *Language Learning*, 71(1), 8–54. <https://doi.org/10.1111/lang.12429>
- \*Khalidieh, S. (2001). The relationship between knowledge of “Icraab,” lexical knowledge, and reading comprehension of non-native readers of Arabic. *The Modern Language Journal*, 85(3), 416–431. <https://doi.org/10.1111/0026-7902.00117>
- \*Kieffer, M. J., & Lesaux, N. K. (2008). The role of derivational morphology in the reading comprehension of Spanish-speaking English language learners. *Reading and Writing: An Interdisciplinary Journal*, 21(8), 783–804. <https://doi.org/10.1007/s11145-007-9092-8>
- \*Kieffer, M. J., & Vukovic, R. K. (2013). Growth in reading-related skills of language minority learners and their classmates: More evidence for early identification and intervention. *Reading and Writing*, 26(7), 1159–1194. <https://doi.org/10.1007/s11145-012-9410-7>

- \*Kim, Y.-S. (2012). The relations among L1 (Spanish) literacy skills, L2 (English) language, L2 text reading fluency, and L2 reading comprehension for Spanish-speaking ELL first grade students. *Learning and Individual Differences*, 22(6), 690–700. <https://doi.org/10.1016/j.lindif.2012.06.009>
- Klauda, S. L., & Guthrie, J. T. (2008). Relationships of three components of reading fluency to reading comprehension. *Journal of Educational Psychology*, 100(2), 310–321. <https://doi.org/10.1037/0022-0663.100.2.310>
- \*Koda, K. (1992). The effects of lower-level processing skills on FL reading performance: Implications for instruction. *The Modern Language Journal*, 76(4), 502–512. <https://doi.org/10.1111/j.1540-4781.1992.tb05400.x>
- Koda, K. (1998). The role of phonemic awareness in second language reading. *Second Language Research*, 14(2), 194–215. <https://doi.org/10.1191/026765898676398460>
- Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., & Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science*, 9, F13–F21. <https://doi.org/10.1111/j.1467-7687.2006.00468.x>
- \*Lam, K., Chen, Xi, Geva, E., Luo, Y. C., & Li, H. (2012). The role of morphological awareness in reading achievement among young Chinese-speaking English language learners: A longitudinal study. *Reading and Writing*, 25(8), 1847–1872. <https://doi.org/10.1007/s11145-011-9329-4>
- \*Larson, J. W. (1983). Skills correlations: A study of three final examinations. *The Modern Language Journal*, 67(3), 228–234. <https://doi.org/10.1111/j.1540-4781.1983.tb01500.x>
- \*Lee, J., & Schallert, D. L. (2014). Literate actions, reading attitudes, and reading achievement: Interconnections across languages for adolescent learners of English in Korea. *The Modern Language Journal*, 98(2), 553–573. <https://doi.org/10.1111/modl.12088>
- \*Lefrancois, P., & Armand, F. (2003). The role of phonological and syntactic awareness in second-language reading: The case of Spanish-speaking learners of French. *Reading and Writing: An Interdisciplinary Journal*, 16(3), 219–246. <https://doi.org/10.1023/A:1022874425314>
- \*Leider, C. M., Proctor, C. P., Silverman, R. D., & Harring, J. R. (2013). Examining the role of vocabulary depth, cross-linguistic transfer, and types of reading measures on the reading comprehension of Latino bilinguals in elementary school. *Reading and Writing*, 26(9), 1459–1485. <https://doi.org/10.1007/s11145-013-9427-6>
- Lems, K. (2003). Adult ESL oral reading fluency and silent reading comprehension (Unpublished doctoral dissertation). National-Louis University.
- Lems, K. (2006). Reading fluency and comprehension in adult English language learners. In T. Rasinski, C. Blachowicz, & K. Lems (Eds.), *Fluency instruction: Research based best practices* (pp. 231–252). Guilford.
- \*Lesaux, N. K., Crosson, A., Kieffer, M. J., & Pierce, M. (2010). Uneven profiles: Language minority learners' word reading, vocabulary, and reading comprehension skills. *Journal of Applied Developmental Psychology*, 31(6), 475–483. <https://doi.org/10.1016/j.appdev.2010.09.004>
- \*Li, M., & Kirby, J. R. (2015). The effects of vocabulary breadth and depth on English reading. *Applied Linguistics*, 36(5), 611–634. <https://doi.org/10.1093/applin/amu007>
- \*Li, T., McBride-Chang, C., & Wong, A. (2012). Longitudinal predictors of spelling and reading comprehension in Chinese as an L1 and English as an L2 in Hong Kong Chinese children. *Journal of Educational Psychology*, 104(2), 286–301. <https://doi.org/10.1037/a0026445>
- Libben, G. (2008). Words, mind, and brain. In P. van Sterkenburg (Ed.), *Unity and diversity of languages* (pp. 111–121). John Benjamins. <https://doi.org/10.1075/z.141.12lib>
- \*Limbird, C. K., Maluch, J. T., Rjosk, C., Stanat, P., & Merckens, H. (2014). Differential growth patterns in emerging reading skills of Turkish-German bilingual and German monolingual primary school students. *Reading and Writing*, 27(5), 945–968. <https://doi.org/10.1007/s11145-013-9477-9>

- \*Lindsey, K. A., Manis, F. R., & Bailey, C. E. (2003). Prediction of first-grade reading in Spanish-speaking English-language learners. *Journal of Educational Psychology*, 95(3), 482–494. <https://doi.org/10.1037/0022-0663.95.3.482>
- Linck, J. A., Osthus, P., Koeth, J. T., & Bunting, M. F. (2014). Working memory and second language comprehension and production: A meta-analysis. *Psychonomic Bulletin and Review*, 21, 861–883. <https://doi.org/10.3758/s13423-013-0565-2>
- \*Mancilla-Martinez, J., & Lesaux, N. K. (2010). Predictors of reading comprehension for struggling readers: The case of Spanish speaking language minority learners. *Journal of Educational Psychology*, 102(3), 701–711. <https://doi.org/10.1037/a0019135>
- \*Mecartty, F. H. (2000). Lexical and grammatical knowledge in reading and listening comprehension by foreign language learners of Spanish. *Applied Language Learning*, 11, 323–348.
- Melby-Lervåg, M., & Lervåg, A. (2011). Cross-linguistic transfer of oral language, decoding, phonological awareness and reading comprehension: A meta-analysis of the correlational evidence. *Journal of Research in Reading*, 34(1), 114–135. <https://doi.org/10.1111/j.1467-9817.2010.01477.x>
- Mizumoto, A., Sasao, Y., & Webb, S. A. (2019). Developing and evaluating a computerized adaptive testing version of the Word Part Levels Test. *Language Testing*, 36(1), 101–123. <https://doi.org/10.1177/0265532217725776>
- \*Morrison, L. (2004). Comprehension monitoring in first and second language reading. *Canadian Modern Language Review*, 61(1), 77–106. <https://doi.org/10.3138/cmlr.61.1.77>
- \*Nassaji, H., & Geva, E. (1999). The contribution of phonological and orthographic processing skills to adult ESL reading: Evidence from native speakers of Farsi. *Applied Psycholinguistics*, 20(2), 241–267. <https://doi.org/10.1017/S0142716499002040>
- Nation, I. S. P. (1983). Testing and teaching vocabulary. *Guidelines*, 5(1), 12–25.
- Nation, I. S. P. (2013). *Learning vocabulary in another language*. Cambridge University Press. <https://doi.org/10.1017/CBO9781139858656>
- Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, 31(7), 9–13.
- \*Noonan, B., Colleaux, J., & Yackulic, R. A. (1997). Two approaches to beginning reading in early French immersion. *Canadian Modern Language Review*, 53(4), 729–742. <https://doi.org/10.3138/cmlr.53.4.729>
- \*Oh, E. (2016). Comparative studies on the roles of linguistic knowledge and sentence processing speed in L2 listening and reading comprehension in an EFL tertiary setting. *Reading Psychology*, 37(2), 257–285. <https://doi.org/10.1080/02702711.2015.1049389>
- \*Olmez, F. (2016). Exploring the interaction of L2 reading comprehension with text- and learner-related factors. *Procedia--Social and Behavioral Sciences*, 232, 719–727. <https://doi.org/10.1016/j.sbspro.2016.10.098>
- \*Pae, H. K., & Sevvick, R. A. (2011). The role of verbal working memory in second language reading fluency and comprehension: A comparison of English and Korean. *International Electronic Journal of Elementary Education*, 2011, 4(1), 47–65.
- \*Park, H. (2001). The effects of L2 linguistic competence on L2 reading. *Journal of Pan-Pacific Association of Applied Linguistics*, 5, 91–103.
- \*Pasquarella, A., Chen, X., Lam, K., Luo, Y. C., & Ramirez, G. (2011). Cross-language transfer of morphological awareness in Chinese-English bilinguals. *Journal of Research in Reading*, 34(1), 23–42. <https://doi.org/10.1111/j.1467-9817.2010.01484.x>
- \*Payne, T. W., Kalibatseva, Z., & Jungers, M. K. (2009). Does domain experience compensate for working memory capacity in second language reading comprehension? *Learning and Individual Differences*, 19(1), 119–123. <https://doi.org/10.1016/j.lindif.2008.05.003>
- Perfetti, C. A. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, 11(4), 357–383. <https://doi.org/10.1080/10888430701530730>

- Perfetti, C. A., & Hart, L. (2002). The lexical basis of comprehension skill. In D. S. Gorfein (Ed.), *Decade of behavior. On the consequences of meaning selection: Perspectives on resolving lexical ambiguity* (p. 67–86). American Psychological Association.  
<https://doi.org/10.1037/10459-004>
- Perfetti, C. A., & Liu, L. (2006). Reading Chinese characters: Orthography, phonology, meaning, and the Lexical Constituency model. In P. Li, L. H. Tan, E. Bates, & O. J. L. Tzeng (Eds.), *The handbook of East Asian psychologists* (pp. 225–236). Cambridge University Press.  
<https://doi.org/10.1017/CBO9780511550751.022>
- \*Perkins, K., Brutton, S. R., & Pohlman, J. T. (1989). First and second language reading comprehension. *RELC Journal*, 20(2), 2–9. <https://doi.org/10.1177/00368828902000201>
- \*Proctor, C., Carlo, M., August, D., & Snow, C. (2005). Native Spanish-speaking children reading in English: Toward a model of comprehension. *Journal of Educational Psychology*, 97(2), 246–256. <https://doi.org/10.1037/0022-0663.97.2.246>
- \*Pulido, D., & Hambrick, D. Z. (2008). The “virtuous” circle: Modeling individual differences in L2 reading and vocabulary development. *Reading in a Foreign Language*, 20, 164–190.
- \*Rai, M. K., Loschky, L. C., & Harris, R. J. (2015). The effects of stress on reading: A comparison of first-language versus intermediate second-language reading comprehension. *Journal of Educational Psychology*, 107(2), 348–363. <https://doi.org/10.1037/a0037591>
- Read, J. (1998). Validating a test to measure depth of vocabulary knowledge. In A. Kunnan (Ed.), *Validation in language assessment* (pp. 41–60). Lawrence Erlbaum Associates.
- Rigg, P. (1977). The miscue ESL project. In H. D. Brown, C. A. Yorio, & R. H. Crymes (Eds.), *Teaching and learning ESL: Trends in research and practice* (pp. 106–118). TESOL.
- \*Royer, J. M., & Carlo, M. S. (1991). Transfer of comprehension skills from native to second language. *Journal of Reading*, 34, 450–455.
- Rydland, V., Aukrust, V. G., & Fulland, H. (2012). How word decoding, vocabulary and prior topic knowledge predict reading comprehension. A study of language-minority students in Norwegian fifth grade classrooms. *Reading and Writing*, 25(2), 465–482.  
<https://doi.org/10.1007/s11145-010-9279-2>
- \*Sang, F., Schmitz, B., Vollmer, H. J., Baumert, J., & Roeder, P. M. (1986). Models of second language competence: A structural equation approach. *Language Testing*, 3(1), 54–79.  
<https://doi.org/10.1177/026553228600300103>
- Segalowitz, N. (2003). Automaticity and second languages. In C. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 382–408). Blackwell.  
<https://doi.org/10.1002/9780470756492.ch13>
- \*Shang, H. (2010). Reading strategy use, self-efficacy and EFL reading. *The Asian EFL Journal Quarterly*, 12, 18–42.
- Shin, J. (2020). A meta-analysis of the relationship between working memory and second language reading comprehension: Does task type matter? *Applied Psycholinguistics*, 41(4), 873–900. <https://doi.org/10.1017/S0142716420000272>
- \*Shiotsu, T. (2010). *Components of L2 reading: Linguistic and processing factors in the reading test performances of Japanese EFL learners*. Cambridge University Press.
- \*Shiotsu, T., & Weir, C. J. (2007). The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance. *Language Testing*, 24(1), 99–128. <https://doi.org/10.1177/0265532207071513>
- \*Siu, C. T., & Ho, C. S. (2015). Cross-language transfer of syntactic skills and reading comprehension among young Cantonese–English bilingual students. *Reading Research Quarterly*, 50(3), 313–336. <https://doi.org/10.1002/rtrq.101>



- Sparks, R. L., & Luebbers, J. (2018). How many U.S. high school students have a foreign language reading “disability”? Reading without meaning and the Simple View. *Journal of Learning Disabilities*, 51(2), 194–208. <https://doi.org/10.1177/0022219417704168>
- Spearman, C. (1904). ‘General intelligence,’ objectively determined and measured. *The American Journal of Psychology*, 15(2), 201–293. <https://doi.org/10.2307/1412107>
- \*Swanson, H. L., Orosco, M. J., Lussier, C. M., Gerber, M., & Guzman-Orth, D. A. (2011). The influence of working memory and phonological processing on English language learner children’s bilingual reading and language acquisition. *Journal of Educational Psychology*, 100(4), 818–856. <https://doi.org/10.1037/a0024578>
- Taguchi, E. (1997). The effects of repeated readings on the development of lower identification skills of FL readers. *Reading in a Foreign Language*, 11(1), 97–119.
- Treiman, R. (1993). *Beginning to spell: A study of first-grade children*. Oxford University Press. <https://doi.org/10.1093/oso/9780195062199.001.0001>
- \*Tsai, Y., Ernst, C., & Talley, P. C. (2010). L1 and L2 strategy use in reading comprehension of Chinese EFL readers. *Reading Psychology*, 31(1), 1–29. <https://doi.org/10.1080/02702710802412081>
- \*Van Gelderen, A., Schoonen, R., de Glopper, K., Hulstijn, J., Simis, A., & Snellings, P. (2004). Linguistic knowledge, processing speed, and metacognitive knowledge in first- and second-language reading comprehension: A componential analysis. *Journal of Educational Psychology*, 96(1), 19–30. <https://doi.org/10.1037/0022-0663.96.1.19>
- \*Walter, C. (2004). Transfer of reading comprehension skills to L2 is linked to mental representations of text and to L2 working memory. *Applied Linguistics*, 25(3), 315–339. <https://doi.org/10.1093/applin/25.3.315>
- \*Wang, M., Cheng, C. X., & Chen, S. W. (2006). Contribution of morphological awareness to Chinese-English biliteracy acquisition. *Journal of Educational Psychology*, 98(3), 542–553. <https://doi.org/10.1037/0022-0663.98.3.542>
- \*Winke, P. (2013). An investigation into second language aptitude for advanced Chinese language learning. *The Modern Language Journal*, 97(1), 109–130. <https://doi.org/10.1111/j.1540-4781.2013.01428.x>
- \*Xue, J., & Jiang, X. (2017). The developmental relationship between bilingual morphological awareness and reading for Chinese EFL adult learners: A longitudinal study. *Reading and Writing*, 30(2), 417–438. <https://doi.org/10.1007/s11145-016-9683-3>
- \*Yamashita, J., & Shiotsu, T. (2017). Comprehension and knowledge components that predict L2 reading: A latent-trait approach. *Applied Linguistics*, 38(1), 43–67. <https://doi.org/10.1093/applin/amu079>
- Yoshikawa, L. & Yamashita, J. (2014). Phonemic awareness and reading comprehension among Japanese adult learners of English. *Open Journal of Modern Linguistics*, 4(4), 471–480. <https://doi.org/10.4236/ojml.2014.44039>
- \*Zhang, D. (2012). Vocabulary and grammar knowledge in second language reading comprehension: A structural equation modeling study. *The Modern Language Journal*, 96(4), 558–575. <https://doi.org/10.1111/j.1540-4781.2012.01398.x>
- Zhang, H., & Koda, K. (2018). Word-knowledge development in Chinese as a heritage language learners: A comparative study. *Studies in Second Language Acquisition*, 40(1), 201–223. <https://doi.org/10.1017/S0272263116000450>

## Appendix A. Acceptable measures of included reading correlates and reading comprehension

**L2 reading comprehension:** To be considered a reading comprehension measure, the measure had to involve oral or silent reading of an L2 passage and answering comprehension questions. L2 reading was defined as reading in a language that is not in the study participant's first language or languages. Studies that involved sentence-level comprehension were not included in this study.

**L2 decoding:** An acceptable decoding measure assesses either silent or oral reading efficiency and/or accuracy of L2 pseudowords (nonexistent words that conform to the L2 orthographical rules) and/or real words.

**L2 phonological awareness:** An acceptable phonological awareness measure involves some type of generation, judgment, and/or manipulation (e.g., deletion, insertion) of sound units such as phoneme, onset, rhyme, coda, and/or syllable. In order to avoid confusing decoding and phonological awareness, we did not include studies that had a phonological processing measure which involved the processing of grapheme.

**L2 vocabulary knowledge:** To be deemed as a measure of vocabulary knowledge, the test must assess one or more aspects of L2 word knowledge; examples of such aspects included but were not limited to the size (breadth), depth, knowledge of word usage, or lexical processing efficiency.

**L2 grammar knowledge:** An acceptable grammar test had to assess some form of L2 morpho-syntactic or syntactic knowledge; such forms of knowledge included but were not limited to the recognition knowledge (e.g., recognizing morphosyntactic or syntactic error in a speeded or in an unspeeded setting) and productive knowledge (e.g., producing a correctly conjugated word for the given context) in the explicit or implicit form.

**L1 reading comprehension:** To be considered as a measure of L1 reading, the measure had to assess comprehension accuracy of reading passages written in the participant's L1.

**L2 listening comprehension:** An L2 listening comprehension measure had to involve assessing passage-level comprehension ability of an aural text.

**Working memory:** All traditional working memory measures such as digit or letter span test and reading span tests were considered acceptable for this study.

**L2 orthographic knowledge:** To be considered as an orthographic knowledge test, the measure has to assess the knowledge of L2-specific orthographic and/or spelling rules at the word or sentence level.

**L2 morphological knowledge:** An acceptable morphological knowledge test assesses the knowledge of word parts and their derivational and inflectional behavior (e.g., affixation, inflection, compounding).

**Metacognition:** A wide range of operational definitions such as perceived and actual use of reading strategies, self-assessment of reading comprehension, comprehension monitoring, and awareness of discourse structures were deemed acceptable for this study.

**Oral reading fluency:** In order for a measure to be considered as oral reading fluency, it had to involve an oral rendition of a connected passage in L2.

## Appendix B. Moderator analysis results of high evidence correlates

### *Decoding*

Moderator variable	k	r [95% CI]	Significant test of difference (Q test)
<i>Age</i>			
Child	21	.624 [.455–.749]	3.062 ( <i>p</i> = .08)
Adolescent/Adult	8	.451 [.341–.550]	
L1-L2 script distance			
AA	21	.614 [.448–.740]	.812 ( <i>p</i> = .367)
AN	8	.507 [.301–.668]	
<i>Language setting</i>			
FL	6	.424 [.310–.527]	4.174* ( <i>p</i> = .041)
SL	23	.622 [.462–.743]	
<i>Measurement type</i>			
Pseudoword	9	.687 [.365–.862]	1.042 ( <i>p</i> = .594)
Real word	9	.537 [.378–.665]	
Pseudoword and real word	11	.533 [.429–.623]	
<i>Language proficiency</i>			
Basic	2	.673 [.387–.841]	0.453 ( <i>p</i> = .501)
Beyond basic	27	.579 [.431–.696]	
<i>Language difference</i>			
II	16	.646 [.458–.779]	1.739 ( <i>p</i> = .187)
IN	13	.498 [.350–.622]	

### *Phonological Awareness*

Moderator variable	k	r [95% CI]	Significant test of difference (Q test)
Age			
Child	18	.616 [.515–.700]	.322 ( <i>p</i> = . 571)
Adolescent/Adult	2	.546 [.269–.739]	
Language difference			
II	9	.580 [.482–.664]	.314 ( <i>p</i> = .575)
IN	11	.636 [.441–.774]	
Script distance			
AA	13	.544 [.452–.625]	2.394 ( <i>p</i> = .122)
AN	7	.722 [.500–.855]	
Measurement type			
In-house	10	.631 [.525–.718]	.040 ( <i>p</i> = .841)
Standardized	8	.651 [.460–.784]	



*Vocabulary Knowledge*

Moderator variable	k	r [95% CI]	Significant test of difference (Q test)
Age			
Child	24	0.698 [.581–.787]	
Adolescent/Adult	27	.746 [.608–.840]	0.367 ( $p = .545$ )
L1-L2 script distance			
AA	32	.748 [.640–.827]	
AN	19	.679 [.501–.803]	0.627 ( $p = .429$ )
Language setting			
FL	27	.797 [.656–.884]	
SL	24	.649 [.571–.716]	3.115 ( $p = .078$ )
Measurement type (Oh [2016] removed)			
In-House	20	.729 [.552–.843]	
Standardized	30	.721 [.614–.802]	.008 ( $p = .928$ )
Measurement type			
Production	6	.856 [.556–.958]	
Selection	45	.716 [.605–.800]	1.178 ( $p = .278$ )
Measurement type			
Contextualized	9	.867 [.680–.948]	
Isolated	40	.689 [.583–.771]	3.171 ( $p = .075$ )
Language proficiency (Shiotsu & Weir [2007]: Sample 2 removed because it was mixed)			
Basic	2	.626 [.239–.842]	
Beyond basic	48	.713 [.624–.784]	0.359 ( $p = .549$ )
Language difference (Horiba Samples 1 and 2 removed because they were both NN)			
II	23	.690 [.567–.783]	
IN	26	.764 [.630–.854]	0.850 ( $p = .356$ )

*Grammar Knowledge*

Moderator variable	k	r [95% CI]	Significant test of difference (Q test)
Age			
Child	5	.561 [.347–.720]	
Adolescent/Adult	21	.726 [.526–.850]	1.684 ( $p = .194$ )
Script distance			
AA	16	.750 [.603–.847]	
AN	10	.600 [.030–.876]	.577 ( $p = .448$ )
Language difference			
II	7	.788 [.461–.927]	
IN	19	.656 [.382–.824]	.638 ( $p = .424$ )
Language setting			
FL	22	.714 [.518–.838]	
SL	4	.597 [.290–.793]	.641 ( $p = .423$ )
Measurement type			
In-house	18	.719 [.490–.855]	
Standardized	8	.644 [.377–.812]	.275 ( $p = .600$ )

Moderator variable	k	r [95% CI]	Significant test of difference (Q test)
Measurement type			
Completion	16	.793 [.604–.898]	
GJT	5	.367 [.150–.550]	
Other	5	.562 [.331–.730]	9.402* ( $p = .009$ )
Language proficiency			
B	2	.900 [neg.686–.999]	
BB	24	.670 [.490–.795]	.310 ( $p = .578$ )

### *L1 Reading Comprehension*

Moderator variable	k	r [95% CI]	Significant test of difference (Q test)
Age			
Child	17	.472 [.374–.560]	
Adolescent/Adult	17	.493 [.241–.683]	.028 ( $p = .866$ )
Script distance			
AA	18	.558 [.365–.705]	
AN	16	.372 [.291–.448]	3.159 ( $p = .076$ )
Language difference			
II	12	.626 [.403–.778]	
IN	22	.369 [.296–.439]	4.571* ( $p = .033$ )
Language setting			
FL	24	.473 [.291–.622]	
SL	10	.514 [.415–.602]	.189 ( $p = .664$ )
Measurement type (Brisbois sample 1 removed)			
In-house	24	.445 [.269–.592]	
Standardized	9	.589 [.488–.675]	2.457 ( $p = .117$ )
L2 proficiency			
Basic	7	.372 [.188–.530]	
Beyond basic	27	.512 [.362–.636]	1.564 ( $p = .211$ )

### *L2 Listening Comprehension*

Moderator variable	k	r [95% CI]	Significant test of difference (Q test)
Age			
Child	7	.848 [.122–.983]	
Adolescent/Adult	13	.778 [.674–.852]	.126 ( $p = .723$ )
Language difference			
II	12	.872 [.641–.958]	
IN	8	.677 [.487–.805]	2.439 ( $p = .118$ )
Script distance			
AA	17	.844 [.669–.930]	
AN	3	.507 [.122–.759]	4.739* ( $p = .029$ )
Language setting			
FL	12	.755 [.644–.835]	
ML	8	.862 [.298–.980]	.375 ( $p = .541$ )

Working Memory

Moderator variable	k	r [95% CI]	Significant test of difference (Q test)
Age			
Child	7	.379 [.251-.495]	
Adolescent/Adult	12	.303 [.163-.430]	.690 ( $p = .406$ )
Language difference			
II	8	.398 [.267-.515]	
IN	11	.278 [.153-.394]	1.819 ( $p = .177$ )
Script distance (Harrington & Sawyer removed because it is the only AS)			
AA	14	.381 [.291-.464]	
AN	4	.232 [-.028-.463]	1.295 ( $p = .255$ )
Language setting			
FL	15	.275 [.170-.373]	
SL	4	.487 [.425-.544]	13.822** ( $p = .000$ )
Testing language (2 samples did not use a linguistic WM measure)			
L1	5	.203 [.041-.355]	
L2	12	.434 [.332-.526]	6.208* ( $p = .013$ )

## CHAPTER 4

# L2 writing

## Theory and research

Rob Schoonen

Radboud University, Nijmegen

This chapter aims to identify sources of individual differences in writing proficiency, and ways to measure these differences. The construct of writing proficiency is delineated and embedded in cognitive approaches to writing research. It is considered to be a multi-faceted construct, which implies that various component skills contribute to the quality of the eventual writing performance. Writing in a second language often builds on L1 writing skills and procedures, but also raises its own questions, such as the role of transfer of (metacognitive) knowledge, and the writer's strength and weaknesses in the target language. The best way to assess a language learner's writing proficiency very much depends on the assessment purposes, and can range from a focused measurement of sub-skills with discrete-point tests, for diagnostic purposes, to a general performance assessment with holistic rating, for admission purposes. A number of characteristics of these measurements are discussed in the second part of this chapter.

### 1. Introduction

Being literate, and more specifically being able to write, still is an essential condition for people to benefit from education and to get ahead in life. Some aspects of human writing are changing rapidly nowadays with new technological developments and ever-changing social media, others will remain largely the same, and will pose the same challenges to newcomers to the domain of literacy. In this chapter, we will focus on these challenges, raising the questions of “what does it need to become a proficient writer?”, and “what are the additional challenges when the language to write in is not your first or dominant language?” The approach we take will be largely theoretical and conceptual, however we will build on empirical research, without the intention to be comprehensive. Extensive and systematic literature reviews regarding writing proficiency and its internal and external correlates can be found in Chapter 5 and 6, respectively.

## 2. Defining (L2) writing proficiency

Any definition of writing proficiency will comprise elements like ‘converting content or meaning into written language’ or ‘translating idea units into written code’. Analytically speaking, we are dealing with a productive language proficiency, ignoring for the time being that writers might –every now and then– reread their text written so far. Considering writing a productive language proficiency implies that meaning (as in ideas or concepts) is transformed into external language, in this case written language or text. Thus, writing proficiency can be regarded as a form of *language proficiency*. This raises the question: where does the proficiency as a *language proficiency* begin and where does it end? Demarcation of the writing process at the beginning and at the end is still problematic. For example, should the ability to create meaning or ideas be considered part of writing proficiency? Is a writer with limited inspiration a poor writer, or just a less creative person? From a literary point of view, the person might be viewed as a poor writer, but from a cognitive-educational or second language acquisition (SLA) point of view, we probably want to be more lenient, because we usually accept that students who have to write about a very unfamiliar topic, let us say *The evolution of the ladybug*, will produce little text. So, generation of content might not be at the core of writing proficiency (cf. Weigle, 2002), or maybe not writing proficiency at all. As demarcation of the initial stages in writing as a language proficiency is debatable, and so may be the final stages, that is the actual writing down, in particular when it involves handwriting. The quality of the transcription is commonly ignored in the appreciation of a person’s writing ability, although it is essential. If the writing cannot be deciphered, there will be no transfer of information or meaning from writer to reader. Still, the physical-motor skill to write legibly is usually not considered part of a person’s writing proficiency. Although there seem to be certain advantages to handwriting compared to typing (Frangou, Ruokamo, Parviainen, & Wikgren, 2018), it will be a matter of time that some form of keyboard has replaced most handwriting. This may solve problems with legibility, but then the question is whether keyboard fluency should be considered part of writing proficiency.

When we exclude the problematic stages that are not or only very loosely related to language use, we could provisionally define writing proficiency as the ability to convert (given) meaning into (legible) script or text. This is a very technical, ‘sterile’ definition, and the construct definition might need to be adapted according to specific instructional or assessment contexts (Grabe & Kaplan, 1996; Weigle, 2002). Nevertheless, in most situations the end product of the writing process, i.e., the text, needs to comply with some kind of optimality criteria that are related to the readability of the text (to be distinguished from legibility). Well written texts should be easy to read and get the message across in an unambiguous way, at least

in factual writing. Poetry or literary texts may benefit from some level of ambiguity or implicitness. In factual writing, the reader should be able to derive the meaning from the text as it was intended by the writer. The question is how can we decide on this quality of texts. Discussing the criteria for evaluating writing performances is opening Pandora's Box. Are there any objective criteria along which texts can be rated? Criteria for optimal writing will largely depend on the rhetorical requirements and the intended readership, if not individual readers. Readers do not always agree upon the quality of texts which makes the assessment of writing proficiency a precarious endeavor. Text features that make texts easier to read are not necessarily the features that are appreciated by raters in their evaluations (McNamara, Crossley, & MacArthur, 2010). Coming up with a construct definition is one thing, but coming up with a workable operational definition is another. For a more extensive discussion on these issues, see Bachman and Palmer (1996) as well as Weigle (2002).

In the remainder of the chapter, we will first focus on the components of (L2) writing proficiency that constitute the complex construct of writing proficiency. Secondly, we will zoom in on issues that are associated with assessment of writing performances, that is eliciting writing products as well as the rating thereof.

### 3. Sources of individual differences in cognitive models of writing

Discussing writing ability as an *ability* implies that writers differ in their capability to produce proper texts in given contexts. The question that bothers theorists and teachers alike is what causes these individual differences. Individual differences can arise from all kinds of stages in the writing process: generating content, structuring it, retrieving the right words, building the appropriate sentences and so forth. This list basically states that all constituting processes in writing are potential sources for individual differences. Research on individual differences and research on cognitive processes in writing should be merged in the way that was already advocated by Cronbach (1957). To come up with possible sources for individual differences, we have to take a closer look at models of writing, that is, models that describe the cognitive processes involved in writing and the linguistic, cognitive and related sources that feed into these writing processes. Comprehensive cognitive models of writing are scarce (Becker, 2006; Deane et al., 2008; MacArthur & Graham, 2016; Schoonen, Snelling, Stevenson, & Van Gelderen, 2009), and these models describing the main cognitive processes involved in writing are rather sketchy (cf. Chenoweth & Hayes, 2001; Hayes, 2012; Hayes & Flower, 1980), without much detail on the resources a writer employs when writing (MacArthur & Graham, 2016) (see Grabe & Kaplan [1996] for an exception when it concerns the writing resources required and see Leggette, Rutherford, Dunsford, & Costello, [2015] and Cumming [2016] for other

models of writing, such as contrastive rhetoric or social cultural models). Many other studies have investigated parts of the writing process, such as planning (Hayes & Nash, 1996; see Torrance, 2016), the act of writing (Fayol, 2016), and especially revising (Flower, Hayes, Carey, Schriver, & Stratman, 1986; Stevenson, Schoonen, & De Glopper, 2006), or specific enabling skills, such as working memory (Kellogg, Whiteford, Turner, Cahill, & Mertens, 2013). In their discussion of research into cognitive writing processes, MacArthur and Graham point out why domains like working memory, self-regulation, and motivation have received special attention in cognitive research. Writers who are able to keep sufficient information activated in their working memory seem to be more successful in their writing than those who cannot; writers who manage the writing process, the writing context, and their own behaviors effectively write better texts than those who have problems in regulating their writing and writing context; and obviously writers who are highly motivated to write generally are more successful than those who are not or less motivated, especially when writing is very demanding. All these subprocesses and components are potential sources for individual differences.

In 2012 Hayes (2012) modeled the cognitive processes of writing similar to Levelt's (1989) model of speaking, a model whose main components are corroborated by an extensive amount of research. Given this suggested similarity between writing and speaking, one could be inclined to adopt a Simple View of Writing (Juel, 1988), analogous to the successful Simple View of Reading (Gough & Tunmer, 1986). In Juel's simple view, writing consists of 'ideation' and 'spelling', ideation being the ability to generate and organize content into text and sentences (p. 438) and largely overlapping with general productive language abilities (as used in speaking). Yet, the content needs to be 'articulated' in the written mode using one's transcription abilities, including both motor and spelling abilities. Several studies have shown, however, that written language should be considered a different register or genre from spoken language (Biber, 1991; Schleppegrell, 2004). Written discourse is usually read at another time and place than where and when it was produced.

Despite the differences between spoken and written language, the analogy between the processing parts of writing and speaking is useful (Grabowski, 1996; Schoonen et al., 2009; Weigle, 2002). We can consult the extensive psycholinguistic research on speaking processes as is described in, and was stimulated by, the model of speaking of Levelt (1989). For example, the writer, just like a speaker, has to think of the message he or she wants to get across. In speaking this planning can be very 'locally determined' when it concerns an interaction with quick exchanges. However, in text production, the writer usually has to make plans for longer stretches of text which requires that the overall goal of the text is kept in mind. In this respect the writer has to plan at multiple levels, that is, for instance, at the discourse level organizing the counter-arguments in argumentative text,

and at the grammatical level formulating the next sentence to express the first counter-argument. This is not to say that speakers do not plan larger pieces of discourse, but writers usually have to deliver the full text, which requires planning at all possible levels of text production. Once the major planning has been done or at least is well on track, the writer has to generate content, i.e., what will be the most appropriate counter-argument for the argumentative text. Most likely planning and content generation go hand-in-hand, mutually influencing each other: planning the information that comes available and searching one's mind to fill possible gaps in the planned structure. This first cluster of processes leads to what Levelt (1989) called a preverbal message, which we can picture as a message in terms of propositions that still need to be translated into language. Then, the writer's linguistic knowledge and skills come into play.

Once the writer knows what to write, the available concepts will activate the corresponding lemmas in the writer's mental lexicon and predications need to be expressed in the most appropriate grammatical structures, including processes of linearization. At this stage, the language is still 'inner speech' and needs to be converted into written language, which means that the written (correct) word *forms* need to be activated or selected. The writer has to make decisions about the spelling of words. Proficient writers will know how to apply phoneme-grapheme correspondence rules or will have developed ready-made word images in visual memory. The balance between rule-application and word images will not only differ between novice and expert writers, but will also be dependent on the language involved. Rule-application is more effective in languages with a shallow orthography than in languages with a so-called deep orthography. In English, the novice writer might write 'I want to buy tea' as *I wont too bi tee*, depending on language exposure and experience with written English language. English is notoriously difficult when it concerns orthography, whereas other languages such as Italian or Finnish are much easier. The orthography can be distinctive in meaning (*read* or *reed*) or contain morphosyntactic information (Dutch: *word* [become, 1sg] or *wordt* [become, 2/3sg]) and thus it does matter for the interpretation of the written message.

In the final step of the writing process the orthographic words need to be externalized by handwriting or keyboarding, that is the motor act of writing. The fluency of the physical writing depends on the motor and/or keyboard skills, but also on the previous stages. If the writer is insecure about the spelling of the words to be used the writing slows down (Fayol, 2016). Slow writing due to lack of motor or keyboard skills can also interfere with the spelling of the words when the writer loses track of the correct spelling of the word(s) involved.

The current presentation of the writing subprocesses is very linear, but in real-life writing there is no need to act fully linearly. Writers work with a different time scale compared to speakers as they do not experience the 'pressure of speech'



as speakers do. Therefore, writers can and usually do work cyclically; the cognitive processes can alternate at any time during the writing, and at any time the writer can take a moment to reread the writing done so far and review whether he or she is on track and has formulated the intended message appropriately. If this is not the case the author can revise parts of the written text immediately or postpone it to the end of the first cycle of writing. We can consider writing as a cyclic process, although not all cycles will be equal or fully accomplished. It is one of the major subskills of writing to be able to regulate one's own writing process efficiently. The production of a larger piece of text can be an overwhelming task and subskills of good self-regulation will be an asset to the writing process. Knowledge of text structure, or the build-up of paragraphs, the awareness to read the text written so far every now and then, and the ability to read it as a naive reader or at least as a reader of the intended readership will greatly assist in making the right decisions during the writing process. In several studies this kind of metacognitive knowledge was shown to be good predictor of writing proficiency (Schoonen et al., 2003; Victori, 1999). Writing without much self-regulation regarding planning, goal-setting, and audience-awareness is what Bereiter and Scardamalia (1987) observed in (young) novice writers, so-called knowledge-telling, or as MacArthur and Graham (2016) call it, a retrieve-and-write approach. More proficient writers mold the information according to their goals and the readers' need ('knowledge transforming', Bereiter & Scardamalia, 1987). It is obvious that the construct of self-regulation is a multi-faceted construct. Zimmerman and Risemberg (1997) distinguished ten self-regulatory techniques, and a meta-analysis of Santangelo, Harris, and Graham (2016) showed that five of these, which could be included in the analyses, showed significant effects on writing scores, with effect sizes ( $d$ ) ranging from .30 (for self-selected models, tutors or books as knowledge sources) to 1.06 (for mental imagery to create a vivid picture of settings or activities to be described) (Santangelo et al., 2016).

#### **4. Process-product**

We started our previous section with the claim that every subprocess in the overall writing process is potentially a source for individual differences and although not every subprocess performance will affect the end product noticeably, we would like to maintain this claim for educational and diagnostic purposes. Furthermore, writing processes draw from linguistic and metacognitive knowledge sources, such as the mental lexicon, syntactic construction repertoires, and regulatory strategies inventories, which makes the act of writing a text a complex enterprise. Grabe and Kaplan provide a nice and extensive "taxonomy of writing skills, knowledge bases

and processes” (Grabe & Kaplan, 1996, p. 216). Table 1 lists only the main categories of this taxonomy, but it still shows the complexity of the writing act and the interaction and involvement of different sources and processes. At the same time, one could object that not all processes show in the end product, the written text. A poorly structured text can be the result of no planning, superficial reviewing or a lack of text structural knowledge, or a combination of the three. In most cases, the reader cannot infer from the text what went wrong and just has to deal with the poorly structured text. In addition, if the reader is also the rater of the writing sample, the score will be negatively affected by the quality of the structure. For the teacher of the student-writer it *does* matter which subprocess caused the structure

**Table 1.** Main categories of Grabe and Kaplan’s Taxonomy of writing skills, knowledge bases and processes (o.c., 1996: 216ff)

	Main categories	Examples of subcategories (out of number provided)
I	Educational settings for writing	Classroom, Office, Home (7)
II	Educational writing tasks	Lecture notes, Letters, Narratives, Argumentative essays, Abstracts, Theses (22)*
III	Educational texts used and produced	Textbooks, Dictionaries, Research journal articles (14)
IV	Topics for academic writing	Personal expressive, Biographies, Topics from professional disciplines (11)
V	The writer’s intention, goals, attributions, and attitudes	Writer’s reinterpretation of the task, Willingness to be understood, Motivation to perform to capacity (13)*
VI	Linguistic knowledge	Knowledge of the written code, Vocabulary, Syntactic/structural knowledge (6)*
VII	Discourse knowledge	Knowledge of informational structuring, Knowledge to recognize main topics, Knowledge of organizing schemes (9)
VIII	Sociolinguistic knowledge	Functional uses of written language, Application and interpretable violation of Gricean maxims, Register and situational parameters (5)*
IX	Further audience considerations	Number of audience, Degree of familiarity with audience, Status of audience with respect to writer (6)
X	Knowledge of the world	Declarative, Episodic, Procedural (3)
XI	Writing process skills	Goal planning routines, Rapid production routines, Revising routines (7)*
XII	Writing process strategies	Monitoring text production, Using invention strategies, Getting feedback from others (20)*

\* Some subcategories are subdivided further, such as different types Argumentative essays, or Functional uses of written language.

to be as it is. Advising the writer to plan or review the text more extensively does not help if the writer lacks the relevant knowledge about text structures. For a full diagnosis of the writer's problems more detailed testing would be necessary. The relationship between writing subprocesses and the quality of the end product is not always straightforward. Studies about this relation are scanty (see Breetvelt, Van den Bergh, & Rijlaarsdam [1994] for one of the first studies explicitly addressing this issue) and from the few studies we learn that it is not just the *frequency* of subprocesses (how often does a writer revise, plan etc.), but it also matters *when* the subprocesses are executed in the overall process of writing. For example, planning near the end of the writing might not be as fruitful as planning in the early stages of the writing (cf. Van den Bergh, Rijlaarsdam, & Van Steendam, 2016).

## 5. L1-L2-FL

So far, we have not been very explicit about the language the presumed writer is writing in. It is obvious that writing in a second or foreign language (henceforth, L2) causes extra challenges. Certainly, there will be extra challenges in the domain of linguistic, discourse, and sociolinguistic knowledge and skills (cf. Table 1). The extent to which there will be increased challenges regarding the non-linguistic knowledge and skills, such as the metacognitive knowledge and skills described above is still subject of discussion. One could claim that the self-regulatory skills are language independent and that these skills easily transfer from L1 to L2, but this may not be as straightforward as it seems. Analogous to the discussions in L1 and L2 reading (cf. Alderson, Huhta, & Nieminen, 2016), we have to take into consideration possible thresholds of L2 proficiency that prevent L1 skills from transferring to L2 writing. Writers that are fully preoccupied with –for example- lexical searches in the L2, may have no more cognitive resources available to think about text structure or self-regulatory writing processes (Schoonen et al., 2003; Whalen & Ménard, 1995). Roca de Lario, Marín, and Murphy (2001) made a temporal analysis of how 21 Spanish EFL writers of three different proficiency levels spent their time writing a Spanish and an English text. It was evident from their think-aloud protocol analysis that the EFL writing involved more problem-solving and less fluent formulation processes than in L1 writing. Roca de Lario and colleagues found that the (percentual) time distribution in L1 and EFL writing was more or less the same, but proficient writers spent less time on formulating their message. Proficient writers' time investment seemed more focused to the middle part of the writing, which was preceded by time devoted to planning and more concerns about writing strategies. Less proficient writers devote most of their initial time to generating text and this process gradually decreases compared to other writing processes. Wang

and Wen (2002) explored the EFL writing of sixteen Chinese writers and found that they used their L1 very frequently in their L2 writing of a narrative and an argumentative text. Writers resorted to their L1 during idea generation and organizing and process controlling, which together constituted almost one third of the processes. In text generating (about two thirds of the processing), the writers used their L1 far less frequently, 12% of the activities in narration and 15% in argumentation, compared to over 80% of the time in process controlling activities. Overall, student writers tended to use L1 slightly more frequently in narration than in argumentation, and advanced students tended to refrain more from the use of their L1 than the intermediate students, although individual variation in the use of L1 was substantial. Woodall (2002) studied language switching in 28 student writers representing various L1-L2 language combinations: Japanese-English, English-Japanese, Spanish-English and English-Spanish. Half of the sample of 28 participants was considered intermediate, the other half advanced, and all participants performed a narrative (letter) and an expository (essay) writing task. Again, writers used their L1 in L2 writing, although the advanced writers used far less L1 than the intermediate L2 writers, both in terms of frequency and duration. The role of L1 was mediated by task, proficiency level, and 'cognate-status' of the languages. For example, less proficient writers used the L1 more frequently in the L2 writing, not only for idea generation and processing control, but also when lexical searches were needed, and when the task was more difficult L1 was used for longer periods. We can imagine that (beginning) L2 writers prefer to deal with the regulatory challenges in their L1, but ultimately the L2 writer has to come up with the right words and sentence structures to formulate the concepts and predications that express the ideas generated.

Most L2 writers became literate in their L1 first, and thus bring their L1 experience to the L2 writing task. It turns out that L2 writers also use their L1 linguistic knowledge to keep their L2 language production going. Murphy and Roca de Larios (2010) have studied specifically the role of L1 in L2 lexical searches. In the small-scale study seven EFL participants wrote an argumentative and a narrative text, and as expected the more difficult (argumentative) text required about three times more lexical searches, on average 1.87 per 100 words, than the writing of the narrative text with 0.6 lexical searches per 100 words. L1 Spanish was frequently used to solve the lexical problems, helping generating lexical units, making evaluations and decisions at different levels and by self-questioning. Although the researchers point to the notable individual differences, the findings are in line with aforementioned findings; writers use their L1 in L2 writing and do so to varying degrees. Writers writing in their third, fourth or *n*-th language have even more resources at hand to solve lexical problems, and they indeed seem to use these previously learned languages (Tullock & Fernández-Villanueva, 2013). In lexical searches the writer can capitalize on cognates. However, syntactic structuring of sentences or

structuring information will be more problematic as previously learned languages might prime structures different from the ones required in the target language (Grabe & Kaplan, 1996; Van Vuuren & Berns, 2018). These cross-linguistic influences will not be limited to the different pieces of L1 and L2 or FL linguistic knowledge a writer has, but will be extended to all sorts of knowledge and routines of writing. The (reduced) list of skills, knowledge bases, and processes or strategies in Table 1 can be seen as a list of potential sources of L1-L2/FL transfer, which in some cases will lead to improved second language writing (positive transfer) and in other to writing problems or errors (negative transfer or cross-linguistic interference).

## 6. Measuring writing proficiency

Previous sections of this chapter showed the complexity of the writing processes and skills, especially when it concerns an L2/FL context. The measurement of writing can be influenced by or even be focused on each of these aspects: the use of writing strategies, the application of certain cognitive processes, various quality aspects of the written text, or just the written end product as whole. Furthermore, the purpose of the measurement adds to the complexity of the decisions and choices to be made when designing a writing proficiency assessment.

For example, diagnostic testing induces different requirements than university entrance testing. Diagnostic testing most likely will be focused on specific features of the writing process or the written product. It also implies that the measurement will provide detailed feedback to the writer and/or assessor. This feedback can be based on skill analyses in so-called cognitive diagnostic assessment (Kim, 2011; Xie, 2017). Xie (2017) had ten teachers assess 472 academic essays by checking them on 35 criteria ('Empirically-derived, Descriptor-based Diagnostic assessment, EDD [Kim, 2011]). The 35 criteria are related to –in this case– five presumed subskills: Content fulfillment, Organizational effectiveness, Grammatical knowledge, Vocabulary, and Mechanics. A Q-matrix described the relationship between the criteria and the subskills, that is, if a certain criterion is met the writer gets credit on one or more of the subskills, as described by the Q-matrix. Although the scoring is rather straightforward (checking the criteria as Yes/No), the results provided useful information about the mastery of the subskills and in Xie's study it showed that students' writing performances (at different levels of performance) could well be described in terms of the subskills, and their writing profiles showed the differential mastery of subskills. Content fulfillment and Vocabulary were the most difficult subskills, and Grammatical knowledge and Mechanics were the easier ones. Whereas the advanced writers mastered all five subskills and the beginners were weak in all five, the intermediate writers had (on average) mastered Grammatical

knowledge and Mechanics, but still had to master the other three subskills. This kind of testing links detailed features of the written text, which can be scored in terms of Yes/No-questions, to subskills that may show differential profiles in writing development. This approach allows for more specific feedback than general, holistic scoring, and as such has similarities with analytic scoring. However, analytic scoring is often more impressionistic than checking 35 criteria.

From an L2 learning perspective, measurements that provide five or fewer general scores might still be too crude. If more fine-grained diagnostic information is required, for instance, information about the mastery of specific linguistic structures or specific spellings, the assessor might want to go for discrete-point testing. This could be a dictation test to test spelling skills or a filling the gap test. These formats allow for very specific testing. If the focus is on the spelling of /f/ sounds, like in *fit*, *tough*, or *move*, the sentences in the dictation test can ask for the transcription of just these kinds of words. If the focus is on morpho-syntactic skills like plural formation, the gaps in the gap-filling test could steer to creating plurals. *First, he only saw one mouse, but then he saw two .. mice..*, or *The farmer could afford to buy more than one cow, so he bought both. Now he has two ... cows...* These might not be the most exciting or challenging test items, but most likely they will provide the assessor with the information for his or her purposes. Test instructions for the test-takers can provide additional guidance. These types of discrete-point tests are not authentic in the sense of being representative of everyday writing (cf. Bachman & Palmer, 1996). However, the ‘authentic’ alternative to discrete-point tests in these cases would be to screen pieces of candidates’ writing for the use of words with /f/ sounds or plural nouns. Apart from the fact that it might be a time-consuming analysis, the assessor will remain ignorant about the times the writer might have avoided words with /f/ sounds or sentences with plural nouns, insecure about the correct spellings. In these cases discrete-point testing is far more efficient and on target. However there is a price to pay for this efficiency. Discrete-point tests tap into specific (linguistic) knowledge very explicitly, which means that the test-taker is very much aware that in each test item a linguistic problem needs to be solved or a linguistic choice has to be made, whereas in authentic writing the test-taker might not be aware of potential problems or choices in the text production. In other words, high performance in discrete-point tests does not necessarily mean that the test-taker will use this linguistic knowledge during writing, especially not with young test-takers. The extent to which discrete-point tests are useful and predictive of general writing performance will depend on various factors such as age and proficiency level of the writer, and the linguistic features involved. In a recent study, Gutierrez (2017) established a positive relationship between explicit linguistic knowledge and accurate use during writing in university students. The students’ explicit knowledge of the Spanish subjunctive as measured with discrete-point-like

tests correlated positively with accurate use of this feature in writing (and not in speaking). The relatively slow pace of writing may provide the writer with the opportunity to use this kind of explicit, declarative knowledge that is measured with discrete-point tests. These specific tests may give relevant information for further language instruction, but when used for high-stakes testing, such as admission or certification testing they also may have undesirable washback effects, that is, students may start practicing discrete-point tests instead of performing full or authentic writing tasks.

From the perspective of the evaluation of various writing processes, the assessor might want to focus on just parts of the writing process. For example, the measurement of whether writers are able to structure their texts in an appropriate way, a writing assignment could ask for just that. This means that content generation should be avoided. One way to do that is to provide all candidates with the right content for the fulfillment of the writing assignment. For instance, one could provide the writers with a list of statements (propositions) and ask the test-taking writers to cluster the statements and to order the clusters in the order as the test-taker would like to have it in the text. This assignment zooms in on text organizational planning without asking the test-taker to go through difficult content generation processes or to ask for actual translation and transcription of content. These kinds of tasks are reminiscent of procedural facilitation as described by Bereiter and Scardamalia (1987). They used these kinds of tasks to help beginning writers by lowering the cognitive load through reducing the number of cognitive processes that need to be executed to fulfill the task, and which also allowed the writers to perform optimally on this specific process (see also Graham, MacArthur, & Schwartz, 1995).

Most writing measurements are based on the assessment of real writing performances, that is, writing proficiency as the ability to perform (semi-)authentic writing tasks, and as such writing assessments can be considered performance assessments (McNamara, 1996; Weigle, 2002). Weigle (2002) uses this term “to describe any assessment procedure that involves either the observation of behavior in the real world or a simulation of a real-life activity – i.e. a performance of the ability being assessed, and the evaluation of the performance by raters.” (p. 46). In the case of writing assessment, it most often concerns a ‘simulation of a real-life activity’, more than real world observations. However, in both cases this description illustrates the multi-facetedness of these writing assessments, that is, we need to be concerned about the validity of at least two facets of the measurement: the simulations or writing tasks and the raters doing the evaluation.

The tasks used to elicit the writing performances should represent the kind of writing we intend to measure, that is, the target language use. This leads us back to the construct definition we had in mind for the assessment. Do we aim



at a narrowly defined construct, such as ‘the ability to convey academic information in written form to peers’ or do we aim at a more general construct (definition), such as ‘the ability to express oneself clearly in written form’? (see Weigle [2002] and Deane et al. [2008] for a more extensive discussion of how to define the construct of writing proficiency in the context of writing assessment). The task we administer to candidates should meet various requirements for proper assessment (Bachman & Palmer, 1996; Schoonen, 2011; Weigle, 2002). In the case of writing tasks the rhetorical setting needs to be specified. The test-taker must be informed about the cause and purpose for the writing, the information available, the readership, and also the time allotted. The cause and purpose of the writing will determine the genre and register that is expected in the writing. Does the writer need to convince his or her readers, or just need to explain some phenomena? The intended audience also determines the register, besides the level of information required. Explaining biochemical processes to lay people requires a different kind of content generation compared to explaining these processes to technically-trained graduate students. It will be evident that writing prompts are very important in designing a writing assessment (Ruth & Murphy, 1988; Weigle, 2002). Underspecified writing assignments, such as “Describe your worst nightmare,” should be avoided, because they immediately raise questions like, why would one describe the nightmare, what would be the purpose, and, most importantly, who will read the description? Apart from these questions, writers with healthy and peaceful dreams will fall behind.

In addition, the design of the writing task is expected to influence the language the test taker uses. Some tasks may invite richer language use than others. In SLA literature, the exact effects of task characteristics are much debated and investigated, especially in the context of two opposing hypotheses, that is, Robinson’s (2011) hypothesis about task complexity effects within the Triadic Componential Framework and Skehan and Forster’s (2001) limited-attentional capacity model. It is hypothesized that some complexity characteristics of the task direct attention and others disperse attention, with positive and negative effects on the language use, respectively. However, these effects were originally postulated for speaking performances, and we have seen that despite the commonalities of speaking and writing, there are relevant differences, too. In a meta-analysis of L2 writing research, Johnson (2017) found that also in writing manipulations of the task complexity may affect the syntactic complexity, accuracy, lexical complexity and fluency (CALF) of language learners’ writings. Although the findings of the meta-analysis were still quite heterogeneous and suffered from speech-oriented measures, it seems that in between-groups comparisons providing planning time is beneficial to the syntactic complexity and the accuracy of the writing, and that task familiarity also increases both syntactic and lexical complexity. Manipulation of the



here-and-now dimension in the task demands affected the syntactic complexity, accuracy, and fluency of the writing. This once more underscores the importance of careful task design.

Writing assignments have to specify the rhetorical tasks that have to be fulfilled in order to qualify as a simulation of a real-life activity, but at the same time writers' performances are very sensitive to task parameters that may disperse or direct attention. The question then is: can a single assignment represent the activities that are implied in the construct? From a theoretical perspective one could say that it depends on the construct definition. If this construct definition is very strict and narrow, for example, 'the ability to write short notes for common people on the weather conditions', then a single assignment could represent the construct as defined, although that still could be problematic if the candidate only has the vocabulary for sunny days, and not for thunder and lightning. It is highly hypothetical that assessors are interested in the candidates' 'ability to write short notes for common people on the weather conditions'. When we broaden the construct to a more realistic construct definition, such as 'the ability to write academic texts for a general audience', it becomes clear that it is highly unlikely that a single assignment can represent the construct sufficiently. Psychometric research has shown that it is indeed very difficult to generalize scores from a single assignment to other writing assignments or writing tasks (Gebril, 2009; Schoonen, 2012). Gebril (2009) found fairly low generalizability coefficients for writing performances of 115 Egyptian students, rated by a single rater, that is, .33 for independent writing tasks and .27 for integrated reading-to-writing tasks. Adding more raters to the design hardly improved the generalizability; scores based on a single performance on an independent task, rated by four raters still had a low generalizability of .50 and on a single reading-to-writing tasks the generalizability of .42, i.e., the expected correlation with the scores for the performances on a randomly selected other task, rated by four randomly selected other raters. Gebril's findings are corroborated by similar findings in other studies with other tasks and participants (Schoonen, 2012; Sudweeks, Reeve, & Bradshaw, 2004). To counter this low generalizability of writing assessments it is of paramount importance to include multiple writing assignments in an assessment. The exact number of tasks required depends on the specific assessment conditions and the characteristics of the writing construct measured, but in general terms it would be good to aim at three to five tasks. Research (Schoonen, 2012) suggests that the generalizability of scores for L2 writing is higher than that of L1 writing scores, implying that three tasks could be acceptable for L2 writing assessment in most cases, whereas at least four or five might be needed in L1 writing assessment, all other things being equal, and assessing a well-defined construct of writing proficiency.

In an ideal situation, carefully designed writing prompts or writing tasks generate as a matter of course the rating criteria that should be applied. If the writing assignment requires the writer to persuade the reader who is a school board member, that the introduction of school uniforms in the school system would be a great step to ban inequality and bullying, then the written text should be evaluated as an attempt to convince school board members. The assignment is considered a real-world task and the fulfillment of the task should be evaluated according to real-world criteria (cf. the strong interpretation of performance assessment, McNamara, 1996; Weigle, 2002). The linguistic quality of the text is ancillary to the goal of the writing, i.e., persuasion. Usually, and especially in the context of second language learning, the linguistic characteristics of the text play a larger role in the evaluation of the text, irrespective of whether they help to persuade or not. The language features of the written text are evaluated in their own right, next to content matters (cf. the weak interpretation of performance assessment, McNamara, 1996; Weigle, 2002). Most scoring rubrics fit the weak interpretation of performance assessment. The oft-cited scoring rubric of Jacobs, Zinkgraf, Wormuth, Hartfiel, and Hughey (1981), the 'ESL composition profile', distinguishes between content, organization, vocabulary, language use, and mechanics, each with four score bands representing: 'very poor', 'fair to poor', 'good to average', and 'excellent to very good'. Almost all score bands allow for further discrimination. For example, 'good to average' vocabulary would give the candidate credit points in the range of 17–14. However, a similar level ('good to average') in the mechanics dimension just gives the writer 4 points. This nicely illustrates the weighing of the scores for the five dimensions or characteristics. Excellent content gives up to 30 points, excellent language use 25, excellent organization and vocabulary each 20, and mechanics 5 points at most. In this example scoring rubric, it is not self-evident that language use, for instance, contributes to the persuasion for 25%, and content for just 30%. Still, using these kinds of scoring rubrics, we assume that the writing and the language used is representative of the candidate's language proficiency, and more specifically writing proficiency.

Usually human raters are called in to evaluate the essays. They have to pretend to be the naive school board member that has to be convinced, or they 'just' have to rate the quality of the essay. To avoid idiosyncratic ratings raters are trained to apply the scoring rubrics as they are intended to be used. These can be a scoring rubric as the analytic scheme of Jacobs et al. (1981), or more holistic ones where the rater can give a (single) general impression score (see Weigle [2002] for examples). If the rating scale is focused on the specifics and the goals of the writing task, it may be considered a primary trait rating scale. In that sense, primary trait rating seems to do most justice to McNamara's (1996) strong interpretation of performance

assessment. Holistic rating and primary trait rating provide the assessor with a single score which may be sufficient for certain research or (national) assessment purposes. However, in educational settings more detailed and instructionally relevant scores are more appropriate, and therefore analytic ratings are better in place. These give multiple scores for various aspects of the written product and these subscores can give diagnostic feedback and give cause for remedial actions. For example, L2 student writers may have developed their writing skills in an uneven way, showing good progress in word choice and lexical diversity, but still using rather simple sentence structures that may not get the message across properly, or the student writer might still use the word order of the L1 which may differ from the word order of the L2. This kind of analytic information is both theoretically and educationally relevant. (See also the aforementioned cognitive diagnostic assessment.)

Claims about writing proficiency can only be validly made when we assume that the writers have shown the best of their abilities, and that the raters are able to recognize optimal text quality for example in terms of organization, language use and vocabulary. This assumption comes under pressure when the analytic rating is taken a step further, and texts are evaluated by counting and computing linguistic features of texts. This is basically also what happens in automatic scoring of texts. McNamara, Crossley, and McCarthy (2010) related various linguistic measures of syntactic complexity, lexical diversity, cohesion and lexical sophistication (word frequency) to writing quality. High scores on the linguistic measures –in general– make texts more difficult to understand, but at the same time make them more highly appreciated because of sophisticated language, which is a paradoxical situation. Nevertheless, linguistic measures as the ones mentioned above are often used in research as indices of proficiency and language development (cf. Troia, Shen, & Brendon, 2019), and correlations between linguistic features and overall writing quality are generally low (McNamara et al., 2010; Troia et al., 2019). One reason for these low correlations could be the fact that writing quality is not linearly related to the linguistic features and writers have different approaches to produce a successful text. Crossley, Roscoe, and McNamara (2014) found that there are linguistically different profiles in writing. They distinguished four, which they labeled as: action and depiction style, academic style, accessible style, and lexical style. These different profiles or styles can be described in terms of (a large number of) linguistic features that were derived from various automatic tools for linguistic analysis like *Coh-Metrix* (Crossley et al., 2014).

Another reason for the low correlations between linguistic features and writing quality rating may be more fundamental. It seems that we are dealing with two different kinds of measuring, or, in the words of Cronbach (1961), we are dealing with measures of ability (i.e., writing ability) by inviting writers to produce their best

text versus measures of typical behavior (i.e., use of linguistic features) for which we did *not* push writers to use as many infrequent words as possible or as complex sentences as possible. For measures of ability there is a strive for high scores and test-takers are aware of that, for tests of typical behavior, however, the measurement is unobtrusive in the sense that the test-taker is unaware of the measurement and there is no, and should not be, a strive for higher scores, such as frequent use of uncommon words. First of all, the writer might deliberately decide to use simple and consistent vocabulary, although s/he has control over infrequent, complex and diverse vocabulary. Secondly, the use of maximally diverse and infrequent vocabulary will most certainly not lead to good writing that gets a message easily across. It is obvious that the use of linguistic features for writing assessment is not without its problems, and there is no compelling causal link between the construct of writing proficiency and the outcomes of a linguistic complexity and diversity analysis that is needed to claim validity (cf. Borsboom, Mellenbergh, & Van Heerden, 2004).

## 7. Future (L2) writing and research

In the previous section, we briefly touched upon automatic scoring. Computational tools can easily provide hundreds of analytic scores for a single, relatively short text (Kyle, Crossley, & Berger, 2018). The automatic scoring has strongly developed, also because of corpus analyses of typical language behavior or language use in different genres and by different subpopulations. Semantic and coherence analyses have been added to the lexical and morpho-syntactic analyses which render automatic scoring quite successful and in that respect comparable to the level of human raters (Weigle, 2010).

However, it is not only the rating that has evolved, also the writing has changed substantially. The change from pen to keyboard is most notably, but also the genres have changed. Students still write papers, but in everyday life letters have been replaced by e-mails, SMS-messages and some communicate via tweets with a larger community. These developments will require new approaches to the study, and possibly the assessment, of writing proficiency. To establish those new assessments, new construct definitions are needed.

## References

- Alderson, J. C., Huhta, A., & Nieminen, L. (2016). Characteristics of weak and strong readers in a foreign language. *The Modern Language Journal*, 100(4), 853–879. <https://doi.org/10.1111/modl.12367>
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.
- Becker, A. (2006). A review of writing model research based on cognitive processes. In A. Horning & A. Becker (Eds.), *Revision: History, theory, and practice* (pp. 25–49). Parlor Press.
- Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Lawrence Erlbaum Associates.
- Biber, D. (1991). *Variation across speech and writing*. Cambridge University Press.
- Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review*, 111(4), 1061–1071. <https://doi.org/10.1037/0033-295X.111.4.1061>
- Breetvelt, I., Van den Bergh, H., & Rijlaarsdam, G. (1994). Relations between writing processes and text quality: When and how? *Cognition and Instruction*, 12(2), 103–123. [https://doi.org/10.1207/s1532690xc1202\\_2](https://doi.org/10.1207/s1532690xc1202_2)
- Chenoweth, N. A., & Hayes, J. R. (2001). Fluency in writing: Generating text in L1 and L2. *Written Communication*, 18(1), 80–98. <https://doi.org/10.1177/0741088301018001004>
- Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, 12(11), 671–684. <https://doi.org/10.1037/h0043943>
- Cronbach, L. J. (1961). *Essentials of psychological testing* (2nd ed.). Harper & Row and John Weatherhill.
- Crossley, S. A., Roscoe, R., & McNamara, D. S. (2014). What is successful writing? An investigation into the multiple ways writers can write successful essays. *Written Communication*, 31(2), 184–214. <https://doi.org/10.1177/0741088314526354>
- Cumming, A. (2016). Theoretical orientations to L2 writing. In R. M. Manchon & P. K. Matsuda (Eds.), *Handbook of second and foreign language writing* (pp. 65–88). Walter de Gruyter. <https://doi.org/10.1515/9781614511335-006>
- Deane, P., Odendahl, N., Quinlan, T., Fowles, M., Welsh, C., & Bivens-Tatum, J. (2008). *Cognitive models of writing: Writing proficiency as a complex integrated skill* (ETS Research Report Series, RR-08-55). ETS.
- Fayol, M. (2016). From language to text. The development and learning of translation. In C. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (2nd ed., pp. 130–143). The Guilford Press.
- Flower, L., Hayes, J. R., Carey, L., Schriver, K., & Stratman, J. (1986). Detection, diagnosis, and the strategies of revision. *College Composition and Communication*, 37(1), 16–55. <https://doi.org/10.2307/357381>
- Frangou, S. M., Ruokamo, H., Parviainen, T., & Wikgren, J. (2018). Can you put your finger on it? The effects of writing modality on Finnish students' recollection. *Writing Systems Research*, 10(2), 82–94. <https://doi.org/10.1080/17586801.2018.1536015>
- Gebril, A. (2009). Score generalizability of academic writing tasks: Does one test method fit it all? *Language Testing*, 26(4), 507–531. <https://doi.org/10.1177/0265532209340188>
- Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education*, 7(1), 6–10. <https://doi.org/10.1177/074193258600700104>

- Grabe, W., & Kaplan, R. B. (1996). *Theory and practice of writing: An applied linguistic perspective*. Longman.
- Grabowski, J. (1996). Writing and speaking: Common grounds and differences toward a regulation theory of written language production. In C. M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences and applications* (pp. 73–91). Lawrence Erlbaum Associates.
- Graham, S., MacArthur, C., & Schwartz, S. (1995). Effects of goal setting and procedural facilitation on the revising behavior and writing performance of students with writing and learning problems. *Journal of Educational Psychology*, 87(2), 230–240. <https://doi.org/10.1037/0022-0663.87.2.230>
- Gutiérrez, X. (2017). Explicit knowledge of the Spanish subjunctive and accurate use in discrete-point, oral production, and written production measures. *Canadian Journal of Applied Linguistics/Revue Canadienne de Linguistique Appliquée*, 20(1), 1–30.
- Hayes, J. R. (2012). Modeling and remodeling writing. *Written Communication*, 29(3), 369–388. <https://doi.org/10.1177/0741088312451260>
- Hayes, J. R., & Flower, L. S. (1980). Identifying the organisation of writing processes. In L. W. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing* (pp. 3–30). Lawrence Erlbaum Associates.
- Hayes, J. R., & Nash, J. G. (1996). On the nature of planning in writing. In C. M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 29–55). Lawrence Erlbaum Associates.
- Jacobs, H. L., Zinkgraf, S. A., Wormuth, D. R., Hartfiel, V. F., & Hughey, J. B. (1981). *Testing ESL composition: a practical approach*. Newbury House.
- Johnson, M. D. (2017). Cognitive task complexity and L2 written syntactic complexity, accuracy, lexical complexity, and fluency: A research synthesis and meta-analysis. *Journal of Second Language Writing*, 37, 13–38. <https://doi.org/10.1016/j.jslw.2017.06.001>
- Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grades. *Journal of Educational Psychology*, 80(4), 437–447. <https://doi.org/10.1037/0022-0663.80.4.437>
- Kellogg, R. T., Whiteford, A. P., Turner, C. E., Cahill, M., & Mertens, A. (2013). Working memory in written composition: A progress report. *Journal of Writing Research*, 5(2), 159–190. <https://doi.org/10.17239/jowr-2013.05.02.1>
- Kim, Y.-H. (2011). Diagnosing EAP writing ability using the reduced reparameterized unified model. *Language Testing*, 28(4), 509–541. <https://doi.org/10.1177/0265532211400860>
- Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*, 50(3), 1030–1046. <https://doi.org/10.3758/s13428-017-0924-4>
- Leggette, H. R., Rutherford, T., Dunsford, D., & Costello, L. (2015). A review and evaluation of prominent theories of writing. *Journal of Applied Communications*, 99(3), 37–52. <https://doi.org/10.4148/1051-0834.1056>
- Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. The MIT Press.
- MacArthur, C. A., & Graham, S. (2016). Writing research from a cognitive perspective. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 24–40). The Guilford Press.
- McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication*, 27(1), 57–86. <https://doi.org/10.1177/0741088309351547>

- McNamara, T. F. (1996). *Measuring second language performance*. Longman.
- Murphy, L., & de Larios, J. R. (2010). Searching for words: One strategic use of the mother tongue by advanced Spanish EFL writers. *Journal of Second Language Writing*, 19(2), 61–81. <https://doi.org/10.1016/j.jslw.2010.02.001>
- Robinson, P. (2011). Second language task complexity, the Cognition Hypothesis, language learning, and performance. In P. Robinson (Ed.), *Second language task complexity: Researching the cognition hypothesis of language learning and performance* (pp. 3–38). John Benjamins. <https://doi.org/10.1075/tblt.2.05ch1>
- Roca de Larios, J., Marín, J., & Murphy, L. (2001). A temporal analysis of formulation processes in L1 and L2 writing. *Language Learning*, 51(3), 497–538. <https://doi.org/10.1111/0023-8333.00163>
- Ruth, L., & Murphy, S. (1988). *Designing writing tasks for the assessment of writing*. Ablex.
- Santangelo, T., Harris, K., & Graham, S. (2016). Self-regulation and writing. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research*, (pp. 174–193). The Guilford Press.
- Schleppegrell, M. J. (2004). *The language of schooling: A functional linguistics perspective*. Routledge. <https://doi.org/10.4324/9781410610317>
- Schoonen, R. (2011). How language ability is assessed. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 701–716). Routledge.
- Schoonen, R. (2012). The validity and generalizability of writing scores: The effect of rater, task and language. In E. van Steendam, M. Tillema, G. Rijlaarsdam, & H. van den Bergh (Eds.), *Measuring writing: Recent insights into theory, methodology and practice* (pp. 1–22). Brill.
- Schoonen, R., Snellings, P., Stevenson, M., & Van Gelderen, A. (2009). Towards a blueprint of the foreign language writer: The linguistic and cognitive demands of foreign language writing. In R. M. Manchón (Ed.), *Writing in foreign language contexts: Learning, teaching, and research* (pp. 77–101). Multilingual Matters. <https://doi.org/10.21832/9781847691859-007>
- Schoonen, R., Van Gelderen, A., De Gloppe, K., Hulstijn, J., Simis, A., Snellings, P., & Stevenson, M. (2003). First language and second language writing: The role of linguistic knowledge, speed of processing, and metacognitive knowledge. *Language Learning*, 53(1), 165–202. <https://doi.org/10.1111/1467-9922.00213>
- Skehan, P., & Foster, P. (2001). Cognition and tasks. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 183–205). Cambridge University Press. <https://doi.org/10.1017/CBO9781139524780.009>
- Stevenson, M., Schoonen, R., & De Gloppe, K. (2006). Revising in two languages: A multi-dimensional comparison of online writing revisions in L1 and FL. *Journal of Second Language Writing*, 15(3), 201–233. <https://doi.org/10.1016/j.jslw.2006.06.002>
- Sudweeks, R. R., Reeve, S., & Bradshaw, W. S. (2004). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing*, 9(3), 239–261. <https://doi.org/10.1016/j.asw.2004.11.001>
- Torrance, M. (2016). Understanding planning in text production. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 72–87). The Guilford Press.
- Troia, G. A., Shen, M., & Brandon, D. L. (2019). Multidimensional levels of language writing measures in grades four to six. *Written Communication*, 36(2), 231–266. <https://doi.org/10.1177/0741088318819473>
- Tullock, B. D., & Fernández-Villanueva, M. (2013). The role of previously learned languages in the thought processes of multilingual writers at the Deutsche Schule Barcelona. *Research in the Teaching of English*, 47(4), 420–441.



- Van den Bergh, H., Rijlaarsdam, G., & Van Steendam, E. (2016). Writing process theory: A functional dynamic approach. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 57–71). The Guilford Press.
- Van Vuuren, S., & Berns, J. (2018). Same difference? L1 influence in the use of initial adverbials in English novice writing. *International Review of Applied Linguistics in Language Teaching*, 56(4), 427–461. <https://doi.org/10.1515/iral-2016-0077>
- Victori, M. (1999). An analysis of writing knowledge in EFL composing: A case study of two effective and two less effective writers. *System*, 27(4), 537–555. [https://doi.org/10.1016/S0346-251X\(99\)00049-4](https://doi.org/10.1016/S0346-251X(99)00049-4)
- Wang, W., & Wen, Q. (2002). L1 use in the L2 composing process: An exploratory study of 16 Chinese EFL writers. *Journal of Second Language Writing*, 11(3), 225–246. [https://doi.org/10.1016/S1060-3743\(02\)00084-X](https://doi.org/10.1016/S1060-3743(02)00084-X)
- Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511732997>
- Weigle, S. C. (2010). Validation of automated scores of TOEFL iBT tasks against non-test indicators of writing ability. *Language Testing*, 27(3), 335–353. <https://doi.org/10.1177/0265532210364406>
- Whalen, K., & Ménard, N. (1995). L1 and L2 writers' strategic and linguistic knowledge: A model of multiple-level discourse processing. *Language Learning*, 45(3), 381–418. <https://doi.org/10.1111/j.1467-1770.1995.tb00447.x>
- Woodall, B. R. (2002). Language-switching: Using the first language while writing in a second language. *Journal of Second Language Writing*, 11(1), 7–28. [https://doi.org/10.1016/S1060-3743\(01\)00051-0](https://doi.org/10.1016/S1060-3743(01)00051-0)
- Xie, Q. (2017). Diagnosing university students' academic writing in English: Is cognitive diagnostic modelling the way forward? *Educational Psychology*, 37(1), 26–47. <https://doi.org/10.1080/01443410.2016.1202900>
- Zimmerman, B. J., & Risemberg, R. (1997). Becoming a self-regulated writer: A social cognitive perspective. *Contemporary Educational Psychology*, 22(1), 73–101. <https://doi.org/10.1006/ceps.1997.0919>





## L2 writing and its internal correlates

### A meta-analysis

Masumi Kojima and Taku Kaneta

Gifu City Women's University / Teikyo University of Science

This study examined the overall average correlation between second/foreign language (L2) writing performance and frequently investigated features of writing (i.e., writing-internal correlates). The correlates of L2 writing performance included objective measures of text features (syntactic complexity, lexical complexity, accuracy, fluency, and cohesion) as well as subjective measures (content, argument, organization, cohesion, coherence, grammar, vocabulary, language use, mechanics, and accuracy). A series of moderator analyses were also carried out for each type of objective measure to examine the effects of participants' age, L2 proficiency, learning context, first language (L1) and L2 distance, task type, writing scoring method, and some characteristics of objective measures. By doing so, the present study aimed to identify key correlates of L2 writing performance and compare their relative importance. To this end, a total of 103 retrieved studies contributed 1,045 effect sizes based on 15,537 independent participants. The results for objective measures demonstrated that fluency had the strongest mean correlation with L2 writing performance ( $r = .570$ ), followed by accuracy ( $r = .477$ ), lexical complexity ( $r = .295$ ), syntactic complexity ( $r = .271$ ), and cohesion ( $r = .198$ ). All subjective measure components had strong mean correlations with L2 writing performance ( $r = .668$  to  $.927$ ), but content and language use features had the strongest effects and cohesion and coherence features showed the least effects. Participants' age, learning context, L1–L2 distance, writing scoring method, and some measurement characteristics were found to be significant moderators for certain components. The findings of this study have implications for L2 instruction suggesting that fluency and language accuracy of L2 writing should be promoted across the various developmental stages of L2 learners in various conditions, whereas lexical and syntactic competence should be more focused upon when instructing child/adolescent or low/intermediate L2 writers.

## 1. Introduction

Researchers have increasingly investigated the relationships between second/foreign language (L2) writing proficiency and its component skills (e.g., Oh, Lee, & Moon, 2015; Schoonen et al., 2003; Schoonen, van Gelderen, Stoel, Hulstijn, & de Glopper, 2011). L2 writing ability comprises various dimensions, including linguistic knowledge, discourse knowledge, sociolinguistic knowledge, and strategic competence (Grabe & Kaplan, 1996). To understand the nature of writing ability, it is useful to hypothesize specific writing sub-skills to investigate writing via those sub-skills. This approach, called a component skills approach, has become popular in various areas of L2 performance (e.g., De Jong, Steinel, Florijn, Schoonen, & Hulstijn, 2012; Jeon & Yamashita, 2014; Schoonen et al., 2003; Schoonen et al., 2011), and can help to identify key sub-skills of L2 writing proficiency and compare their relative importance. If significant associations can be found between L2 writing proficiency and its sub-skills, this knowledge will help practitioners improve the effectiveness of future writing instruction, assessment, and research, and will provide researchers with better understanding of the writing ability construct and its component sub-skills.

L2 writing sub-skills have been investigated via specific textual features of the written production (e.g., Bulté & Housen, 2014; Kim & Crossley, 2018; Kyle & Crossley, 2018; Lee, Gentile, & Kantor, 2010; Perkins, 1980; Verspoor, Schmid, & Xu, 2012; Yu, 2010). Here, the underlying assumption is that the textual features show the degree to which linguistic and discourse knowledge of each L2 learner is successfully implemented in the language production (Housen, Kuiken, & Vedder, 2012). The present study synthesizes previous studies that present correlations between overall L2 writing performance and textual features of the written production which include content, argument, organization, cohesion, coherence, grammar, vocabulary, language use, mechanics, complexity, accuracy, and fluency. These textual features are sometimes subjectively rated by human judges, but in cases of quantifiable features (e.g., complexity, accuracy, fluency, and cohesion), objective measures (e.g., frequencies or ratios of target forms per word or per comparable linguistic unit, such as a clause or sentence) are frequently employed. We synthesize both subjective and objective text-based measures, calling them “internal correlates,” and differentiate them from text-external correlates, such as L2 grammar or vocabulary knowledge. We call the latter “external correlates” and discuss them in Chapter 6.

In the literature, L2 writing performance is usually assessed using holistic or analytic (multi-trait) rating scales. In holistic rating, a single score is assigned to a given composition according to the rater’s overall impression of it, without any attempt to isolate and rate specific elements (Weigle, 2002). In analytic scoring, compositions are rated on several aspects of writing or criteria rather than given a

single score. The present study operationally defines overall writing performance in terms of holistic scores or composite scores of analytic scoring based on judgments of human raters.

Unfortunately, the relationship between L2 writing proficiency and writing-internal correlates has remained rather unclear. Even as some studies have reported strong correlations between them, other studies have found only marginal effects at best. As Larsen-Freeman (2009) points out, this is partly because “language performance and development are complex, nonlinear, dynamic, socially situated processes” (p. 588). Given this, she emphasizes the importance of capturing the dynamic interplay between an L2 learner’s performance and the other contexts in which the learner’s data are collected. Besides specific L2 writing skills, variables such as learners’ cognitive development, overall L2 proficiency, first language (L1) and L2 distance, learning environment, task type, writing scoring method, and measurement characteristics (e.g., assessing global complexity or phrasal complexity) are presumed to affect study outcomes. There has been considerable research interest in how these variables affect L2 writing performance and its internal correlates (e.g., Berman & Verhoeven, 2002; Hinkel, 2002; Ortega, 2003; Thewissen, 2013). However, few studies combine these elements in a single study. This is another motivation behind the present study.

We employ meta-analytic techniques to determine the importance of the relative contributions of internal correlates to the overall L2 writing performance. Findings of individual studies are often easily attributable to chance variability as well as the idiosyncratic characteristics of study design, analysis, and research setting (Cooper, 2010; Norris & Ortega, 2000). Thus, to search for answers within a given domain, findings from primary research are best utilized as evidence in secondary research (Norris & Ortega, 2000) rather than taken simply at face value. The meta-analytic method also enables us to investigate the effects of potential moderators of the research results, such as writer’s L1, age, L2 proficiency, learning context, task type, and measurement characteristics. In other words, meta-analytic techniques allow us to examine the characteristics of individual studies and summarize them in moderator analyses, increasing the validity of the overall findings and potentially identifying additional relevant factors.

As regards previous studies, Wolfe-Quintero, Inagaki, and Kim (1998) reviewed 39 L2 writing studies that examined the strength of the relationship between a number of complexity, accuracy, and fluency (CAF) metrics and proficiency levels. Although their study is insightful, the L2 proficiency level was conceptualized in inconsistent ways, including writing evaluation, standardized tests, program levels, school levels, classroom grades, short-term change in classes, and comparisons with native speakers. As the researchers admitted, the constructs of L2 proficiency might have been quite different from one another. In addition, the reviewed studies

employed a variety of statistical tests, such as *t*-tests, ANOVA, and correlations, making their direct comparison difficult. Furthermore, an increasing number of studies have examined the strength of the relationship between L2 writing quality and its textual features including not only CAF but also other discourse features such as cohesion (Kojima, 2020). New CAF indices have been also developed, especially lexical and syntactic indices.

Crossley and colleagues have developed computational tools to measure various linguistic features including CAF and cohesion, which they investigated in relationship with writing quality (e.g., Crossley, Kyle, Allen, Guo, & McNamara, 2014; Crossley & McNamara, 2012; Guo, Crossley, & McNamara, 2013; Kim & Crossley, 2018; Kyle & Crossley, 2016, 2017, 2018). Although their contribution in the L2 writing research field is substantial, L2 writing studies are only a part of their huge research portfolio, which also includes L1 writing and L2 speaking studies, and the same L2 writing samples were repeatedly used in some of these studies. Thus, their studies are not free from chance variability and idiosyncratic characteristics of study design, analysis, and research settings.

Kojima and Kaneta (2020) conducted a meta-analysis synthesizing studies published by 2015 and focused on the relationship between L2 writing performance and objective CAF measures in the texts of English as a second (ESL) or foreign language (EFL) writers who were secondary school students or older. Their main research interest was the comparison of the integrated correlations between ESL/EFL writing performance and objective CAF features in the study texts. They found that fluency had the strongest mean correlation with holistic text quality ( $r = .60$ ) and syntactic complexity had the least trivial effect ( $r = .15$ ). The present study expanded their research scope to add elementary school students as participants and accepted various types of text-based measures assessed both objectively and subjectively and various target languages including English. We also accepted the total scores of analytic scoring dimensions as well as holistic scores as a criterion variable of L2 writing performance and added some potential moderators (e.g., writing scoring method and characteristics of text-based measures). Synthesizing past studies along with the studies of Wolfe-Quintero, Crossley and colleagues and of Kojima and Kaneta, the present study seeks more comprehensive and generalizable results than any previous study.

## 2. Background to the meta-analysis

The present meta-analysis focuses on the relative strengths of association between L2 writing performance and the following five types of objective measures: syntactic complexity, accuracy, lexical complexity, fluency, and cohesion, and subjective

measures assessing rhetorical and idea constructs (content, argument, organization, cohesion, and coherence) and linguistic constructs (grammar, vocabulary, language use, mechanics, and accuracy). These internal correlates of L2 writing performance were initially taken from the literature, and then revised through a literature search undertaken for the present study. Thus, these writing internal correlates have been frequently investigated by previous studies, but not all types of objective and subjective measures were covered. In this section, we will review the writing internal correlates included in the present study. Details for how each variable was coded will be presented in the Method section.

## 2.1 Review of objective measures

### 2.1.1 *Syntactic complexity*

Development of grammar knowledge is one of the central issues in the study of L2 acquisition. In L2 writing research, syntactic features of L2 texts have been intensively investigated as an index of L2 writing proficiency and overall L2 development (e.g., Flahive & Snow, 1980; Kyle & Crossley, 2018; Martínez, 2018; Ortega, 2003; Yang, Lu, & Weigle, 2015). In those studies, syntactic complexity of L2 texts has been widely assessed assuming that L2 learners come to be able to produce more syntactically complex sentences as their language develops.

Syntactic features of L2 writers are sometimes evaluated intuitively by human raters, but in far more cases, they have been assessed using objective measures, such as mean length of production unit (T-units, clauses, or sentences) and amount of subordination. More specific measures are based on frequency of certain grammatical features, such as passive or progressive verb forms, or their ratios per word, clause, or sentence. Traditionally, most popular measures have been T-unit length and ratio of clausal subordination, assuming that increased subordination is typical in advanced writing. However, Biber, Gray, and Poonpon (2011), based on their corpus-based evidence, criticize this practice, arguing that clausal complexification characterizes spoken registers, whereas phrasal elaboration, specifically complex noun phrases, is the main source of syntactic complexity in written academic texts. Norris and Ortega (2009) also argue that development of syntactic complexity proceeds from coordination through subordination to lexical density and more complex phrases. These arguments have been recently supported by several L2 writing studies (Biber, Gray, & Staples, 2016; Kyle & Crossley, 2018).

Research has also shown that syntactic complexity in L2 writing may be affected by various learner- and context-related variables. One such variable is the L1 of L2 learners. Lu and Ai (2015), for example, found that EFL learners with different L1 backgrounds, even for those at comparable proficiency levels, showed different patterns of syntactic complexity in their written texts evaluated using 14 measures.

It can be hypothesized that L1–L2 language distance may affect the strength of correlation between L2 writing performance and syntactic features.

Other variables such as age, learning environment, and writing task may also affect the syntactic complexity of L2 learners (Hinkel, 2002; Hunt, 1965; Ortega, 2003). Ortega (2015) points out that few studies combine these elements in a single study. More studies are needed that investigate moderating effects of various factors on the relationship between L2 writing performance and its syntactic features.

### 2.1.2 *Lexical complexity*

The association between L2 productive vocabulary and writing performance is well established (e.g., Engber, 1995; Kyle & Crossley, 2016; Linnarud, 1986; Yu, 2010). The underlying assumption is that learners' vocabulary use will reflect their lexical knowledge and is closely aligned with writing proficiency levels (Jarvis, 2013; Kim, Crossley, & Kyle, 2018; Kyle & Crossley, 2016; Laufer & Nation, 1995). One of the most commonly used terms for lexical features in L2 texts is lexical richness (i.e., the overall quality of vocabulary found in a language sample; Kim et al., 2018; Read, 2000). There are three main types of lexical richness measures: lexical diversity (i.e., how varied the vocabulary a writer uses), lexical sophistication (i.e., how sophisticated the vocabulary a writer uses; Jarvis, 2013; Kim et al., 2018; Skehan, 2009), and lexical density (i.e., the percentage of content words in the text; Read, 2000). Engber (1995) found significant correlations between L2 learners' lexical diversity and holistic writing scores, but an insignificant result for lexical density measure. Laufer and Nation (1995) observed significant differences in lexical sophistication of written compositions among three proficiency groups.

Recently, Crossley and colleagues have expanded the scope of lexical sophistication beyond frequency of words. They argue that sophisticated words are usually more specific, less concrete, less imageable, less familiar, contextually less diverse, more widely used in academic contexts, have fewer phonological and orthographical neighbors, and elicit slower response time in lexical decision tasks (Kim et al., 2018; Kyle & Crossley, 2016). They also argue that sophisticated lexical items can be extended into multiword units (Crossley, Salsbury, & McNamara, 2015). They have developed computational tools to measure these aspects of lexical sophistication and investigated how these indices correlate with L2 writing performance (e.g., Crossley & McNamara, 2012; Guo et al., 2013; Kyle & Crossley, 2016). These correlations tend to vary from small to large depending on what types of indices are used. Other variables such as learner's age, L1–L2 distance, and task type are presumed to affect the relationship between lexis and L2 learners' writing performance (Jarvis, 2002; Park, 2013). We will discuss these issues later in the Review of moderator variables section.

### 2.1.3 *Accuracy*

Accuracy or correctness refers to the extent to which an L2 learner's performance (and the L2 system that underlies this performance) deviates from a norm (usually the native speaker; Housen et al., 2012; Wolfe-Quintero et al., 1998). Such deviations from the norm are called "errors," and have been traditionally investigated as indicators of L2 language development. According to Housen et al. (2012), accuracy is partly determined by the developmental stage of L2 learners and partly by the degree to which their L2 linguistic knowledge is successfully implemented under the restrictions of cognitive limitations. Widely used accuracy measures include percentage of clauses or T-units which are error free and errors per certain number of words. There are also more specific measures, such as frequency or rate of target-like verbal morphology or target-like plurals.

The association between accuracy and L2 writing proficiency has been widely investigated (e.g., Brodkey & Young, 1981; Lee et al., 2010; Perkins, 1980; Verspoor et al., 2012). Brodkey and Young (1981) reported that the rate of correct usage in L2 compositions correlated positively and significantly with overall writing quality, and successfully discriminated among four very narrow levels of proficiency. Perkins (1980) examined 10 text-based objective measures in L2 texts and found that only those based on absence of errors discriminated among holistic evaluations of the compositions.

However, Pallotti (2009) points out that the gravities of errors differ in terms of their comprehensibility and communicative effectiveness, as well as depending on whether the erroneous structure is developmentally easy or more advanced. Merely counting the number of errors in a given text can obscure these issues. Furthermore, little is known about the extent to which moderating variables such as learner's age, L1, and proficiency level may affect accuracy in L2 writing. We will examine this issue later in the Review of moderator variables section.

### 2.1.4 *Fluency*

It has been repeatedly observed that fluency is a strong indicator of L2 writing proficiency (e.g., Guo et al., 2013; Kojima & Kaneta, 2020; Lee et al., 2010). Fluency is commonly characterized as the ability to produce L2 with native-like rapidity, pausing, hesitation, or reformulation (Housen et al., 2012; Skehan, 1998; Wolfe-Quintero et al., 1998). Housen et al. (2012) state that, whereas syntactic complexity, lexis, and accuracy are considered to relate primarily to the current state of a learner's L2 knowledge and the level of analysis of linguistic information, fluency is primarily related to learners' control over their linguistic L2 knowledge as reflected in the speed and ease of their productions. Schoonen et al. (2011) argued that if writers have reached a high level of fluency in, for example, word retrieval or



sentence construction, these components of writing will not hinder (or will do so to a lesser degree) higher-level processes such as content generation or monitoring of the pragmatic appropriateness of the text. Thus, fluent writers can allot more of their limited attentional resources to higher-level processes of writing, which contributes to better writing quality.

There have been several discussions concerning the operationalization of fluency, or how it can be validly, reliably, and efficiently measured. Usually, fluent writers are expected to compose discourse rapidly, coherently, appropriately, and creatively (Wolfe-Quintero et al., 1998). However, such a characterization encompasses complexity and accuracy as well as discourse-related criteria. Wolfe-Quintero et al. (1998) used the term in a narrower sense to cover only the rate (speed) and length (quantity) of output. In contrast, Norris and Ortega (2009) argued that measures based on length of language unit (e.g., mean length of T-units/clauses) should be considered complexity measures and not fluency measures, based on an empirical study by Oh (2006). Following Norris and Ortega's (2009) view of fluency, we operationalized fluency as amount of output (total number of words, T-units, and sentences produced within an allotted time). Length-based measures were, on the other hand, classified as syntactic complexity measures.

Although fluency is considered a strong indicator of L2 writing performance, correlations have varied from study to study. Variables such as learner's age, L1–L2 distance, and task type are presumed to affect fluency in L2 learners' written production (Bouwer, Béguin, Sanders, & van den Bergh 2015; Jarvis, 2002). We will discuss these issues later in the Review of moderator variables section.

### 2.1.5 *Cohesion*

An increasing number of studies have examined L2 writers' cohesion in relation to their text quality (e.g., Crossley & McNamara, 2012; Crossley, Kyle, and McNamara, 2016a; Kim & Na, 2009; Liu & Braine, 2005). Proficient writers generally produce coherent texts with appropriate cohesive devices and organize discourse and express ideas effectively. Cohesion and coherence are closely related but different concepts. Coherence refers to the relationships that link the meanings of the sentences in a text. These links may be based on the shared knowledge between the writers and the readers, and a link of grammatical or lexical ties may not exist (Halliday & Hasan, 1976). Cohesion refers to the grammatical and/or lexical relationships between the different elements of a text. The most commonly examined cohesive devices include Halliday and Hasan's (1976) five types of cohesive ties: reference, conjunction, substitution, ellipsis, and lexical cohesion. More recently, Crossley, Kyle, and McNamara (2016b) have proposed three different types of cohesion: local cohesion at the sentence level (e.g., connectives and lexical overlap for sentences), global cohesion at the paragraph level (e.g., lexical overlap for paragraphs), and

text cohesion at the entire text level (e.g., word repetition). Crossley et al. (2016a) argue that most L2 cohesion studies have focused on local and text cohesion, but those cohesive ties do not always positively correlate with L2 writing quality. L1 writing studies suggest that the use of local cohesion cues starts at the early stages of writing development, but in university level academic writing, L1 writers seem to depend less on local and text cohesive devices and more on global ones (Crossley et al., 2016b). L2 studies of global cohesion remain scarce, but Crossley et al. (2016a) showed that their L2 learners exhibited growth at the local, global, and text levels of cohesion across a semester-long upper-level English for academic purposes course. The developmental pattern of L2 writers' cohesion stands in need of further clarification.

## 2.2 Review of subjective measures

Several studies have examined discrete skills of L2 writing performance determined by human raters' intuitive judgement. For example, trained raters read each writing sample independently and assign it a holistic score based on discrete features such as content, organization, vocabulary, and cohesion according to a scoring rubric. This method is often employed as a part of analytic scoring of written production. For example, the most widely used analytic scoring scheme, the ESL Composition Profile (Jacobs, Zinkgraf, Wormuth, Hartfiel, & Hughey, 1981) has five discrete dimensions: content, organization, vocabulary, language use, and mechanics. These analytic scores can be more useful for providing diagnostic feedback to L2 learners than holistic scores because different aspects of L2 writing ability usually develop at different rates (Weigle, 2002).

The present study examined 10 constructs assessed by subjective measures. They were classified into two dimensions of L2 writing: rhetorical/idea constructs (content, argument, organization, cohesion, and coherence) and linguistic constructs (grammar, vocabulary, language use, mechanics, and accuracy). These discrete scores have been compared to overall holistic scores of writing samples or composite scores of each analytic rating by examining the contribution of each discrete feature to the overall writing quality. However, the rating dimensions are often highly correlated with each other and with holistic scores or composite scores, and thus Bacha (2001) has stated that its usefulness is limited. Malvern and Richards (2002) also argued that human judgements are often affected by halo effects and are not good at independently assessing particular aspects of L2 performances. As Meara and Bell (2001) stated, objective measures can be used to provide support for human raters' intuitive judgements in writing assessments. Thus, it is worthwhile to examine both objective and subjective measures in their relationship to overall L2 writing performance. If their results are consistent with each other, they will be

robust results. Moreover, some features, such as content of writing, are difficult to capture using objective measures. Examining subjective measures, we can compare the importance of linguistic and rhetorical/idea features of writing.

## 2.3 Review of moderator variables

The present meta-analysis investigated the effects of seven moderating variables theoretically predicted to be relevant. These were participants' age, L2 proficiency, L1–L2 distance, task type, learning context, writing scoring method, and measurement characteristics. We present the rationales for adopting these as moderator variables in this section.

Some other moderating variables also exist in the literature, such as rater background (e.g., experienced vs. novice, native speaker vs. non-native speaker) and rubric type (e.g., content focused vs. language focused). However, none of these proved suitable for our meta-analysis because our preliminary coding indicated that the raters were experienced or trained in most cases, and their first languages were not generally mentioned. As for rubric types, we did not accept rhetorical or linguistic-feature-only rubrics but adopted more balanced ones to minimize the potential heterogeneity among the primary studies.

### 2.3.1 Age

Learners' age is well established as being strongly related to their cognitive and L1 and L2 language development. Age usually has positive relationships with syntactic and lexical features (Berman & Verhoeven, 2002; Hunt, 1965) as well as discourse features (McCutchen & Perfetti, 1982; Yde & Spoelders, 1985). L2 adult writers are cognitively more mature and have considerably richer L1 literacy experiences than younger writers. Thus, adult writers could compensate for weak L2 writing skills with cognitive maturity and L1 literacy skills such as using background knowledge and writing strategies. In addition, a preview of our data indicates that adult learners tend to engage in academic writing, while younger learners focus more on narrative/personal writing. In academic writing, not only grammar and vocabulary knowledge but also various other skills and knowledge bases such as discourse organization skills and content knowledge (Grabe & Kaplan, 1996) are important, presumably making the role of linguistic complexity and accuracy smaller than in non-academic writing. This would weaken the correlations between L2 writing quality and L2 linguistic features of texts for adult writers compared to younger writers. Therefore, we adopted age as a moderator to be investigated in the present study.

### 2.3.2 *L2 proficiency*

Many researchers have been interested in how L2 proficiency plays a role in the development of L2 writing skills (e.g., Biber et al., 2016; Crossley, Salsbury, & McNamara, 2012; Flahive & Snow, 1980; Schoonen et al., 2011; Thewissen, 2013; Wolfe-Quintero et al., 1998). Thewissen (2013) showed that progression in language accuracy is associated with L2 writing proficiency for intermediate L2 learners, but it stabilizes for advanced L2 learners. Several researchers (e.g., Kyle & Crossley, 2018; Norris & Ortega, 2009) have pointed out that popular syntactic complexity measures such as subordination are effective when differentiating among intermediate L2 learners but not among advanced learners. The study of Roca de Larios, Manchón, Murphy, and Marín (2008) demonstrated that their L2 learners spent about 60% of the total composing time on language formulation, and less proficient L2 writers devoted more than 80% of their time to it, leaving them with little time for planning and revision. These studies suggest that insufficient L2 linguistic knowledge and disfluency of producing language would be major issues for non-proficient L2 writers. Thus, linguistic features of texts would explain overall text quality better for low/intermediate learners than for advanced learners. We test this hypothesis in our moderator analyses.

### 2.3.3 *Learning context*

L2 writing researchers (e.g., Ortega, 2003; Sasaki, 2009) have repeatedly observed that learning context greatly influences the nature of L2 writing development. They have pointed out that L2 competence may progress more slowly and develop less fully in foreign-language instructional settings than in second-language ones. The difference might be attributable to motivational factors as well as the amount of L2 exposure in the two environments. In other words, the L2 proficiency of foreign language learners would generally be lower than that of second language learners, and thus their L2 linguistic knowledge would still be developing and explain greater proportion of individual differences in L2 writing performance. It is hypothesized on this basis that linguistic text features would be more important and more strongly associated with L2 writing performance for foreign language learners than for second language learners. Thus, we chose learning context as a potential moderator to be analyzed.

### 2.3.4 *L1–L2 distance*

L2 acquisition research in the last 50 years has revealed a substantial influence of learners' L1 on L2 acquisition (e.g., Koda, 2005; Luk & Shirai, 2009; Murakami & Alexopoulou, 2016). For example, learners' L1 tends to affect their L2 writing, especially in terms of discourse construction features (Hinkel, 2002; Leki, Cumming,

& Silva, 2008). Language errors are also said to be frequently caused by L1 transfer (Leki et al., 2008; Murakami & Alexopoulou, 2016). From these study results, we can speculate that a cross-linguistic transfer effect (Koda, 2005) may mediate L2 writing performance differently depending on the distance between L1 and L2. We hypothesized if L1 and L2 are linguistically similar, it would promote cross-linguistic transfer of writing component skills, reducing individual differences in those component skills, thus resulting in a smaller correlation between L2 writing performance and its internal features. The present meta-analysis investigates moderating effects of learners' L1–L2 distance testing this hypothesis.

### 2.3.5 *Task type*

Previous studies suggest that task type affects text features. For example, certain syntactic and lexical features are commonly associated with formal academic prose (e.g., passive verb forms, third-person pronouns, phrasal complexity; Biber et al., 2011; Kyle & Crossley, 2018; Hinkel, 2002), whereas other features are frequently encountered in personal writing (e.g., first- and second-person pronouns, progressive verb forms; Hinkel, 2002). Different levels of cognitive demand associated with different writing tasks may also cause differences in outputs. For example, Bouwer et al. (2015) reported that the narrative writings of their Iranian EFL learners were more fluent than their academic writings, but in contrast, the latter were syntactically more complex than the former. The taxonomy of academic writing skills, knowledge bases, and processes proposed by Grabe and Kaplan (1996) indicates that not only linguistic knowledge but also discourse knowledge, sociolinguistic knowledge, and knowledge of the world are involved in academic writing. Therefore, we hypothesize that the role of linguistic complexity, accuracy, and fluency would be smaller in academic writing than in narrative/personal writing, but the role of rhetorical features such as cohesion would be greater in academic writing since academic texts are expected to be logical and cohesive. Thus, we adopted task type as a moderator to examine this hypothesis.

### 2.3.6 *Writing scoring method*

In the literature, two main types of writing scoring methods have been used to assess overall L2 text quality: holistic scoring and analytic scoring. Holistic scoring involves assigning a single score to a text based on the overall impression of the text (Weigle, 2002). In some cases, criteria for scoring are not explicitly stated, but in far more cases, a scoring rubric is employed, and raters are trained to adhere to the rubric when scoring. An example of a holistic scoring rubric in L2 is the one used for the TOEFL iBT independent writing text, formerly known as the Test of Written English (TWE). These holistic scoring rubrics usually contain descriptors

of both linguistic and rhetorical qualities of L2 compositions. Perkins (1983) stated that “of all of the composition evaluation schemes available today, holistic scoring has the highest construct validity when overall attained writing proficiency is the construct to be assessed” (p. 652).

In analytic scoring, scripts are rated on several aspects of writing or criteria, rather than being given a single score according to the overall quality of a text (Weigle, 2002). The ESL Composition Profile (Jacobs et al., 1981) is one of the most widely used analytic scales in L2 writing. The primary advantage of an analytic scoring scheme over a holistic one is that it provides more useful diagnostic information about L2 learners’ writing ability (Weigle, 2002).

Authenticity is usually considered higher for holistic scoring than for analytic scoring. For example, Perkins (1983) pointed out the drawbacks of analytic scoring, saying that the features to be analyzed are predetermined, isolated from context, and scored separately. In addition, the scoring weight of a particular category is fixed and cannot be adjusted for different texts. White (1984) also argued that holistic scoring evaluates a whole text rather than simply parts of a text, thus reflecting most closely the authentic, personal reaction of a reader to a text. However, Weigle (2002) stated that experienced raters may read a composition holistically and adjust analytic scores to match what they expect the total score to come out to be. If so, the correlations between holistic scores or total scores of analytic dimensions on the one hand and text features on the other hand would not be significantly different. We will examine this hypothesis in our moderator analyses.

### 2.3.7 *Measurement types of complexity and fluency*

As we discussed in the reviews of syntactic and lexical complexity and fluency, we differentiated three types of syntactic complexity measures: global, phrasal, and specific measures; three types of lexical complexity measures: lexical diversity, sophistication, and density; and two types of fluency measures: word-based and unit-based. Although we discussed three types of cohesion measures (i.e., local, global, and text cohesion), because most of the primary studies have focused on local cohesion, we did not analyze cohesion measurement type as a moderator.

### 3. Research questions for the meta-analysis

The following research questions were investigated by the present study.

1. What are the relative strengths of association between discourse-level, communicative L2 writing performance and the following five constructs assessed by objective measures: syntactic complexity, lexical complexity, accuracy, fluency, and cohesion, and 10 constructs assessed by subjective measures: content, argument, cohesion, coherence, organization, grammar, vocabulary, language use, mechanics, and accuracy?
2. Do potential moderators such as participants' age, L2 proficiency, learning context, task type, L1–L2 distance, and writing scoring method and measurement characteristics systematically influence the relationship between L2 writing performance and its internal correlates?

### 4. Method

#### 4.1 Literature search and inclusion criteria

The literature search was conducted several times between April 2018 and August 2020. During this time, the previous search of Kojima and Kaneta (2020) was updated, and the scope of the study was expanded. The literature search for the present study was conducted in two ways. First, we electronically and manually examined articles in the 24 most relevant journals in the field of applied linguistics, education, L1 and L2 writing, and literacy studies. These were *Annual Review of Applied Linguistics*, *Language Teaching*, *Studies in Second Language Acquisition*, *International Review of Applied Linguistics in Language Teaching*, *Assessing Writing*, *English for Specific Purposes*, *Journal of English for Academic Purposes*, *Journal of Second Language Writing*, *System*, *Applied Linguistics*, *ELT Journal*, *Language Testing*, *Language Teaching Research*, *RELC Journal*, *Second Language Research*, *Language Assessment Quarterly*, *TESOL Quarterly*, *Modern Language Journal*, *Foreign Language Annals*, *Language Learning*, *International Journal of Applied Linguistics*, *Language Learning & Technology*, *Applied Language Learning*, and *Canadian Modern Language Review*. In this stage, we did not specify any specific internal or external correlates, but just searched for studies that reported correlations between overall writing performance and at least one variable. These variables included various text features, test scores, and scales of questionnaires. They were then classified into writing internal or external correlates. Second, we used seven electronic databases (ERIC, LLBA, ProQuest Central, ProQuest D&T, Web



of Science, MLA International Bibliography, and Google Scholar) to locate relevant studies for inclusion in the meta-analysis. In this stage, we included both refereed and non-refereed (e.g., research reports, conference papers) articles. Studies published by August 2018 were searched using various combinations of key terms related to L2 writing, complexity, accuracy, fluency, cohesion, and rhetorical features. These key terms were decided based on the initial search of the 24 journals, the thesauruses supplied in the databases, books, meta-analyses, narrative reviews in the field, and the authors' experiences.

Through the journal and database search, when abstracts indicated that the articles might be relevant to the current research synthesis, full texts were retrieved for further examination. As a result, we obtained 311 full study reports in addition to the 356 study reports collected by Kojima and Kaneta (2020); these 311 studies were further examined and the originally collected 356 studies were reexamined to check their eligibility. To be included in the synthesis, a study had to meet all the following criteria:

1. The participants were L2 learners. If the participants were L1 and L2 writers mixed, and if their data were not separately analyzed, the study was excluded.
2. The participants were engaged in an L2 writing task as an act of communication, such as writing an academic, narrative, or personal essay; describing a picture; or responding to a film. If they were just asked to rewrite an existing passage in simplified form, or translate L1 into L2, the study was excluded. Phrase- or sentence-level writing was also excluded. Cooperative writing tasks and integrated tasks, such as reading-listening-writing tasks, were also excluded, because we focus on component skills of L2 writing and their individual differences.
3. The study employed holistic scores of overall writing quality or composite scores of analytic rating dimensions, which included both linguistic and rhetorical features.
4. The study reported correlations between overall writing quality scores (i.e., holistic scores or the total scores of analytic scoring dimensions) and at least one writing-internal variable.
5. The study was published in English.

After finalizing the initial list of eligible studies, additional studies were also found through citation chasing using the initial list. In this stage, non-electronic papers, such as empirical studies published as book chapters and monographs, were also included. As a result, an additional 87 studies, both refereed and non-refereed, were retrieved and assessed for eligibility.



After screening, 126 studies initially met our criteria. Of these, four pairs of studies (Barkaoui, 2010a, 2010b; Engber, 1992, 1995; Kroll, 1982, 1990; Lee, Gentile, & Kantor, 2008, 2010) reported duplicate data; we excluded one of each of these pairs (i.e., four studies). In addition, 19 studies were excluded because their targeted text-based features were relatively rare and only few studies were available (i.e., less than 3 studies) for an effect size aggregation with them (for more information on the excluded studies, see Kojima, 2020). Ultimately, the sample consisted of 103 studies (see Appendix A), including 52 studies synthesized in Kojima and Kaneta (2020). These studies provided a total of 15,537 independent participants and 1,045 correlations between L2 writing performance and objective/subjective measures.

#### 4.2 Acceptable measures of L2 writing performance and internal correlates

As we have already elaborated on the construction of writing performance and internal correlates to be examined in the present study, we describe their acceptable measures only briefly in Appendix B.

#### 4.3 Coding the primary studies

Of the 103 studies, 52 were already coded by Kojima and Kaneta (2020). Twelve of the remaining 51 studies (23.5%) were independently coded for various features (e.g., sample size; participants' L1, L2, age, proficiency, learning context, task types, writing scoring method; measurement characteristics; and statistical information to be analyzed) by the first and second authors, both of whom are experienced L2 writing researchers. Inter-coder reliability (percentage agreement) for the coding ranged from 92% to 100% across all the coded variables. Four trained research assistants then coded the remaining 39 studies, and both the first and second authors double-checked all the coding and made necessary corrections. Any disagreements were discussed and resolved.

Table 1 summarizes the chosen moderator variables and the codes for their sub-categories. We employed two- or three-level codes for each moderator, both because in many cases detailed information was not available in the primary studies and to maintain adequate power for the intended moderator analysis. Studies that did not report sufficient information for coding were excluded from the moderator analysis.

Here are some specific features of the coding that may be of relevance. Regarding L1, in many primary studies, participants' ethnicity and/or locations of data collection were explicitly stated, but participants' L1s were not. To maximize the number of studies submitted for moderator analysis while maintaining adequate relevance, we utilized the available information. For example, if participants were stated to be

**Table 1.** Coding table for moderators

Variables	Codes	Examples
Age	1. Child/ Adolescent	Primary/secondary schoolers; average age under 18
	2. Adult	Undergraduate/graduate students; average age 18 or above
L1–L2 distance	1. Shorter	Both Indo-European
	2. Longer	Combinations of Indo-European and non-Indo-European
Proficiency	1. Low/ Intermediate	A1 to B1 in CEFR; under 530 in TOEFL PBT; All child/ adolescent learners in foreign language contexts
	2. Advanced	B2 to C2 in CEFR; 530 or above in TOEFL PBT
Learning context	1. Second language (SL)	Minority language speakers; international students
	2. Foreign language (FL)	Foreign language learners
Writing task	1. Academic	Argumentative/persuasive/expository/thesis writing
	2. Narrative/ Personal	Describing pictures/silent films/personal experiences/ personal tastes
Writing scoring method	1. Composite scores of analytic scoring	ESL Composition Profile (Jacobs et al., 1981)
	2. Holistic scoring	TOEFL TWE
Fluency measures	1. Word-based	Words per minute/text; number of content words
	2. Language-unit- based	Number of sentences/clauses/T-units
Syntactic complexity measures	1. Global complexity	Mean length of sentence/T-unit/clause; mean number of clauses per T-unit
	2. Phrasal complexity	Mean length of noun phrase; preposition per nominal
	3. Specific measures	Frequencies/ratios of passives/articles/to infinitive/third person -s
Lexical complexity measures	1. Diversity	TTR, type, Guiraud index, D, MTLD, HD-D
	2. Sophistication	Advanced word ratio; mean length of words; LFP (Laufer & Nation, 1995); number of multi-word units; word familiarity/concreteness/imageability
	3. Density	Lexical words per words; nouns per words; verbs per words

“Chinese students,” their L1 was assumed Chinese. If a study was conducted in a foreign language environment, the dominant language of the area was coded for the participants’ L1. If a study examined L2 learners of mixed L1s and did not report separate correlations for each L1 group, they were not included in the moderator analysis. Mixed groups of other moderators (age, L2 proficiency, learning context, writing task) were also excluded from moderator analyses. Regarding proficiency, authors’ description of participants’ proficiency and/or scores of standardized tests were used for coding. For example, if participants scored 530 or higher for TOEFL PBT or equivalently in other standardized tests, they were coded as Advanced, and if they scored lower than that, they were coded as Low/Intermediate. All child/adolescent learners in foreign language contexts were coded as Low/Intermediate because their proficiencies are generally low and their standardized test scores were rarely reported. As for international students in tertiary education, we checked their institutions’ admission policies and if L2 learners were required to score 530 or higher for TOEFL PBT or equivalently in other standardized tests, they were coded as Advanced, if not, they were treated as a mixed/unknown proficiency group.

#### 4.4 Research synthesis

Initially, we coded the intact correlation coefficient values as reported by the primary researchers. Among them, some values were negative due to the nature of the measures (i.e., based on errors instead of accuracy; ratio of basic vocabulary instead of sophisticated vocabulary; some psychometric indices of Coh-Metrix and TAALES, such as word frequency, word familiarity, word meaningfulness, word concreteness, word imageability, word polysemy values). In such cases, inversed values were calculated before aggregation in the meta-analysis. For a longitudinal study that reported data collected at multiple time points, only the data from the first time point were included in the meta-analysis in order to avoid the effect of time-varying irrelevant variables. Some studies reported partly duplicative data. For example, in some cases, data for the whole group of the participants as well as for each sub-group were reported. In such a case, we used the data derived from the whole group only, unless those sub-groups were classified by our hypothesized moderators (e.g., L2 proficiency).

When a primary study reported correlations between L2 writing and various measures of a construct (e.g., several syntactic complexity measures), we did not average the correlations but treated them as multilevel data and conducted meta-analyses via multilevel linear mixed-effects models using R (version 3.6.3) and its metafor package (version 2.1-0) (Viechtbauer, 2010). We treated each effect size nested within a sample, which was in turn nested within a study. When the same participants were involved in different studies, they were given the same study and

sample IDs in the analyses (i.e., one set of IDs for Evola, Mamer, & Lentz [1980] & Kaczmarek [1980]; another set of IDs for Crossley et al. [2014]; Guo et al. [2013]; Kim & Crossley [2018], and Kyle & Crossley [2016, 2017, 2018]). In this way, we addressed the issues of effect size dependency (multiple effect sizes from a single study) more appropriately than a fixed-effects or random-effects model. The statistical analyses included effect size aggregation for the association between L2 writing performance and each of the 15 constructs (i.e., 5 constructs assessed by objective measures and 10 constructs assessed by subjective measures). Fisher's  $Z$  values converted from the correlation coefficients retrieved from the primary studies were weighted by the inverse-variance method using the metafor package. The weighted  $Z$  values were then utilized to estimate an average weighted effect size and its 95% confidence interval (CI) in each aggregation. Publication or availability bias was examined using a funnel plot which displays the distribution of effect sizes obtained from primary studies. We also employed fail-safe  $N$  (Rosenthal, 1979), which suggests the number of non-significant studies necessary to make the result non-significant. A  $Q$  test was employed to determine heterogeneity for each aggregation. If the  $Q$  value was significant, it indicated that possible moderators were not considered in the model. We then conducted a series of moderator analyses, entering hypothesized moderators in turn into the initial model. A  $Q_M$  test was served to probe whether each hypothesized moderator was statistically significant. Subset differences in a moderator variable with three levels were examined by changing the reference level in turn. The significance level was specified as  $p < .05$  throughout the analyses. Each integrated  $Z$ -value and its 95% CI were transformed back to  $r$  when reporting the result. When interpreting, we followed Plonsky and Oswald (2014) and considered correlations of .25, .40, and .60 to indicate small, medium, and large effects, respectively.

## 5. Results

Table 2 shows the results of aggregated correlations between L2 writing performance and objective/subjective measures and their related statistics. Overall, the results showed no significant evidence of a file-drawer problem. All the values of fail-safe  $N$  exceeded Rosenthal's (1979)  $5k + 10$  criterion, where  $k$  indicates the number of included studies (number of correlations in our case). All the heterogeneity tests (i.e.,  $Q$  tests) were significant suggesting possible moderators were not considered in the models. However, we conducted moderator analyses only for objective measures because the number of studies with subjective measures was relatively small (i.e., below 15 studies in each construct), in order to maintain adequate statistic power for the analyses.

Table 2. L2 writing and its internal correlates: Overall, component, and subcomponent analyses

Component	Subcomponent	No. of studies	No. of independent participants	No. of dependent participants	No. of rs	R	95%	CI	Min r	Max r	Test for moderators (QM)	Fair-safe <sup>a</sup>
<i>Objective Measures</i>												
Syntactic Complexity	–	32	6,539	74,319	245	.271***	.125	.405	–.441	.970		226,225
	Age	30	5,641	50,235	185				–.423	.970	2.132	107,906
	Age: Adult	24	4,806	40,757	147	.217*	.041	.380	–.423	.970		
	Age: Child/Adolescent	6	835	9,478	38	.477**	.160	.705	–.003	.930		
	L1–L2 distance	13	3,792	38,287	113				–.217	.753	4.705*	79,174
	L1–L2: Longer	7	2,852	28,615	73	.173**	.063	.278	–.217	.680		
	L1–L2: Shorter	6	940	9,672	40	.346***	.230	.452	–.003	.753		
	Proficiency	16	1,862	15,005	92				–.150	.740	1.540	32,522
	Proficiency: Low/Intermediate	6	968	9,816	44	.235***	.111	.352	–.110	.730		
	Proficiency: Advanced	11	894	5,189	48	.141**	.034	.244	–.150	.740		
	Context	29	5,205	45,101	187				–.217	.970	7.228**	139,327
	Context: SL	17	1,544	7,049	76	.112	–.088	.303	–.190	.740		
	Context: FL	12	3,661	38,052	111	.490***	.292	.648	–.217	.970		
	Task	25	5,771	67,078	210				–.441	.970	0.002	157,078
	Task: Academic	18	4,903	58,230	164	.298**	.108	.467	–.441	.753		
	Task: Narrative	9	904	8,848	46	.301**	.088	.488	–.160	.970		
	Scoring	31	6,439	74,119	243				–.441	.970	8.601**	218,346
	Scoring: Analytic	7	1,851	23,610	46	.570***	.342	.735	–.217	.970		
	Scoring: Holistic	24	4,588	50,509	197	.151	–.006	.300	–.441	.740		
	Measure	32	6,539	74,319	245				–.441	.970	35.502***	226,225
	Measure: Global	29	6,298	18,096	110	.289***	.143	.423	–.339	.930		
	Measure: Phrasal	9	4,246	14,623	38	.218**	.066	.359	–.423	.460		
	Measure: Specific	17	4,478	41,600	97	.248**	.099	.386	–.441	.970		

(continued)

Table 2. (continued)

Component	Subcomponent	No. of studies	No. of independent participants	No. of dependent participants	No. of rs	R	95%	CI	Min r	Max r	Test for moderators (Q <sub>M</sub> )	Fail-safe N <sup>a</sup>
Lexical Complexity	–	37	7,196	73,264	282	.295***	.234	.354	–.657	.853		459,150
	Age	34	4,764	44,841	230				–.657	.853	7.033**	222,288
	Age: Adult	28	3,638	31,023	155	.258***	.186	.327	–.657	.769		
	Age: Child/Adolescent	7	1,126	13,818	75	.435***	.325	.534	–.366	.853		
	L1–L2	23	3,790	25,168	172				–.657	.853	1.365	97,170
	L1–L2: Longer	17	2,842	11,803	97	.267***	.177	.353	–.657	.771		
	L1–L2: Shorter	8	948	13,365	75	.352***	.227	.465	–.366	.853		
	Proficiency	17	2,672	24,608	108				–.366	.853	2.630	46,848
	Proficiency: Low/Intermediate	8	949	11,595	71	.356***	.231	.469	–.366	.853		
	Proficiency: Advanced	10	1,723	13,013	37	.201**	.053	.341	–.180	.602		
	Context	34	5,862	39,370	218				–.657	.853	1.694	170,245
	Context: SL	12	979	2,905	38	.250***	.140	.355	–.180	.577		
	Context: FL	22	4,883	36,465	180	.335***	.264	.401	–.657	.853		
	Task	33	5,339	59,340	260				–.657	.853	1.449	376,750
	Task: Academic	25	4,420	46,194	171	.295***	.225	.362	–.602	.680		
	Task: Narrative	11	1,027	13,146	89	.351***	.260	.436	–.657	.853		
	Scoring	37	7,196	73,264	282				–.657	.853	1.037	459,150
	Scoring: Analytic	9	527	1,520	30	.352***	.224	.468	–.340	.830		
	Scoring: Holistic	29	6,699	71,744	252	.281***	.213	.347	–.657	.853		
	Measure	37	7,196	73,264	282				–.657	.853	26.858***	459,150
	Measure: Diversity	26	5,278	14,214	92	.285***	.224	.343	–.657	.853		
	Measure: Sophistication	29	6,473	58,334	177	.317***	.259	.374	–.366	.830		
	Measure: Density	6	484	716	13	.134*	.031	.234	–.200	.309		

(continued)

Table 2. (continued)

Component	Subcomponent	No. of studies	No. of independent participants	No. of dependent participants	No. of rs	R	95%	CI	Min <i>r</i>	Max <i>r</i>	Test for moderators ( <i>Q<sub>M</sub></i> )	Fail-safe <i>N</i> <sup>a</sup>
Accuracy	–	27	2,738	10,508	128	.477***	.373	.570	.028	.930		82,530
	Age	26	2,689	10,459	127				.028	.930	1.174	80,129
	Age: Adult	20	2,018	4,961	60	.431***	.306	.542	.028	.866		
	Age: Child/Adolescent	6	671	5,498	67	.555***	.352	.708	.068	.930		
	L1–L2 distance	12	1,488	5,487	41				.028	.866	0.478	8,310
	L1–L2: Longer	6	701	1,071	14	.433**	.178	.633	.028	.770		
	L1–L2: Shorter	6	787	4,416	27	.540***	.311	.710	.068	.866		
	Proficiency	18	1,807	7,257	58				.043	.819	1.062	12,005
	Proficiency: Low/Intermediate	9	1,127	4,756	32	.457***	.333	.566	.068	.819		
	Proficiency: Advanced	10	680	2,501	26	.381***	.249	.499	.043	.659		
	Context	27	2,738	10,508	128				.028	.930	1.524	82,530
	Context: SL	14	1,203	4,927	85	.415***	.260	.550	.043	.830		
	Context: FL	13	1,535	5,581	43	.539***	.396	.656	.028	.930		
	Task	23	2,554	10,240	122				.028	.930	0.325	76,562
	Task: Academic	14	1,428	3,814	53	.467***	.344	.574	.043	.819		
	Task: Narrative	11	1,154	6,426	69	.490***	.367	.597	.028	.930		
	Scoring	27	2,738	10,508	128				.028	.930	3.850*	82,530
Fluency	Scoring: Analytic	3	167	498	8	.698***	.461	.843	.068	.930		
	Scoring: Holistic	24	2,571	10,010	120	.443***	.333	.541	.028	.866		
	–	30	6,310	9,866	73	.570***	.463	.661	–.060	.930		67,991
	Age	26	4,203	6,922	65				–.060	.930	1.174	34,666
	Age: Adult	18	2,967	5,279	39	.595***	.465	.700	–.060	.930		
	Age: Child/Adolescent	9	1,236	1,643	26	.488***	.282	.650	.095	.880		
	L1–L2 distance	18	3,012	4,777	41				–.060	.880	0.642	16,368
	L1–L2: Longer	14	2,560	4,203	26	.567***	.418	.686	–.060	.880		
	L1–L2: Shorter	6	452	574	15	.497***	.285	.662	.174	.859		

(continued)

Table 2. (continued)

Component	Subcomponent	No. of studies	No. of independent participants	No. of dependent participants	No. of rs	R	95% CI	Min r	Max r	Test for moderators (Q <sub>M</sub> )	Fail-safe N <sup>a</sup>
Cohesion	Proficiency	12	766	982	24			-.060	.930	0.170	3,492
	Proficiency: Low/Intermediate	6	370	483	15	.629***	.310 .821	-.060	.880		
	Proficiency: Advanced	6	396	499	9	.547**	.182 .779	.147	.930		
	Context	26	3,970	6,456	65			-.060	.930	0.117	37,291
	Context: SL	11	1,119	1,840	27	.594***	.417 .728	.170	.930		
	Context: FL	15	2,851	4,616	38	.557***	.405 .680	-.060	.880		
	Task	28	5,832	9,388	71			-.060	.930	0.001	65,008
	Task: Academic	19	4,963	7,892	36	.583***	.465 .680	-.060	.930		
	Task: Narrative	11	897	1,496	35	.585***	.451 .693	.147	.859		
	Scoring	30	6,310	9,866	73			-.060	.930	0.215	67,991
	Scoring: Analytic	5	325	385	7	.511**	.178 .739	.147	.780		
	Scoring: Holistic	25	5,985	9,481	66	.580***	.463 .677	-.060	.930		
	Measure	28	6,102	9,658	70			-.060	.930	137.562***	59,940
	Measure: Word	28	5,782	6,918	55	.590***	.490 .674	-.060	.930		
	Measure: Unit	8	2,342	2,740	15	.325***	.186 .451	.154	.780		
	–	21	3,568	47,381	201	.198***	.096 .296	-.402	.821		8,329
	Age	20	3,088	41,621	187			-.402	.821	2.062	7,306
	Age: Adult	19	2,744	39,213	180	.225***	.120 .325	-.402	.821		
	Age: Child/Adolescent	1	344	2,408	7	-.122	-.530 .331	-.279	.124		
	L1–L2 distance	15	2,068	12,388	100			-.279	.821	11.573***	7,612
	L1–L2: Longer	12	1,852	11,856	89	.170**	.059 .277	-.279	.821		
	L1–L2: Shorter	3	216	532	11	.548***	.368 .688	.120	.756		
	Proficiency	6	301	3,878	75			-.260	.821	0.033	3,108
	Proficiency: Low/Intermediate	2	59	358	14	.162	-.143 .440	-.193	.556		
	Proficiency: Advanced	5	242	3,520	61	.193	-.005 .377	-.260	.821		
	Context	18	2,234	15,185	144			-.279	.821	3.327	8,586

(continued)



Table 2. (continued)

Component	Subcomponent	No. of studies	No. of independent participants	No. of dependent participants	No. of rs	R	95%	CI	Min <i>r</i>	Max <i>r</i>	Test for moderators ( <i>Q<sub>M</sub></i> )	Fail-safe <i>N</i> <sup>a</sup>
	Context: SL	5	363	4,942	66	.053	-.163	.265	-.260	.590		
	Context: FL	13	1,871	10,243	78	.281***	.159	.394	-.279	.821		
	Task	19	3,174	44,923	193				-.402	.821	0.050	9,906
	Task: Academic	16	3,020	44,239	179	.213***	.095	.325	-.402	.821		
	Task: Narrative	3	154	684	14	.179	-.105	.436	-.080	.750		
	Scoring	20	3,468	47,081	198				-.402	.821	0.168	6,101
	Scoring: Analytic	5	411	4,868	63	.136	-.059	.321	-.260	.590		
	Scoring: Holistic	15	3,057	42,213	135	.182**	.071	.289	-.402	.821		
<i>Subjective Measures</i>												
Content	–	8	788	915	10	.927***	.892	.951	.868	.970		6,357
Argument	–	4	401	591	8	.841***	.659	.930	.610	.960		2,094
Coherence	–	3	692	692	3	.668***	.450	.811	.472	.770		325
Cohesion	–	3	243	491	11	.688**	.259	.891	.310	.890		813
Organization	–	14	2,405	2,722	21	.878***	.812	.921	.460	.964		27,011
Grammar	–	9	2,708	4,787	15	.837***	.737	.901	.500	.953		30,204
Vocabulary	–	13	2,271	3,231	17	.888***	.825	.929	.510	.964		22,262
Language use	–	7	679	806	9	.920***	.862	.954	.809	.990		4,142
Mechanics	–	9	1,758	1,905	13	.766***	.524	.894	-.830	.932		7,663
Accuracy	–	4	581	1,131	9	.781**	.361	.938	.350	.967		2,436

*Note*\*  $p < .05$ .\*\*  $p < .01$ .\*\*\*  $p < .001$ .a. Fail-safe *N* (Rosenthal, 1979). CI = Confidence interval. SL = Second language. FL = Foreign language.

## 6. Discussion

The primary goal of this meta-analysis, as formulated in our first research question, was to quantify the typical strength of the associations between the overall L2 writing performance and its writing-internal correlates (objective/subjective measures). The meta-analytic methodology allows us to benchmark how strong these relationships are on average. All objective and subjective measures had significant correlations with L2 writing performance. Among objective measures, fluency had the strongest average correlation with L2 writing performance ( $r = .570$  [95% CI: .463, .661]), followed by accuracy ( $r = .477$  [.373, .570]) and syntactic and lexical complexity ( $r = .271$  [.125, .405] and  $r = .295$  [.234, .354], respectively). Cohesion had the weakest and most trivial effect on L2 writing performance ( $r = .198$  [.096, .296]). Examination of 95% CIs revealed that fluency had a significantly stronger effect on L2 writing quality than syntactic/lexical complexity and cohesion, and accuracy had a significantly stronger mean correlation than lexical complexity and cohesion. In contrast, all subjective measures had strong average correlations with overall L2 writing performance, which ranged from  $r = .668$  for coherence to  $r = .927$  for content. We will further discuss the results for each correlate below, including the results of moderator analyses as formulated in the second research question.

### 6.1 Objective measures

#### 6.1.1 *Syntactic complexity*

Syntactic complexity had a significant but small mean correlation with overall L2 writing performance ( $r = .271$  [.125, .405]). This means that syntactic complexity explains only 7.34% of the variance caused by individual differences in overall L2 writing performance. This result might be surprising because syntactic complexity measures have frequently been used as indices of L2 learners' overall writing proficiency in the literature (see Ortega, 2003). However, the integrated mean correlations were medium when writers were children or adolescents ( $r = .477$ ), were foreign language learners ( $r = .490$ ), or when L2 writing performance was assessed by the composite scores of analytic scoring dimensions ( $r = .570$ ). The effect of syntactic complexity on L2 text quality was substantial in these cases. Thus, moderator variables should be considered. We will discuss each of our hypothesized moderators in turn.

L2 learning context was a significant moderator, and the integrated mean correlation between syntactic complexity and overall L2 writing performance was stronger for foreign language learners ( $r = .490$ ) than for second language learners

( $r = .112$ ). This result supported our hypothesis, which assumed L2 proficiency of foreign language learners would generally be lower than that of second language learners, and thus their L2 syntactic knowledge would still be developing and explain greater proportion of individual differences in L2 writing performance.

In contrast, participants' age and L2 proficiency were not significant moderators. We hypothesized that syntactic complexity would be more important and more strongly associated with L2 writing quality for child/adolescent writers than for adult writers, and for low/intermediate L2 learners than for advanced ones. In fact, the average correlation between syntactic complexity and overall L2 writing performance was stronger for child/adolescent learners ( $r = .477$ ) than for adult learners ( $r = .217$ ), and for low/intermediate L2 learners ( $r = .235$ ) than for advanced learners ( $r = .141$ ). However, these differences were statistically insignificant. These null results seem to be attributable to the wide ranges of the 95% CI for the younger and low/intermediate learners, mainly caused by the small number of studies. In fact, the difference between these null results and the significant result of the learning context-based moderator effect can be explained by the number of studies involved (i.e., 17 studies for the second language group, 12 studies for the foreign language group, and six studies for each of the lower proficiency and younger groups). As Kojima and Kaneta (2020) pointed out, far more L2 writing researchers have investigated adult L2 writers (i.e., undergraduate and postgraduate students) than younger writers, although the latter can be developmentally more interesting and worth examining. Regarding L2 proficiency, there were only a small number of studies providing the proficiency information of participants. In particular, as such information for child/adolescent students was scarce, we classified them into low/intermediate learners if they were in foreign language settings. Still, there were only six studies available for the category. The moderator effect of L2 proficiency was difficult to capture in this analysis, but L2 learning context appears to have served as an alternative indicator of L2 proficiency.

L1–L2 distance turned out to be a significant moderator. We hypothesized that if L1–L2 distance is shorter, that would promote cross-linguistic transfer of writing components, reducing individual differences in those component skills, thus resulting in a smaller correlation between overall L2 writing quality and its syntactic complexity features. Contrary to our expectation, the mean correlation was significantly stronger for the shorter L1–L2 distance group ( $r = .346$ ) than for the longer distance group ( $r = .173$ ). It seems even when L1 and L2 are linguistically similar, L2 learners cannot simply transfer their L1 syntactic knowledge to L2, and thus their individual differences in syntactic complexity are substantial. A cross-linguistic transfer effect seems to strengthen the association between syntactic complexity features and overall L2 writing performance for L2 learners with closer L1.

Task type was not a significant moderator ( $r = .298$  for academic writing and  $r = .301$  for narrative/personal writing). We assumed that various subskills are involved in academic writing, diminishing the impact of syntactic complexity on the overall quality of L2 texts. A closer look at the primary studies revealed that in narrative/personal writing studies, global complexity measures were mainly used, whereas in academic writing studies more phrasal complexity measures and specific measures were used. In both types of writing, developmentally suitable syntactic complexity measures seem to account for overall L2 writing performance.

Writing scoring method was a significant moderator, and syntactic complexity of texts correlated more strongly with the total scores of analytic dimensions ( $r = .570$ ) than with holistic scores of overall text quality ( $r = .151$ ). The latter result is consistent with the finding of Kojima and Kaneta (2020), who examined the effect of syntactic complexity on holistic rating of the quality of texts written by ESL/EFL learners. It seems that syntactic complexity has only trivial effects on human raters' intuitive holistic scoring of L2 text quality, which is considered closer to readers' evaluation in the real world than analytic scoring. For example, the ESL Composition Profile (Jacobs et al., 1981) awards the "language use" dimension of a text up to 25 points out of a total score of 100. In fact, the scoring rubric of Jacobs et al. (1981) indicates that "language use" includes descriptions for syntactic complexity as well as grammatical accuracy features. The weight of syntactic complexity in overall text quality might become too great in such a rating scale compared to holistic scoring. Although Weigle (2002) suggested that experienced raters may read an L2 composition holistically and adjust analytic scores to match their expectations of what the total score would be, the present study demonstrated that there was significant difference between holistic and analytic scoring in terms of the typical weight of syntactic complexity in overall text quality.

Measurement type was also a significant moderator, and global complexity measures yielded a significantly stronger mean correlation with overall writing quality ( $r = .289$ ) than that shown by specific measures and phrasal complexity measures ( $r = .248$  and  $r = .218$ , respectively), and so did specific measures in comparison with phrasal complexity measures. Biber et al. (2011) argue that phrasal complexity is the main source of syntactic complexity in written academic texts. In our analysis, phrasal complexity measures were exclusively used to analyze academic texts, whereas global and specific measures were used to examine both academic and narrative/personal texts. Thus, it might not be an appropriate comparison. However, even if we compare the result of phrasal complexity ( $r = .218$ ) with that of academic writing ( $r = .298$ ), phrasal complexity did not show a stronger effect on overall text quality. Thus, it seems that phrasal complexity has only a trivial effect on the quality of academic writing. If we focus on the academic writing of advanced L2 learners,

phrasal complexity might be more important than global complexity. However, because of the small number of primary studies involved, it was difficult to examine such an interactional effect in the present study. We need to cumulate more primary studies to conduct such a comparison in a future study.

### 6.1.2 *Lexical complexity*

Lexical complexity also demonstrated a significant but small mean correlation with overall L2 writing quality ( $r = .295$  [.234, .354]). Thus, lexical complexity explained only 8.7% of individual variability of L2 writing performance. However, among child/adolescent learners, lexical complexity had a medium effect on average ( $r = .435$ ), whereas for advanced learners it had a below small effect ( $r = .201$ ). Thus, we need to consider the effects of moderator variables. We will discuss the results of the moderator analyses below.

Age was a significant moderator, and the integrated mean correlation was stronger for child/adolescent learners ( $r = .435$ ) than for adult learners ( $r = .258$ ). This result supported our hypothesis, and it seems that L2 lexical knowledge of younger learners would still be developing and their individual variability of lexical complexity in texts explained substantial amount of their variability in overall L2 writing performance.

In contrast, L2 proficiency and language learning context were insignificant moderators. As expected, the mean correlation between lexical complexity and overall L2 text quality was stronger for low/intermediate learners ( $r = .356$ ) than for advanced learners ( $r = .201$ ), and for foreign language learners ( $r = .335$ ) than for second language learners ( $r = .250$ ), but these differences were statistically insignificant. Considering that L2 proficiencies of second language learners are usually higher than those of foreign language learners, the L2 proficiency effect on lexical complexity seems to be trivial. It seems that lexical competence is also important for advanced learners whose writing tasks are typically academic ones requiring low-frequency, sophisticated vocabulary. As this type of vocabulary knowledge can vary greatly across individuals, even after reaching advanced proficiency, L2 vocabulary development seems to continue and correlate with writing performance as much as for low/intermediate learners.

L1–L2 distance did not show a significant moderating effect either. Contrary to our expectation, the shorter distance group yielded a higher mean correlation between lexical complexity and overall text quality ( $r = .352$ ) than that shown by the longer distance group ( $r = .267$ ), although the difference was statistically insignificant. Even among the shorter L1–L2 distance group, individual differences of L2 lexical competence seem to be substantial and significantly associate with overall L2 text quality.

Task effect was not significant either. We had assumed that various text features would be involved in academic writing as the taxonomy of academic writing skills presented by Grabe and Kaplan (1996) suggests, and that the relative importance of lexical complexity would be smaller in overall performance of academic writing than in narrative/personal writing. In fact, the effect of lexical complexity was smaller in academic writing ( $r = .295$ ) than in narrative/personal writing ( $r = .351$ ), but the difference was statistically insignificant. In academic writing, low frequency and sophisticated vocabulary is required, and individual differences of lexical complexity in texts seem to be associated with overall L2 writing performance as much as in narrative/personal writing.

Scoring method also turned out to be an insignificant moderator ( $r = .352$  for analytic scoring and  $r = .281$  for holistic scoring). The ESL Composition Profile (Jacob et al., 1981), for example, awards the vocabulary dimension of a text up to 20 points out of a total score of 100. This type of weight might not be significantly different from intuitive holistic scoring of overall text quality.

Measurement type was a significant moderator, and lexical sophistication had significantly stronger mean effect ( $r = .317$ ) on overall L2 text quality than that shown by lexical diversity measures ( $r = .285$ ) and lexical density measures ( $r = .134$ ), and that of lexical diversity measures was significantly greater than that of lexical density measures. Lexical sophistication measures utilize information on word frequency in a large corpus or psycholinguistic properties of words (e.g., word familiarity), and thus seem to have stronger effect on overall L2 writing performance than do lexical diversity and density measures. In contrast, lexical density measures had only a trivial mean effect. In fact, some primary studies even reported negative effects. If L2 learners drop articles and do not use prepositions, lexical density scores would be high but would not be associated with high-quality L2 writing. Thus, lexical density may not be a valid L2 developmental measure, especially for low-level L2 learners.

### 6.1.3 *Accuracy*

Accuracy had a significant and medium correlation with overall L2 writing performance ( $r = .477$  [.373, .570]), which was significantly stronger than that of lexical complexity and cohesion as components of overall L2 writing performance. In other words, accuracy explained 22.8% of the variance caused by individual differences in overall L2 writing performance, which is substantial. This result supported the observation of some primary studies (e.g., Brodkey & Young, 1981; Perkins, 1980) that accuracy successfully discriminates texts written by different L2 writing proficiency groups.

Regarding moderator analyses, writing scoring method was a significant moderator, and the effect of accuracy on overall text quality was stronger when text quality was evaluated by the total scores of analytic scoring dimensions ( $r = .698$ ) than by holistic scoring ( $r = .443$ ). Thus, accuracy explained on average 48.7% of the variance of those analytic composite scores. The ESL Composition Profile (Jacobs et al., 1981), for example, does not include an accuracy dimension but comprises language use, vocabulary, and mechanics dimensions. It seems accuracy of these linguistic features affects the resulting composite scores substantially. In contrast, the shared variance between accuracy and human raters' intuitive holistic scores was only 19.6%. The latter result is consistent with the finding of Kojima and Kaneta (2020). Although Weigle (2002) suggested that experienced raters may read an L2 composition holistically and adjust analytic scores to match their expectation of what the total score would be, the present study demonstrated that the typical weight of accuracy in total scores of analytic scoring is significantly different from that in holistic scoring.

In contrast, age, L2 proficiency, language learning context, and writing task type were not significant moderators. We hypothesized that the L2 linguistic knowledge of child/adolescent, low/intermediate, and foreign language learners would generally not be developed fully, and thus accuracy would explain individual variability of L2 writing performance better than more mature counterparts. As we expected, accuracy was associated with overall L2 writing performance more strongly for child/adolescent learners ( $r = .555$ ) than for adult learners ( $r = .431$ ), for low/intermediate L2 learners ( $r = .457$ ) than for advanced learners ( $r = .381$ ), and for foreign language learners ( $r = .539$ ) than for second language learners ( $r = .415$ ). However, these differences were statistically insignificant. A closer examination of primary studies indicated that adult learners engaged in academic writing more frequently than younger learners, and so did advanced and second language learners than their counterparts. In fact, the effect of accuracy for overall academic writing tasks ( $r = .467$ ) was almost equivalent for overall narrative/personal writing performance ( $r = .490$ ). Considering the nature of academic writing, in which detailed information needs to be conveyed more precisely than in non-academic writing, accuracy seems to be important in advanced academic writing as well, resulting in the insignificance of these results.

L1–L2 distance was not a significant moderator either. Contrary to our assumption, the closer L1–L2 distance group yielded a higher average correlation ( $r = .540$ ) than the longer distance group ( $r = .433$ ), although the difference was statistically insignificant. As Jeon and Yamashita (2014) pointed out in their meta-analysis of L2 reading and its correlates, even within the same language family, morphosyntactic systems of different languages can vary greatly. It seems that even L2 learners with

closer L1 need to be sensitive to the L2 language system and selectively transfer their L1 knowledge to L2, causing individual variability in linguistic accuracy in their writing, as much as L2 learners with longer L1–L2 distance.

#### 6.1.4 *Fluency*

The integrated average correlation between fluency and L2 writing performance was significant and quite strong ( $r = .570$  [.463, .661]), and significantly greater than that of syntactic/lexical complexity and cohesion as components of overall L2 writing performance. In this study, fluency measures were limited to those related to the speed and quantity of text production, but it is noteworthy that a large effect was seen, explaining about one-third of the variance (i.e., 32.5%) of L2 writing performance. As Housen et al. (2012) and Wolfe-Quintero et al. (1998) pointed out, in order to produce language fluently, L2 learners not only need linguistic knowledge but also to proceduralize that knowledge and smoothly convert ideas they want to convey into L2 forms appropriately. Therefore, various knowledge resources and cognitive processes seem to be involved in writing fluency, resulting in the stronger effect on overall L2 text quality than other more isolated linguistic features of texts (i.e., lexical/syntactic complexity and cohesion).

Regarding moderator analysis, measurement type was a significant moderator and word-based fluency measures (e.g., words per minute or text) had a stronger mean correlation ( $r = .590$ ) with overall L2 text quality than unit-based fluency measures (e.g., number of sentences/clauses per minute of text) did ( $r = .325$ ). It seems that unit-based fluency measures penalize L2 writers who produce longer language units because the longer the language units they produce, the fewer the number of these units within an allocated time of writing. As our results of syntactic complexity indicate that more mature L2 writers tend to produce longer language units, unit-based fluency measures seem to be negatively influenced by such syntactic maturity.

In contrast, age, L2 proficiency, learning context, and task type were not significant moderators. We hypothesized that L2 linguistic knowledge of younger or lower-proficiency learners would still be developing and would not yet be fully proceduralized, and thus individual variability of fluency would be greater and more strongly associated with L2 writing performance than for more mature or advanced L2 learners. As we expected, the association between fluency and overall L2 writing performance was stronger for low/intermediate learners ( $r = .629$ ) than for advanced learners ( $r = .547$ ), although the difference was statistically insignificant. Contrary to our expectation, the effect of fluency on overall L2 writing performance was stronger for adult learners ( $r = .595$ ) than for younger learners ( $r = .488$ ), and for second language learners ( $r = .594$ ) than for foreign language



learners ( $r = .557$ ), although these differences were also statistically insignificant. It seems that fluency is important for various developmental stages of L2 learners with various conditions. These results seem to be consistent with the findings of task type as a moderator. The effect of fluency on overall L2 writing performance was almost equivalent ( $r = .583$  for academic writing and  $r = .585$  for narrative/personal writing). Typically, more mature or advanced L2 learners engage in academic writing, and younger or lower proficient L2 learners in narrative/personal writing. In academic writing, writers need to develop their arguments fully, including, for example, supporting data and rebuttals. Therefore, a sufficient amount of language output would be a premise for high-quality academic writing, resulting in an insignificant difference from the outcome of narrative/personal writing.

L1–L2 language distance was not a significant moderator either. We hypothesized that it would be more difficult for L2 learners with longer L1–L2 distance to produce an L2 text fluently than for L2 learners with closer L1, and thus fluency would explain greater individual variability of overall L2 writing performance for the former than for the latter. In fact, the former group yielded a higher mean effect size ( $r = .567$ ) than the latter ( $r = .497$ ), but the difference was statistically insignificant. It seems that even for L2 learners with closer L1, the systems of two languages can be substantially different, which would cause individual differences of L2 linguistic knowledge and the extent to which these knowledge and processes are proceduralized, resulting in the insignificant difference from that of the longer L1–L2 distance group.

Scoring method was not a significant moderator either. Although Weigle (2002) stated that holistic scores correlate with superficial aspects such as text length more strongly than do analytic scores, our results showed that fluency (i.e., text length) was quite strongly correlated with the total scores of analytic scoring ( $r = .511$ ), and did not significantly differ from the mean correlation with holistic scoring ( $r = .580$ ). The ESL Composition Profile (Jacobs et al., 1981), for example, does not include a fluency dimension, but it does have content and organization dimensions. To get high scores in these dimensions, writers would need to fully develop their arguments and produce a sufficient amount of language. Thus, fluency seems to be important for the composite scores of analytic scoring as much as for holistic scoring.

### 6.1.5 *Cohesion*

The integrated mean correlation between cohesion and overall L2 writing performance was significant but trivial ( $r = .198$  [.096, .296]), and significantly smaller than the effect of fluency and accuracy on L2 text quality. Cohesion is considered important because proficient writers generally produce coherent texts with adequate cohesive devices and organize discourse effectively. However, its role in

overall text quality seems to be smaller than that of writing fluency and accuracy. As Crossley et al. (2016a) argued, most L2 cohesion studies have focused on local and text cohesion, but these cohesive ties did not always positively correlate with L2 writing quality. More studies with measures of global cohesive ties are needed, and the role of cohesion in overall L2 writing performance should be further explored.

As for moderator analyses, L1–L2 distance was a significant moderator, and the mean effect of cohesion on overall text quality was greater for L2 learners with closer L1 ( $r = .548$ ) than for the longer L1–L2 distance group ( $r = .170$ ). This result is contrary to our expectations, but consistent with the results for syntactic complexity. A cross-linguistic transfer effect seems to strengthen the association between cohesion features and overall L2 writing performance for L2 learners with closer L1. In contrast, for L2 learners with longer L1–L2 distance, the rhetorical convention of organizing texts using cohesive devices could be considerably different between L1 and L2, and the use of cohesive devices seems to not always be associated with a higher quality of L2 texts.

Task type, participants' age, and L2 proficiency were not significant moderators. We assumed that cohesion would be more important in academic writing than in narrative/personal writing. In fact, cohesion had a stronger effect on academic writing performance ( $r = .213$ ) than on narrative/personal writing performance ( $r = .179$ ), and the former correlation was statistically significant whereas the latter was insignificant. However, the difference of the two correlations was statistically insignificant. Similarly, adult L2 learners demonstrated a significant and stronger mean correlation ( $r = .225$ ) than younger learners, who showed a non-significant mean correlation ( $r = -.122$ ), although the difference was not statistically significant. In contrast, advanced L2 learners yielded a stronger mean correlation ( $r = .193$ ) than low/intermediate learners ( $r = .162$ ), but both the correlations and the difference were insignificant. A closer look at primary studies indicated that adult and advanced learners tend to engage in academic writing, whereas child/adolescent and low/intermediate learners engage in narrative/personal writing, and the numbers of studies that examined cohesion in narrative/personal texts or texts written by child/adolescent or low/intermediate learners were much smaller (i.e., one to three studies). These small numbers of studies seem to have caused these insignificant results. Although primary studies have focused on cohesion in adult or advanced learners' academic texts, in order to understand the developmental patterns of cohesive devices, more studies would need to examine cohesion in non-academic texts or in texts written by younger or lower proficient L2 learners.

Regarding L2 learning context, the foreign language group showed a small but significant mean correlation ( $r = .281$ ), whereas the second language group showed an insignificant mean correlation ( $r = .053$ ). Acquiring L2 rhetorical conventions such as cohesion might be more difficult in foreign language environments than

in second language environments, and individual variability of cohesive features in texts might be greater and more strongly associated with the quality of L2 texts in the former case than in the latter. However, the difference between the two effect sizes was statistically insignificant, which seems to be attributable to the small number of studies of second language learners (i.e., five studies). Although cohesion studies have rapidly increased in the last 10 years (Kojima, 2020), these studies typically examined cohesion in academic texts written by adult or advanced foreign language learners. Cohesion studies need to investigate more diverse L2 writers.

Scoring method was not a significant moderator either. Cohesion had a trivial effect on L2 writing performance assessed by both holistic scoring ( $r = .182$ ) and total scores of analytic scoring ( $r = .136$ ). Most analytic scoring methods do not include a cohesion dimension, and the effect of cohesion on their composite scores seems to be as trivial as that on holistic scoring.

## 6.2 Subjective measures

All discrete features of L2 text quality judged by human raters subjectively had strong correlation with overall L2 writing performance ( $r = .668$  to  $.927$ ). These results seem to be consistent with the argument of Bacha (2001), who stated that those rating dimensions are often highly correlated with holistic scores, and thus their usefulness is limited. However, examination of their 95% CIs revealed some significant differences among them. The mean effect of the content feature on overall L2 text quality ( $r = .927$  [.892, .951]) was significantly stronger than that of cohesion ( $r = .688$  [.259, .891]), coherence ( $r = .668$  [.450, .811]), and mechanical features ( $r = .766$  [.524, .894]) of texts, and the effects of language use ( $r = .920$  [.862, .954]) and lexical features ( $r = .888$  [.825, .929]) were significantly stronger than that of the coherence feature of texts ( $r = .668$ ). The results of both objective and subjective measures consistently suggest that the cohesion feature is less important than other features of texts in overall L2 writing performance. However, comparing the effects of linguistic and rhetorical/idea features on overall text quality, their 95% CIs largely overlap and there seems to be an insignificant difference between them. Therefore, both linguistic and rhetorical/idea features of texts seem to equally contribute to overall text quality. However, we need to note that most of the primary studies we integrated examined correlations between each of the discrete dimensions of analytic scoring and their composite scores rather than holistic scores of overall text quality. The ESL Composition Profile (Jacobs et al., 1981) gives content dimension of writing the greatest weight (i.e., 25% of the total score), followed by the language use dimension (i.e., 20% of the total score). As their weights were larger than those for other dimensions, there is no wonder they had stronger association with the total scores. Few studies have compared those discrete features with holistic scoring

of overall text quality. Yun (2005) did such a comparison and reported correlations of  $r = .868$  between the content feature and holistic text quality and  $r = .809$  between the language use feature and holistic text quality. The effect of content and language use features can be smaller in holistic scoring than in total scores of analytic features. However, because of the small number of primary studies, we did not conduct moderator analyses of subjective measures or differentiated writing scoring methods. We call for more primary studies to investigate the weight of discrete features of texts judged by human raters compared to the overall holistic scores of texts. In this way, we can seek a more appropriate weight of each discrete feature in an analytic scoring scheme.

## 7. Conclusion

The present study systematically reviewed past studies and investigated the relationship between L2 writing and its internal correlates. By doing so, this meta-analysis aimed to identify key correlates of L2 writing proficiency and compare their relative importance. Considering the nature of language performance and development, which is complex and nonlinear, we also examined some hypothesized moderators: participants' age, L2 proficiency, language learning context, L1–L2 distance, task type, writing scoring method, and some measurement characteristics. Among objective measures, fluency was the strongest construct associated with overall L2 writing performance, followed by accuracy and syntactic/lexical complexity. Cohesion had only a trivial effect on overall text quality. Among the subjective measures, content and language use features had the strongest effects on overall L2 writing performance, followed by vocabulary, organization, argument, grammar, accuracy, and mechanics. Cohesion and coherence features had the least effects. Overall, both linguistic and rhetorical/idea features equally contributed to L2 text quality, although the role of cohesion and coherence might be less important than other features.

The second aim of this meta-analysis was to identify moderator variables that systematically affect the correlational outcomes. Our results demonstrated that syntactic complexity correlates with overall text quality for less proficient foreign language learners more strongly than for more proficient second language learners, whereas the effect of participants' age was insignificant. In contrast, lexical complexity affects overall text quality for younger learners more strongly than for adult learners, whereas L2 proficiency effect was insignificant. These results suggest that individual differences of syntactic competence are greater for lower-proficient L2 writers than for advanced learners, whereas the importance of L2 lexical competence seems to be more consistent from low to advanced stages of L2 writing

development. The effect of cognitive maturity on L2 lexical competence seems to be greater than L2 proficiency effect.

Conversely, the effects of age and L2 proficiency were not evident on fluency and accuracy, which showed a relatively strong association with overall L2 writing performance across various age and L2 proficiency groups. L1 effect was significant for syntactic complexity and cohesion features of L2 texts, and when participants' L1 and L2 were closer, those text features were associated with overall L2 writing performance more strongly than for participants with longer L1–L2 distance. A cross-linguistic transfer effect seems to strengthen those associations. Writing scoring method was also a significant moderator, and the weight of syntactic complexity and language accuracy of texts was greater in composite scores of analytic scoring dimensions than in holistic scores of overall text quality. Therefore, the total scores of analytic scoring seem to be more language-focused and significantly different from holistic scoring, which is considered more authentic and closer to the personal reaction of a reader to a text. Some measurement characteristics were also significant moderators. Thus, the present study demonstrated that study outcomes can vary greatly depending on which aspect of a construct is focused on by a measure. Therefore, researchers need to carefully consider and select appropriate measurements in accordance with their research purposes.

This study also revealed some relatively under-investigated but potentially important features of L2 writing performance. Cohesion is such a component. Although cohesion studies have rapidly increased in the last 10 years (Kojima, 2020), these studies mainly target academic texts written by adult or advanced foreign language learners. To understand developmental pattern of cohesion in L2 texts, more diverse L2 writers' texts should be examined. Second, because of the small number of studies, we did not conduct moderator analyses for subjective measures but pointed out the importance of investigating the weight of discrete features of texts analytically scored by human raters compared to overall holistic scores of texts. In this way, a more authentic weight of each discrete feature of analytic scoring will emerge. Therefore, we would encourage L2 writing researchers to devote greater effort to studies of these under-investigated writing components targeting a variety of participants.

An implication of this study for teachers and instructors is that fluency and language accuracy of L2 writing should be promoted across various developmental stages of L2 learners in various conditions, whereas lexical and syntactic competence should be more focused upon when instructing child/adolescent or low/intermediate L2 writers. As the content of writing is strongly connected with overall L2 writing performance, L2 language teachers should improve L2 learners' topic knowledge and the quality of their argument through instruction. Content-based instruction (Stoller, 2004) and an integrated communication skills approach (Koda & Yamashita, 2019) will be effective.

Despite these useful implications for researchers and practitioners, the present study is not without limitations. First, some potential moderators were not considered. Specifically, the second language group included minority language speakers with primary or secondary education as well as international students with tertiary education or those who were preparing for it without differentiating them, although the majority fell into the latter category. A future study will need to consider this point. Some measurement characteristics were not differentiated either. Accuracy measures included grammatical, lexical, and mechanical accuracy measures, but they were not differentiated because mixed measures were most frequently used. Cohesion measures included global, local, and text cohesive devices, but they were not differentiated either, because most of the primary studies employed local cohesive devices. Lexical complexity measures were classified into three groups: lexical diversity, sophistication, and density, but there was some variation among them. For example, the most widely used lexical diversity measure, type-token ratio (TTR), is known for its text length dependency and low validity, but it was not differentiated from other, more reliable diversity measures. Lexical sophistication measures included not only the ratio of advanced words (i.e., low frequent words in a larger corpus), but also such measures as psycholinguistic property indices (e.g., word familiarity, word concreteness) based on the claim of Crossley and associates (e.g., Kim et al., 2018; Kyle & Crossley, 2016). Considering the significant effects of measurement types as demonstrated by the present study, the effects of measurement characteristics should be explored further.

Second, interactional effects of potential moderators were not considered. For example, phrasal complexity measures had a smaller effect on overall text quality than global and specific measures, but if we had focused on the academic writing of advanced writers, the results might have been different. To investigate the interactional effects of moderators, more primary studies with diverse participants and study features need to be considered. Including objective information on L2 proficiency is also strongly encouraged. Third, the correlations between L2 writing and its components integrated by this study were not corrected for attenuation. Thus, the true mean correlations could have been larger than reported here.

Despite these limitations, we believe that our study offers useful insights into the relationship between L2 writing performance and its internal correlates to better understand individual differences in L2 writing proficiency. As a future direction toward a comprehensive model of L2 writing proficiency, meta-analytic structural equation modeling (SEM; Jak, 2015) of L2 writing components should be pursued. Using this technique, researchers can examine how each component jointly explain overall L2 writing performance, and a more comprehensive picture of L2 writing proficiency will emerge.

## References

- Bacha, N. (2001). Writing evaluation: What can analytic versus holistic essay scoring tell us? *System*, 29(3), 371–383. [https://doi.org/10.1016/S0346-251X\(01\)00025-2](https://doi.org/10.1016/S0346-251X(01)00025-2)
- Barkaoui, K. (2010a). Do ESL essay raters' evaluation criteria change with experience? A mixed-methods, cross-sectional study. *TESOL Quarterly*, 44(1), 31–57. <https://doi.org/10.5054/tq.2010.214047>
- Barkaoui, K. (2010b). Explaining ESL essay holistic scores: A multilevel modeling approach. *Language Testing*, 27(4), 515–535. <https://doi.org/10.1177/0265532210368717>
- Berman, R., & Verhoeven, L. (2002). Cross-linguistic perspectives on the development of text-production abilities: Speech and writing. *Written Language & Literacy*, 5(1), 1–43. <https://doi.org/10.1075/wll.5.1.02ber>
- Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly*, 45(1), 5–35. <https://doi.org/10.5054/tq.2011.244483>
- Biber, D., Gray, B., & Staples, S. (2016). Predicting patterns of grammatical complexity across language exam task types and proficiency levels. *Applied Linguistics*, 37(5), 639–668. <https://doi.org/10.1093/applin/amu059>
- Bouwer, R., Béguin, A., Sanders, T., & Van den Bergh, H. (2015). Effect of genre on the generalizability of writing scores. *Language Testing*, 32(1), 83–100. <https://doi.org/10.1177/0265532214542994>
- Brodkey, D., & Young, R. (1981). Composition correctness scores. *TESOL Quarterly*, 15(2), 159–167. <https://doi.org/10.2307/3586407>
- Bulté, B., & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing*, 26, 42–65. <https://doi.org/10.1016/j.jslw.2014.09.005>
- Cooper, H. (2010). *Research synthesis and meta-analysis: A step-by-step approach* (Vol. 2). Sage.
- Crossley, S. A., Kyle, K., Allen, L. K., Guo, L., & McNamara, D. S. (2014). Linguistic microfeatures to predict L2 writing proficiency: A case study in automated writing evaluation. *The Journal of Writing Assessment*, 7(1), 1–34.
- Crossley, S. A., Kyle, K., & McNamara, D. S. (2016a). The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. *Journal of Second Language Writing*, 32, 1–16. <https://doi.org/10.1016/j.jslw.2016.01.003>
- Crossley, S. A., Kyle, K., & McNamara, D. S. (2016b). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*, 48(4), 1227–1237. <https://doi.org/10.3758/s13428-015-0651-7>
- Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, 35(2), 115–135. <https://doi.org/10.1111/j.1467-9817.2010.01449.x>
- Crossley, S. A., Salsbury, T., & McNamara, D. S. (2012). Predicting the proficiency level of language learners using lexical indices. *Language Testing*, 29(2), 243–263. <https://doi.org/10.1177/0265532211419331>
- Crossley, S. A., Salsbury, T., & McNamara, D. S. (2015). Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. *Applied Linguistics*, 36(5), 570–590. <https://doi.org/10.1093/applin/amt056>
- De Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition*, 34(1), 5–34. <https://doi.org/10.1017/S0272263111000489>



- Engber, C. A. (1992). A study of lexis and the relationship to quality in written texts of second language learners of English (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (303986307)
- Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing*, 4(2), 139–155. [https://doi.org/10.1016/1060-3743\(95\)90004-7](https://doi.org/10.1016/1060-3743(95)90004-7)
- Evola, J., Mamer, E., & Lentz, B. (1980). Discrete-point versus global scoring for cohesive devices. In J. W. Oiler & K. Perkins (Eds.), *Research and language testing* (pp. 177–181). Newbury House.
- Flahive, D. E., & Snow, B. G. (1980). Measures of syntactic complexity in evaluating ESL compositions. In J. W. Oiler & K. Perkins (Eds.), *Research and language testing* (pp. 171–176). Newbury House.
- Grabe, W., & Kaplan, R. B. (1996). *Theory and practice of writing: An applied linguistic perspective*. Longman.
- Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing*, 18(3), 218–238. <https://doi.org/10.1016/j.asw.2013.05.002>
- Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. Longman.
- Hinkel, E. (2002). *Second language writers' text: Linguistic and rhetorical features*. Routledge. <https://doi.org/10.4324/9781410602848>
- Housen, A., Kuiken, F., & Vedder, I. (2012). Complexity, accuracy and fluency: Definitions, measurement and research. In A. Housen, F. Kuiken, & F. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 1–19). John Benjamins. <https://doi.org/10.1075/llt.32.01hou>
- Hunt, K. W. (1965). *Grammatical structures written at three grade levels* (Vol. 3). National Council of Teachers of English.
- Jacobs, H. L., Zinkgraf, S. A., Wormuth, D. R., Hartfiel, V. F., & Hughey, J. B. (1981). *Testing ESL composition: A practical approach*. Newbury House.
- Jak, S. (2015). *Meta-analytic structural equation modelling*. Springer. <https://doi.org/10.1007/978-3-319-27174-3>
- Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing*, 19(1), 57–84. <https://doi.org/10.1191/0265532202lt2200a>
- Jarvis, S. (2013). Capturing the diversity in lexical diversity. *Language Learning*, 63(S1), 87–106. <https://doi.org/10.1111/j.1467-9922.2012.00739.x>
- Jeon, E.-H., & Yamashita, Y. (2014). L2 reading comprehension and its correlates: A meta-analysis. *Language Learning*, 64(1), 160–212. <https://doi.org/10.1111/lang.12034>
- Kaczmarek, C. M. (1980). Scoring and rating essay tasks. In J. W. Oiler & K. Perkins (Eds.), *Research and language testing* (pp. 151–159). Newbury House.
- Kim, M., & Crossley, S. A. (2018). Modeling second language writing quality: A structural equation investigation of lexical, syntactic, and cohesive features in source-based and independent writing. *Assessing Writing*, 37, 39–56. <https://doi.org/10.1016/j.asw.2018.03.002>
- Kim, M., Crossley, S. A. & Kyle, K. (2018). Lexical sophistication as a multidimensional phenomenon: Relations to second language lexical proficiency, development, and writing quality. *The Modern Language Journal*, 102(1), 120–141. <https://doi.org/10.1111/modl.12447>
- Kim, Y. A., & Na, Y. H. (2009). Cohesive devices and quality of argumentative writing produced by Korean EFL learners. *Studies in English Education*, 14(2), 1–29.
- Koda, K. (2005). *Insights into second language reading: A cross-linguistic approach*. Cambridge University Press. <https://doi.org/10.1017/CBO9781139524841>



- Koda, K., & Yamashita, J. (2019). *Reading to learn in a foreign language: An integrated approach to foreign language instruction and assessment*. Routledge.
- Kojima, M. (2020). A study synthesis on the relationship between second language writing performance and text features: Focusing on text-based measures and study features. *Learner Corpus Studies in Asia and the World*, 5, 1–24. <https://doi.org/10.24546/81012486>
- Kojima, M., & Kaneta, T. (2020). Raitingu hyoka to gengoteki shihyo no kankei: Meta bunseki ni yoru kenkyu seika no togo [The relationship between writing performance and linguistic indices: A meta-analysis]. In Y. Ishii & Y. Kondo (Eds.), *Eigo Kyoiku ni okeru jido saiten: Genjo to kadai* [Automated scoring in English language education: Its current situation and issues] (pp. 33–72). Hituzi Shobo.
- Kroll, B. (1982). Levels of error in ESL composition (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (303230742)
- Kroll, B. (1990). What does time buy? ESL student performance on home versus class compositions. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 140–154). Cambridge University Press. <https://doi.org/10.1017/CBO9781139524551.014>
- Kyle, K., & Crossley, S. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing*, 34, 12–24. <https://doi.org/10.1016/j.jslw.2016.10.003>
- Kyle, K., & Crossley, S. (2017). Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing*, 34(4), 513–535. <https://doi.org/10.1177/0265532217712554>
- Kyle, K., & Crossley, S. A. (2018). Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. *The Modern Language Journal*, 102(2), 333–349. <https://doi.org/10.1111/modl.12468>
- Larsen-Freeman, D. (2009). Adjusting expectations: The study of complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, 30(4), 579–589. <https://doi.org/10.1093/applin/amp043>
- Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, 16(3), 307–322. <https://doi.org/10.1093/applin/16.3.307>
- Lee, Y.-W., Gentile, C., & Kantor, R. (2008). Analytical scoring of TOEFL CBT essays: Scores by humans and e-rater, *TOEFL Research Report*, 81, ETS RR-08-01. Educational Testing Service. <https://doi.org/10.1002/j.2333-8504.2008.tb02087.x>
- Lee, Y.-W., Gentile, C., & Kantor, R. (2010). Toward automated multi-trait scoring of essays: Investigating links among holistic, analytic, and text feature scores. *Applied Linguistics*, 31(3), 391–417. <https://doi.org/10.1093/applin/amp040>
- Leki, I., Cumming, A., & Silva, T. (2008). *A synthesis of research on second language writing in English*. Routledge.
- Linnarud, M. (1986). *Lexis in composition: A performance analysis of Swedish learners' written English*. CWK Gleerup.
- Liu, M., & Braine, G. (2005). Cohesive features in argumentative writing produced by Chinese undergraduates. *System*, 33(4), 623–636. <https://doi.org/10.1016/j.system.2005.02.002>
- Lu, X., & Ai, H. (2015). Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds. *Journal of Second Language Writing*, 29, 16–27. <https://doi.org/10.1016/j.jslw.2015.06.003>
- Luk, Z. P. S., & Shirai, Y. (2009). Is the acquisition order of grammatical morphemes impervious to L1 knowledge? Evidence from the acquisition of plural -s, articles, and possessive's. *Language Learning*, 59(4), 721–754. <https://doi.org/10.1111/j.1467-9922.2009.00524.x>

- Malvern, D., & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing*, 19(1), 85–104. <https://doi.org/10.1191/0265532202lt2210a>
- Martínez, A. C. L. (2018). Analysis of syntactic complexity in secondary education EFL writers at different proficiency levels. *Assessing Writing*, 35, 1–11. <https://doi.org/10.1016/j.asw.2017.11.002>
- McCutchen, D., & Perfetti, C. A. (1982). Coherence and connectedness in the development of discourse production. *Text-Interdisciplinary Journal for the Study of Discourse*, 2(1–3), 113–140. <https://doi.org/10.1515/text.1.1982.2.1-3.113>
- Meara, P., & Bell, H. (2001). P\_Lex: A simple and effective way of describing the lexical characteristics of short L2 texts. *Prospect*, 16(3), 5–19.
- Murakami, A., & Alexopoulou, T. (2016). L1 influence on the acquisition order of English grammatical morphemes: A learner corpus study. *Studies in Second Language Acquisition*, 38(3), 365–401. <https://doi.org/10.1017/S0272263115000352>
- Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, 50(3), 417–528. <https://doi.org/10.1111/0023-8333.00136>
- Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, 30(4), 555–578. <https://doi.org/10.1093/applin/amp044>
- Oh, E., Lee, C. M., & Moon, Y. I. (2015). The contributions of planning, L2 linguistic knowledge and individual differences to L2 writing. *The Journal of Asia TEFL*, 12(2), 45–85.
- Oh, S. (2006). Investigating the relationship between fluency measures and second language writing placement test decisions (Unpublished master's thesis). University of Hawai'i. Retrieved from <http://hdl.handle.net/10125/20203>
- Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, 24(4), 492–518. <https://doi.org/10.1093/applin/24.4.492>
- Ortega, L. (2015). Syntactic complexity in L2 writing: Progress and expansion. *Journal of Second Language Writing*, 29, 82–94. <https://doi.org/10.1016/j.jslw.2015.06.008>
- Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics*, 30(4), 590–601. <https://doi.org/10.1093/applin/amp045>
- Park, S. K. (2013). Lexical analysis of Korean university students' narrative and argumentative essays. *English Teaching (영어교육)*, 68(3), 131–157.
- Perkins, K. (1980). Using objective methods of attained writing proficiency to discriminate among holistic evaluations. *TESOL Quarterly*, 14(1), 61–69. <https://doi.org/10.2307/3586809>
- Perkins, K. (1983). On the use of composition scoring techniques, objective measures, and objective tests to evaluate ESL writing ability. *TESOL Quarterly*, 17(4), 651–671. <https://doi.org/10.2307/3586618>
- Plonsky, L., & Oswald, F. L. (2014). How big is “big”? Interpreting effect sizes in L2 research. *Language Learning*, 64(4), 878–912. <https://doi.org/10.1111/lang.12079>
- Read, J. (2000). *Assessing vocabulary*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511732942>
- Roca de Larios, J., Manchón, R., Murphy, L., & Marín, J. (2008). The foreign language writer's strategic behaviour in the allocation of time to writing processes. *Journal of Second Language Writing*, 17(1), 30–47. <https://doi.org/10.1016/j.jslw.2007.08.005>

- Rosenthal, R. (1979). The “file drawer problem” and tolerance for null results. *Psychological Bulletin*, 86, 638–641. <https://doi.org/10.1037/0033-2909.86.3.638>
- Sasaki, M. (2009). Changes in English as a foreign language students’ writing over 3.5 years: A sociocognitive account. In R. M. Manchón (Ed.), *Writing in foreign language contexts: Learning, teaching, and research* (pp. 49–76). Multilingual Matters. <https://doi.org/10.21832/9781847691859-006>
- Schoonen, R., Van Gelderen, A., De Glopper, K., Hulstijn, J., Simis, A., Snellings, P., & Stevenson, M. (2003). First language and second language writing: The role of linguistic knowledge, speed of processing, and metacognitive knowledge. *Language Learning*, 53(1), 165–202. <https://doi.org/10.1111/1467-9922.00213>
- Schoonen, R., Van Gelderen, A., Stoel, R. D., Hulstijn, J., & De Glopper, K. (2011). Modeling the development of L1 and EFL writing proficiency of secondary school students. *Language Learning*, 61(1), 31–79. <https://doi.org/10.1111/j.1467-9922.2010.00590.x>
- Skehan, P. (1998). *A cognitive approach to language learning*. Oxford University Press.
- Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, 30(4), 510–532. <https://doi.org/10.1093/applin/ampo47>
- Stoller, F. L. (2004). Content-based instruction: Perspectives on curriculum planning. *Annual Review of Applied Linguistics*, 24, 261–283. <https://doi.org/10.1017/S0267190504000108>
- Thewissen, J. (2013). Capturing L2 accuracy developmental patterns: Insights from an error-tagged EFL learner corpus. *The Modern Language Journal*, 97(S1), 77–101. <https://doi.org/10.1111/j.1540-4781.2012.01422.x>
- Verspoor, M., Schmid, M. S., & Xu, X. (2012). A dynamic usage based perspective on L2 writing. *Journal of Second Language Writing*, 21(3), 239–263. <https://doi.org/10.1016/j.jslw.2012.03.007>
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3), 1–48. <https://doi.org/10.18637/jss.v036.i03>
- Weigle, S. C. (2002). *Assessing Writing*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511732997>
- White, E. M. (1984). Holisticism. *College Composition and Communication*, 35(4), 400–409. <https://doi.org/10.2307/357792>
- Wolfe-Quintero, Y., Inagaki, S., & Kim, H.-Y. (1998). *Second language development in writing: Measures of fluency, accuracy, & complexity*. University of Hawai’i Press.
- Yang, W., Lu, X., & Weigle, S. C. (2015). Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. *Journal of Second Language Writing*, 28, 53–67. <https://doi.org/10.1016/j.jslw.2015.02.002>
- Yde, P., & Spoelders, M. (1985). Text cohesion: An exploratory study with beginning writers. *Applied Psycholinguistics*, 6(4), 407–415. <https://doi.org/10.1017/S0142716400006330>
- Yu, G. (2010). Lexical diversity in writing and speaking task performances. *Applied Linguistics*, 31(2), 236–259. <https://doi.org/10.1093/applin/ampo24>
- Yun, Y. (2005). Factors explaining EFL learners’ performance in a timed essay writing test: A structural equation modeling approach (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (3199191)

## Appendix A. 103 studies included in the meta-analysis

- Abu-Rabia, S. (2003). The influence of working memory on reading and creative writing processes in a second language. *Educational Psychology*, 23(2), 209–222.  
<https://doi.org/10.1080/01443410303227>
- Aziz, L. J. (1995). A model of paired cognitive and metacognitive strategies: Its effect on second language grammar and writing performance (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (9531651)
- Barkaoui, K. (2010). Do ESL essay raters' evaluation criteria change with experience? A mixed-methods, cross-sectional study. *TESOL Quarterly*, 44(1), 31–57.  
<https://doi.org/10.5054/tq.2010.214047>
- Barkaoui, K., & Knouzi, I. (2018). The effects of writing mode and computer ability on L2 test-takers' essay characteristics and scores. *Assessing Writing*, 36, 19–31.  
<https://doi.org/10.1016/j.asw.2018.02.005>
- Bestgen, Y. (2017). Beyond single-word measures: L2 writing assessment, lexical richness and formulaic competence. *System*, 69, 65–78. <https://doi.org/10.1016/j.system.2017.08.004>
- Black, E. M. (1997). A text analysis of the argumentative writing in English by Spanish and English bilingual college students and by English monolingual college students in the South-west United States (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (304383507)
- Breland, H. M., & Jones, R. J. (1984). Perceptions of writing skills. *Written Communication*, 1(1), 101–119. <https://doi.org/10.1177/0741088384001001005>
- Browne, S. R. (1990). Social cognition as a predictor of the writing quality of students using English as a second language in freshman English composition (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (303879213)
- Bulté, B., & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing*, 26, 42–65.  
<https://doi.org/10.1016/j.jslw.2014.09.005>
- Campbell, B. E. (1998). Coherence in the expository essays of intensive ESL students: A textual analysis of topical development (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (9824614)
- Campbell, H., Espin, C. A., & McMaster, K. (2013). The technical adequacy of curriculum-based writing measures with English learners. *Reading and Writing*, 26(3), 431–452.  
<https://doi.org/10.1007/s11145-012-9375-6>
- Carlson, S. B., Bridgeman, B., Camp, R., & Waanders, J. (1985). Relationship of admission test scores to writing performance of native and nonnative speakers of English. *ETS Research Report Series 1985*, 1, i–137. <https://doi.org/10.1002/j.2330-8516.1985.tb00106.x>
- Chao, Y.-C. J. (2003). Contrastive rhetoric, lexico-grammatical knowledge, writing expertise, and metacognitive knowledge: An integrated account of the development of English writing by Taiwanese students (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (3119448)
- Chiang, S. Y. (1999). Assessing grammatical and textual features in L2 writing samples: The case of French as a foreign language. *The Modern Language Journal*, 83(2), 219–232.  
<https://doi.org/10.1111/0026-7902.00017>
- Chon, Y. V., & Shin, D. (2009). Collocations in L2 writing and rater's perceived writing proficiency. *응용언어학*, 25(1), 101–129.
- Cleary, C. (1988). Testing lower intermediate writing: A comparison of two scoring methods. *British Journal of Language Teaching*, 26(2), 75–80.

- Crossley, S. A., Kyle, K., Allen, L. K., Guo, L., & McNamara, D. S. (2014). Linguistic microfeatures to predict L2 writing proficiency: A case study in automated writing evaluation. *The Journal of Writing Assessment*, 7(1), 1–34.
- Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. *Journal of Second Language Writing*, 32, 1–16. <https://doi.org/10.1016/j.jslw.2016.01.003>
- Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, 35(2), 115–135. <https://doi.org/10.1111/j.1467-9817.2010.01449.x>
- Daif-Allah, A. S., & Albeshier, K. (2013). The use of discourse markers in paragraph writings: The case of preparatory year program students in Qassim University. *English Language Teaching*, 6(9), 217–227. <https://doi.org/10.5539/elt.v6n9p217>
- Daller, H., & Phelan, D. (2007). What is in a teacher's mind? Teacher ratings of EFL essays and different aspects of lexical richness. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 93–115). Cambridge University Press. <https://doi.org/10.1017/CBO9780511667268.016>
- Demetriotu, T. (2016). Predicting IELTS ratings using vocabulary measures. University of the West of England (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (10135088)
- Douglas, S. R. (2010). Non-native English speaking students at university: Lexical richness and academic success (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (848966609)
- East, M. (2006). The impact of bilingual dictionaries on lexical sophistication and lexical accuracy in tests of L2 writing proficiency: A quantitative analysis. *Assessing Writing*, 11(3), 179–197. <https://doi.org/10.1016/j.asw.2006.11.001>
- El-Bacha, N. N. S. (1997). Patterns of lexical cohesion in EFL texts: A study of the compositions of students at the Lebanese American University (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (U095226)
- Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing*, 4(2), 139–155. [https://doi.org/10.1016/1060-3743\(95\)90004-7](https://doi.org/10.1016/1060-3743(95)90004-7)
- Evola, J., Mamer, E., & Lentz, B. (1980). Discrete-point versus global scoring for cohesive devices. In J. W. Oiler & K. Perkins (Eds.), *Research and language testing* (pp. 177–181). Newbury House.
- Feurer, H. (1996). Bilingual education among minority nationalities in China: A study of the Naxi Pilot Project at Yilong, Yunnan. *REL C Journal*, 27(1), 1–22. <https://doi.org/10.1177/00368829602700101>
- Fischer, R. A. (1984). Testing written communicative competence in French. *The Modern Language Journal*, 68(1), 13–20. <https://doi.org/10.1111/j.1540-4781.1984.tb01539.x>
- Flahive, D., & Bailey, N. (1993). Exploring reading/writing relationships in adult second language learners. In J. Carson & I. Leki (Eds.), *Reading in the composition class: Second language perspectives* (pp. 128–140). Heinle and Heinle.
- Flahive, D., & Snow, B. G. (1980). Measures of syntactic complexity in evaluating ESL compositions. In J. W. Oiler & K. Perkins (Eds.), *Research and language testing* (pp. 171–176). Newbury House.
- Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing*, 18(3), 218–238. <https://doi.org/10.1016/j.asw.2013.05.002>

- Gyllstad, H., Granfeldt, J., Bernardini, P., & Källkvist, M. (2014). Linguistic correlates to communicative proficiency levels of the CEFR: The case of syntactic complexity in written L2 English, L3 French and L4 Italian. *EUROSLA Yearbook*, 14(1), 1–30.  
<https://doi.org/10.1075/eurosla.14.01gyl>
- Hamp-Lyons, L., & Henning, G. (1991). Communicative writing profiles: An investigation of the transferability of a multiple-trait scoring instrument across ESL writing assessment contexts. *Language Learning*, 41(3), 337–373. <https://doi.org/10.1111/j.1467-1770.1991.tb00610.x>
- Harrison, G. L., Goegan, L. D., Jalbert, R., McManus, K., Sinclair, K., & Spurling, J. (2016). Predictors of spelling and writing skills in first- and second-language learners. *Reading and Writing*, 29(1), 69–89. <https://doi.org/10.1007/s11145-015-9580-1>
- Hassan, B. A. (2001). *The relationship of writing apprehension and self-esteem to the writing quality and quantity of EFL university students*. Retrieved from ERIC database. (ED459671)
- Hirose, K. (2003). Comparing L1 and L2 organizational patterns in the argumentative writing of Japanese EFL students. *Journal of Second Language Writing*, 12(2), 181–209.  
[https://doi.org/10.1016/S1060-3743\(03\)00015-8](https://doi.org/10.1016/S1060-3743(03)00015-8)
- Hongwei, W. (2012). Coh-Metrix: A computational tool to discriminate writing qualities. *International Education Studies*, 5(2), 204–215. <https://doi.org/10.5539/ies.v5n2p204>
- Imao, Y. (2010). Investigating the construct of lexico-grammatical knowledge in an academic ESL writing test (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (3450984)
- Jafarpur, A. (1991). Cohesiveness as a basis for evaluating compositions. *System*, 19(4), 459–465.  
[https://doi.org/10.1016/0346-251X\(91\)90026-L](https://doi.org/10.1016/0346-251X(91)90026-L)
- James, M. O. (1992). *L2 writing fluency: A pilot study*. Retrieved from ERIC database. (ED350870)
- Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing*, 19(1), 57–84. <https://doi.org/10.1191/0265532202lt2200a>
- Jung, Y. J., Crossley, S. A., & McNamara, D. S. (2015). Linguistic features in MELAB writing task performances. *CaMLA Working Papers*, 5, 1–17. Retrieved from ResearchGate. <https://www.researchgate.net/>
- Kaczmarek, C. M. (1980). Scoring and rating essay tasks. In J. W. Oiler & K. Perkins (Eds.), *Research and language testing* (pp. 151–159). Newbury House.
- Kang, H. S. (2011). The relationship between different dimensions of lexical proficiency and writing quality of Korean EFL learners. *응용언어학*, 27(3), 81–104.
- Kim, M., & Crossley, S. A. (2018). Modeling second language writing quality: A structural equation investigation of lexical, syntactic, and cohesive features in source-based and independent writing. *Assessing Writing*, 37, 39–56. <https://doi.org/10.1016/j.asw.2018.03.002>
- Kim, Y. A., & Na, Y. H. (2009). Cohesive devices and quality of argumentative writing produced by Korean EFL learners. *Studies in English Education*, 14(2), 1–29.
- Kobrin, J. L., Deng, H., & Shaw, E. J. (2007). Does quantity equal quality? The relationship between length of response and scores on the SAT essay. *Journal of Applied Testing Technology*, 8(1), 1–15.
- Kroll, B. (1990). What does time buy? ESL student performance on home versus class compositions. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 140–154). Cambridge University Press. <https://doi.org/10.1017/CBO9781139524551.014>
- Krzemińska-Adamek, M. (2016). Lexis in writing: Investigating the relationship between lexical richness and the quality of advanced learners' texts. In M. Pawlak (Ed.), *Classroom-oriented research: Reconciling theory and practice* (pp. 185–196). Springer.  
[https://doi.org/10.1007/978-3-319-30373-4\\_12](https://doi.org/10.1007/978-3-319-30373-4_12)



- Kyle, K., & Crossley, S. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing*, 34, 12–24.  
<https://doi.org/10.1016/j.jslw.2016.10.003>
- Kyle, K., & Crossley, S. (2017). Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing*, 34(4), 513–535. <https://doi.org/10.1177/0265532217712554>
- Kyle, K., & Crossley, S. A. (2018). Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. *The Modern Language Journal*, 102(2), 333–349.  
<https://doi.org/10.1111/modl.12468>
- Larsen-Freeman, D., & Strom, V. (1977). The construction of a second language acquisition index of development. *Language Learning*, 27(1), 123–134.  
<https://doi.org/10.1111/j.1467-1770.1977.tb00296.x>
- Lee, H. K. (2012). Exploring the relationship among L1 writing, L2 writing, and L2 linguistic proficiency depending on L2 topic difficulty. *Asia-Pacific Education Researcher (De La Salle University Manila)*, 21(3), 576–586.
- Lee, H. S. (2013). The correlation between matriculation students' lexical richness and their writing scores (Master's Thesis). Retrieved from <http://studentsrepo.um.edu.my/5431/>
- Lee, Y. (2010). Concept mapping strategy to facilitate foreign language writing: A Korean application (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (3429061)
- Lee, Y.-W., Gentile, C., & Kantor, R. (2010). Toward automated multi-trait scoring of essays: Investigating links among holistic, analytic, and text feature scores. *Applied Linguistics*, 31(3), 391–417. <https://doi.org/10.1093/applin/amp040>
- Li, Y. (1998). Using task-based e-mail activities in developing academic writing skills in English as a second language (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (304414086)
- Li, Y. (2000). Assessing second language writing: The relationship between computerized analysis and rater evaluation. *ITL-International Journal of Applied Linguistics*, 127(1), 37–51.  
<https://doi.org/10.1075/itl.127-128.02li>
- Linnarud, M. (1986). *Lexis in composition: A performance analysis of Swedish learners' written English*. CWK Gleerup.
- Liu, M., & Braine, G. (2005). Cohesive features in argumentative writing produced by Chinese undergraduates. *System*, 33(4), 623–636. <https://doi.org/10.1016/j.system.2005.02.002>
- Llach, M. P. A. (2005). The relationship of lexical error and their types to the quality of ESL compositions: An empirical study. *Porta Linguarum: Revista Internacional de Didáctica de las Lenguas Extranjeras*, 3, 45–57. <https://doi.org/10.30827/Digibug.29120>
- Llach, M. P. A. (2007). Lexical errors as writing quality predictors. *Studia Linguistica*, 61(1), 1–19.  
<https://doi.org/10.1111/j.1467-9582.2007.00127.x>
- Lutviana, R., Kadarisman, A. E., & Laksmi, E. D. (2015). Correlation between lexical richness and overall quality of argumentative essays written by English department students. *Jurnal Pendidikan Humaniora*, 3(1), 41–51.
- Martínez, A. C. L. (2004). Discourse markers in the expository writing of Spanish university students. *Ibérica: Revista de la Asociación Europea de Lenguas para Fines Específicos (AELFE)*, 8, 63–80.
- Martínez, A. C. L. (2018). Analysis of syntactic complexity in secondary education EFL writers at different proficiency levels. *Assessing Writing*, 35, 1–11.  
<https://doi.org/10.1016/j.asw.2017.11.002>

- Matthews, J., & Wijeyewardene, I. (2018). Exploring relationships between automated and human evaluations of L2 texts. *Language Learning & Technology*, 22(3), 143–158. <https://doi.org/10125/44661>
- McNeill, B. R. (2005). A comparative statistical assessment of different types of EFL writing by Japanese college students (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (301659887)
- Mellor, A. (2010). Automatic essay scoring for low level learners of English as a second language (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (10797955)
- Modhish, A. S. (2012). Use of discourse markers in the composition writings of Arab EFL learners. *English Language Teaching*, 5(5), 56–61. <https://doi.org/10.5539/elt.v5n5p56>
- Mullen, K. (1980). Evaluating writing proficiency in ESL. In J. W. Oiler, Jr. & K. Perkins (Eds.), *Research in language testing* (pp. 160–170). Newbury House.
- Muñoz-Luna, R., & Taillefer, L. (2014). A mathematical model for academic genre awareness. *Revista Española de Lingüística Aplicada*, 27(2), 469–491. <https://doi.org/10.1075/resla.27.2.11mun>
- O'Loughlin, K. J. (1994). The assessment of writing by English and ESL teachers. *Australian Review of Applied Linguistics*, 17(1), 23–44. <https://doi.org/10.1075/aral.17.1.020lo>
- Park, J.-H. (2017). Syntactic complexity as a predictor of second language writing proficiency and writing quality (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (10248579)
- Park, S. K. (2013). Lexical analysis of Korean university students' narrative and argumentative essays. *English Teaching (영어교육)*, 68(3), 131–157.
- Perkins, K. (1980). Using objective methods of attained writing proficiency to discriminate among holistic evaluations. *TESOL Quarterly*, 14(1), 61–69. <https://doi.org/10.2307/3586809>
- Pongsiriwet, C. (2001). Relationships among grammatical accuracy, discourse features, and the quality of second language writing: The case of Thai EFL learners (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (304728088)
- Reid, J. (1986). Using the writer's workbench in composition teaching and testing. In C. Stansfield (Ed.), *Technology and language testing* (pp. 167–188). TESOL.
- Reynolds, D. W. (2001). Language in the balance: Lexical repetition as a function of topic, cultural background, and writing development. *Language Learning*, 51(3), 437–476. <https://doi.org/10.1111/0023-8333.00161>
- Sadeghi, K., & Dilmaghani, S. K. (2013). The relationship between lexical diversity and genre in Iranian EFL learners' writings. *Journal of Language Teaching and Research*, 4(2), 328–334. <https://doi.org/10.4304/jltr.4.2.328-334>
- Schuler, P. (1992). Assessment of English proficiency: A computer-assisted approach (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (304017162)
- Seibert, L. C., & Goddard, E. R. (1935). A more objective method of scoring compositions. *The Modern Language Journal*, 20(3), 143–150. <https://doi.org/10.1111/j.1540-4781.1935.tb02669.x>
- Seifeddin, A. H., & Ebedy, H. G. M. (2016). The effects of the frequency of lexical errors on the quality of EFL learners' writing through email communication. *Journal of Research in Curriculum, Instruction and Educational Technology*, 2(3), 67–91.
- Shea, M. C. (2011). Cohesion in second language writing (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (3466231)
- Shi, L., & Qian, D. (2012). How does vocabulary knowledge affect Chinese EFL learners' writing quality in web-based settings? Evaluating the relationships among three dimensions of vocabulary knowledge and writing quality. *Chinese Journal of Applied Linguistics*, 35(1), 117–127. <https://doi.org/10.1515/cjal-2012-0009>



- Shin, Y. (2008). The effects of planning on L2 writing: A study of Korean learners of English as a foreign language (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (3315989) <https://doi.org/10.17077/etd.4deo2eme>
- Song, M. (2007). A correlational study of the holistic measure with the index measure of accuracy and complexity in international English-as-a-second-language (ESL) student writings (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (3246048)
- Tracy, G. E. (1989). The effects of sentence-combining practice on syntactic maturity and writing quality in ESL students in freshman composition (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (303791751)
- Uzun, K. (2017). The use of conjunctions and its relationship with argumentative writing performance in an EFL setting. *Journal of Teaching English for Specific and Academic Purposes*, 5(2), 307–315. <https://doi.org/10.22190/jtesap1702307u>
- Verspoor, M., Schmid, M. S., & Xu, X. (2012). A dynamic usage based perspective on L2 writing. *Journal of Second Language Writing*, 21(3), 239–263. <https://doi.org/10.1016/j.jslw.2012.03.007>
- Wistner, B. (2014). Effects of metalinguistic knowledge and language aptitude on second language learning (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (3611192)
- Yan, J., & Xu, X. Y. (2017). The Relationship between syntactic complexity and writing quality of Chinese EFL learners. In H. Z. Lin (Ed.), *Proceedings of the 3rd Annual International Conference on Social Science and Contemporary Humanity Development* (pp. 331–337). <https://doi.org/10.2991/sschd-17.2017.65>
- Yang, W. (2014). Mapping the relationships among the cognitive complexity of independent writing tasks, L2 writing quality, and complexity, accuracy and fluency of L2 writing (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (1617543253)
- Yang, W., Lu, X., & Weigle, S. C. (2015). Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. *Journal of Second Language Writing*, 28, 53–67. <https://doi.org/10.1016/j.jslw.2015.02.002>
- Yang, W., & Sun, Y. (2012). The use of cohesive devices in argumentative writing by Chinese EFL learners at different proficiency levels. *Linguistics and Education*, 23(1), 31–48. <https://doi.org/10.1016/j.linged.2011.09.004>
- Yousofi, N., & Bahrmlou, K. (2014). Assessing writing quality: Vocabulary profiles in place of holistic measures. *The Iranian EFL Journal Special Edition of 2014*, 10(6), 323–344.
- Yu, G. (2010). Lexical diversity in writing and speaking task performances. *Applied Linguistics*, 31(2), 236–259. <https://doi.org/10.1093/applin/amp024>
- Yun, Y. (2005). Factors explaining EFL learners' performance in a timed essay writing test: A structural equation modeling approach (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (3199191)
- Zhai, L. (2016). A study on Chinese EFL learners' vocabulary usage in writing. *Journal of Language Teaching and Research*, 7(4), 752–759. <https://doi.org/10.17507/jltr.0704.16>
- Zhang, A. (2010). Use of cohesive ties in relation to the quality of compositions by Chinese college students. *Journal of Cambridge Studies*, 5(2–3), 78–86. <https://doi.org/10.17863/CAM.1358>
- Zhang, M. (2000). Cohesive features in the expository writing of undergraduates in two Chinese universities. *RELJ Journal*, 31(1), 61–95. <https://doi.org/10.1177/003368820003100104>
- Zhang, X. (1993). English collocations and their effect on the writing of native and non-native college freshmen (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (9319454)

## Appendix B. Acceptable measures of L2 writing performance and internal correlates

### L2 Writing Performance

Holistic scores of overall writing quality or composite scores of analytic rating dimensions were used as an index of L2 writing performance. Acceptable measures of composite scores had to include both linguistic and rhetorical features of each text as components. Discrete scores such as holistic language accuracy were not accepted as a measure of overall L2 writing performance.

### Objective internal correlates

#### *Syntactic complexity*

An acceptable syntactic complexity measure was one calculated from the mean length of production unit (e.g., T-units, clauses, or phrases), or based on the amount of subordination (e.g., mean number of clauses per T-unit or per c-unit), frequency of specific grammatical features (e.g., frequency of passive or progressive verb forms), or ratio of those features per word, clause, or sentence, or phrasal complexity measures (e.g., prepositions per nominal).

#### *Accuracy*

An acceptable accuracy measure was derived from frequency or percentage of error counts (e.g., frequency of error-free clauses, percentage of error-free T-units, errors per 100 words). Errors could include grammatical, lexical, or mechanical errors, or specific type of errors (e.g., frequency of target-like verbal morphemes or ratio of target-like to all verbal morphemes), as well as inappropriate or non-standard usage of language.

#### *Lexical complexity*

To be used as a lexis measure, the measure had to assess either lexical diversity, lexical sophistication, including psycholinguistic property measures, or lexical density (i.e., content word–total word ratio). Lexical diversity measures are usually based on the ratio of different words (types) to total number of words (tokens), known as the type–token ratio (TTR), or improved measures based on it but altered to overcome its text-dependency problems (e.g., D proposed by Malvern & Richards, 1997; MTLD by McCarthy & Jarvis, 2010). Lexical sophistication is commonly measured by ratio of low-frequency words to total words (e.g., Daller, van Hout, & Treffers-Daller, 2003; Kojima & Yamashita, 2014; Laufer & Nation, 1995; Meara & Bell, 2001). Mean length of words or mean syllable numbers per word were also considered an index of lexical sophistication. Psycholinguistic property measures, such as word familiarity and word concreteness, typically implemented in Coh-Metrix (McNamara, Graesser, McCarthy, & Cai, 2014) and TAALES (Kyle, Crossley, & Berger, 2018), were also accepted as lexical sophistication measures.

#### *Fluency*

An acceptable fluency measure calculated either number or rate of words, words with particular part of speech (e.g., nouns), clauses, T-units, or sentences per text or allocated time.

#### *Cohesion*

An acceptable cohesion measure was one calculated from frequency, ratio, or type of cohesive devices (e.g., connectives or pronouns), or from lexical or grammatical overlap (lexical overlap for sentences, paragraphs, or entire text, or aspect repetition). Other cohesive indices calculated by computational tools such as Coh-Metrix (McNamara et al., 2014) or TAACO (Crossley, Kyle, & McNamara, 2016) were also accepted.

### *Subjective internal correlates*

To be included as a measure of subjective internal correlate, human raters had to assess discrete skills of L2 writing performance intuitively. Discrete features included 10 constructs: content, argument, organization, cohesion, coherence, grammar, vocabulary, language use, mechanics, and accuracy. The dimensions of content, argument, organization, and coherence included human rated scores of content, argumentation, organization, and coherence, respectively. Acceptable cohesion measures included human rated scores of cohesion and cohesion devices. As for grammar, we accepted human rated scores of grammar, syntax, syntactic structure, and sentence variability. The dimension of vocabulary included human rated scores of vocabulary range, lexical variation, lexical sophistication, and lexical proficiency. An acceptable language use measure included human rated language or language use scores. As for mechanics, we accepted human rated mechanics, spelling, and punctuation scores. An acceptable accuracy measure included human rated scores of language accuracy or language appropriateness.

## References for Appendix B

- Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*, 48(4), 1227–1237. <https://doi.org/10.3758/s13428-015-0651-7>
- Daller, H., Van Hout, R., & Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals. *Applied linguistics*, 24(2), 197–222. <https://doi.org/10.1093/applin/24.2.197>
- Kojima, M., & Yamashita, J. (2014). Reliability of lexical richness measures based on word lists in short second language productions. *System*, 42, 23–33. <https://doi.org/10.1016/j.system.2013.10.019>
- Kyle, K., Crossley, S. A., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*, 50(3), 1030–1046. <https://doi.org/10.3758/s13428-017-0924-4>
- Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, 16(3), 307–322. <https://doi.org/10.1093/applin/16.3.307>
- Malvern, D., & Richards, B. (1997). A new measure of lexical diversity. In A. Ryan & A. Wray (Eds.), *Evolving models of language* (pp. 58–71). Multilingual Matters.
- McCarthy, P. M., & Jarvis, S. (2010). MTL-D, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods* 42(4), 381–392. <https://doi.org/10.3758/BRM.42.2.381>
- McNamara, D. S., Graesser, A. C., McCarthy, P., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511894664>
- Meara, P., & Bell, H. (2001). P\_Lex: A simple and effective way of describing the lexical characteristics of short L2 texts. *Prospect*, 16(3), 5–19.

## L2 writing and its external correlates

### A meta-analysis

Masumi Kojima, Yo In'nami and Taku Kaneta

Gifu City Women's University / Chuo University /  
Teikyo University of Science

The present meta-analysis examined the overall average correlation between second/foreign language (L2) writing performance and each of 11 key predictor variables frequently investigated in the research domain, and compared the relative importance of those variables. A series of moderator analyses were also carried out to examine the effects of age, L2 proficiency, first language (L1) and L2 distance, learning context, and measurement characteristics for six high-evidence correlates (L2 grammar, L2 vocabulary, L2 reading, L2 speaking, L1 writing, and motivational constructs). By doing so, we examined various L2 proficiency models, writing models, and hypotheses, particularly the core-periphery model proposed by Hulstijn (2015). To this end, a total of 103 retrieved studies contributed 377 effect sizes based on 112,475 independent participants. The results showed that L2 reading and speaking achievement had strong average correlations with L2 writing performance, whereas L1 writing performance had a medium correlation with it. L2 linguistic knowledge (grammar, vocabulary, transcription, decoding) had more medium to strong effects on L2 writing performance than language-general cognitive skills and motivational constructs, which had only weak effects. The effects of metacognitive knowledge were trivial and insignificant. L2 proficiency, age, and certain measurement characteristics were found to be significant moderators for certain components.

#### 1. Introduction

Compared to second/foreign language (L2) reading, listening, and speaking research, cognitively oriented research on L2 writing is still in its infancy, and little attention has been paid to individual differences in L2 writing proficiency. Regardless, an increasing number of studies have investigated the relationship between L2 writing ability and its component skills (Oh, Lee, & Moon, 2015; Schoonen, 2019; Schoonen, van Gelderen, Stoel, Hulstijn, & de Glopper, 2011). Writing is a complex process that requires the co-ordination of various sub-processes and resources

including linguistic, discourse, sociolinguistic, and meta-cognitive knowledge, as well as strategic competence (Grabe & Kaplan, 1996; Hayes, 2012). In order to understand the nature of writing ability and individual differences, it is useful to hypothesize specific writing sub-skills or components and to investigate writing via those components. This is known as the component skills approach, which has become popular in various areas of L2 performance (e.g., De Jong, Steinel, Florijn, Schoonen, & Hulstijn, 2012; Jeon & Yamashita, 2014; Schoonen et al., 2011). The component skills approach helps to identify key predictors of L2 performance and compare their relative importance. If predictive relationships can be found between L2 writing proficiency and its sub-skills, then this knowledge will provide useful information for understanding the construct of writing ability and its component sub-skills, and will help researchers and practitioners to improve the effectiveness of future writing instructions, assessments, and research.

The aim of the present meta-analysis is two-fold. First, this study synthesizes the correlations between L2 writing performance and each of its components taken from primary studies. By doing so, we can examine various L2 proficiency models (Bachman & Palmer, 1996; Canale & Swain, 1980; Hulstijn, 2015) and first language (L1) and L2 writing models (Grabe & Kaplan, 1996; Hayes, 2012; Kellogg, 1996; Zimmerman & Risemberg, 1997), particularly the core-periphery model of Hulstijn (2015). Hulstijn hypothesizes that L2 linguistic knowledge and speed in the phonetic-phonological, morpho-phonological, morpho-syntactic, and lexical/pragmatic domains are the core components of L2 proficiency, whereas more language-general cognitive and metacognitive knowledge, such as strategic competencies, are peripheral components. The present meta-analysis goes a step further from the existing narrative reviews to test Hulstijn's hypothesis systematically. Second, we investigate the effects of some potential moderators, which have been posited in the literature to affect the strength of association between L2 writing proficiency and its correlates. By doing so, we can examine various hypotheses including the threshold hypothesis (Alderson, 1984; Schoonen, Snellings, Stevenson, & Van Gelderen, 2009), which posits an L2 proficiency threshold that is to be attained to successfully apply L1 experiences and metacognitive knowledge to L2 performance.

## **2. Background to the meta-analysis**

The present meta-analysis focuses on 11 L2 writing external correlates that are reviewed in this section. Correlation does not imply causation, but it makes it possible to predict the value of one variable from the value of another. Thus, we call the 11 correlates predictor variables, which were initially taken from the literature and from L1 and L2 writing models (Grabe & Kaplan, 1996; Hayes, 1996,

2012; Kellogg, 1996; Leki, Cumming, & Silva, 2008; Zimmerman & Risemberg, 1997), and then revised in the process of the literature search undertaken for the present meta-analysis. Thus, the 11 correlates have been frequently investigated in previous studies, but not all writing components have been covered. They can be divided into L2 linguistic knowledge (grammar, vocabulary, transcription, decoding); cognitive and metacognitive ability (language aptitude, working memory, and metacognitive knowledge); motivational constructs including intrinsic/integrative motivation, attitude, self-efficacy, goal orientation, and anxiety; and companion proficiency variables (L2 reading, L2 speaking, and L1 writing). Six of the 11 correlates (L2 grammar, L2 vocabulary, L2 reading, L2 speaking, L1 writing, and motivational constructs) have been investigated more frequently than the other five (i.e., L2 transcription, L2 decoding, language aptitude, working memory, and metacognitive knowledge). Thus, we call the former *high-evidence correlates* and the latter *low-evidence correlates*. We found some other low-evidence correlates in the literature. These include L2 listening comprehension, L2 phonological awareness, L1 reading comprehension, L1 vocabulary knowledge, L1 grammar knowledge, L1 decoding skills and various dimensions of L2 speaking performance (e.g., pronunciation, fluency, accuracy, and complexity). Despite their importance, they were not included in the present meta-analysis because they are considered components of other skills (i.e., L2 reading, L1 writing, and L2 speaking), and thus, their roles in L2 writing proficiency appear to be rather indirect. The 11 chosen correlates were assessed and scored independently from writing performance itself in primary studies. Thus, we call them *external correlates* and differentiate them from *internal correlates* which are taken and assessed as part of writing performance itself (e.g., writing speed, manner of revision, accuracy of written text). For the analysis of and discussion on writing internal correlates, see Chapter 5.

Various variables can affect the relationship between L2 writing proficiency and its correlates. Nevertheless, the quality of descriptions for moderator variables in the primary studies is often insufficient for coding and classifying them, as some meta-analysts have pointed out (e.g., Jeon & Yamashita, 2014; Norris & Ortega, 2000). Thus, the present study chose five moderators from the literature. These were age of L2 writers, their L2 proficiency, L1–L2 language distance, L2 learning environment, and various measurement characteristics. We decided to conduct a series of moderator analyses only for the high-evidence correlates in order to maintain the adequate statistical power of the analyses.

## 2.1 Review of high-evidence correlates

### 2.1.1 *L2 grammar*

Various L2 proficiency models proposed so far include L2 grammatical competence (Bachman & Palmer, 1996; Canale & Swain, 1980; Carroll, 1972; Hulstijn, 2015). In particular, written texts tend to be characterized by longer clauses and more syntactically complex language than spoken texts (Weigle, 2002). Grammar knowledge is, therefore, considered a crucial part of L2 writing proficiency. Several studies have reported that more proficient L2 writers produce more syntactically complex sentences than less proficient L2 writers (Bulté & Housen, 2014; Flahive & Snow, 1980). In many L2 writing assessment schemes, syntactic variety and grammatical appropriateness are described as premises of good writing (Weigle, 2002). Although there have been few real experimental intervention studies to prove a cause-and-effect relationship between L2 grammar knowledge and L2 writing proficiency, without grammar knowledge it is impossible to build sentences in a meaningful way. The study of Roca de Larios, Manchón, Murphy, and Marín (2008) demonstrated that their L2 learners spent about 60% of total composing time on language formulation, and less proficient L2 writers devoted more than 80% of their time to it, leaving them with little time for planning and revision. Their study and the inhibition hypothesis of Schoonen et al. (2003, 2009) suggest that insufficient grammar knowledge and effortful sentence construction can be big obstacles for nonproficient L2 writers and can inhibit them from paying more attention to higher-level conceptual aspects of L2 writing, such as planning and revision.

While Hulstijn (2015) classified components of L2 proficiency dichotomously (i.e., core vs. periphery), and classified morphosyntactic knowledge and speed along with lexical counterparts as core components of L2 proficiency, Schoonen et al. (2009) summarized the results of their NELSON projects and stated that grammatical knowledge and processing speed were more strongly related to writing proficiency than lexical proficiency and processing speeds. These studies suggest that there may be differences within core components.

The term *grammar* can refer to a wide range of linguistic knowledge, and there are various kinds of grammar tests (Purpura, 2004). However, our meta-analysis restricts the term to refer to morphosyntactic knowledge (e.g., tense, aspect, article, word order, part of speech), since in L2 writing research, morphosyntactic knowledge is most commonly referred to in grammar tests.

The present study examines the size of the integrated correlation between L2 grammar and L2 writing performance derived from past studies and compares it with those of other predictor variables including L2 lexical knowledge. We also examine the effects of theoretically motivated moderator variables which we will discuss in our review of moderator variables.



### 2.1.2 *L2 vocabulary*

A significant feature of written texts, especially academic texts, is that they tend to contain a wider variety of words, and words that are more sophisticated, than in spoken texts (Weigle, 2002). In many L2 writing assessment schemes, lexical variation and appropriacy are evaluated as part of their holistic or analytic assessment (Weigle, 2002). However, acquiring a rich vocabulary is a big challenge for L2 learners. L2 learners differ most noticeably from native speakers in terms of the vocabulary they use in written texts (Laufer, 1998). There has been incremental evidence to suggest that more proficient L2 writers display richer vocabularies in the written texts that they produce than less proficient L2 writers (Bulté & Housen, 2014; Kyle & Crossley, 2016).

Although there may be a reciprocal relationship between L2 vocabulary and L2 writing, L2 vocabulary knowledge is undoubtedly an essential component of L2 writing. Schoonen et al. (2009) argue that vocabulary knowledge facilitates writing in two ways. First, it may help L2 writers to retrieve an appropriate word without having to simplify the concept they want to express. In this way, the quality of the language formulation improves directly and attention does not need to be devoted to producing simplified expressions. Second, a rich vocabulary may enable a faster lexical retrieval process, whereby a number of tentative language formulations can be generated rapidly, freeing the writer to pay more attention to selecting the most appropriate one. The writer may also devote more attention to various aspects of writing, such as discourse organization and text content. As a result, the quality of writing will be improved. Several studies have supported the view of Schoonen et al. and demonstrated the importance of lexical knowledge and efficient language formulation in L2 writing (Roca de Larios et al., 2008; Schoonen et al., 2003).

Vocabulary knowledge is multidimensional, and thus, our meta-analysis accepts a wide range of lexical knowledge dimensions and their measurements. There are basically two issues related to vocabulary measurement methods: assessing vocabulary knowledge breadth vs. depth, and elicitation of productive responses vs. receptive ones (Nation, 2013; Read, 2000; Schmitt, 2010). Few studies have compared the links between L2 writing performance and various aspects of L2 vocabulary knowledge.

One of the focuses in our research is to examine the moderating effects of certain variables including vocabulary measurement methods. Another focus is to examine the hypothesis presented by Hulstijn (2015) positing L2 linguistic cognition, including vocabulary knowledge, as a core component of L2 performance. To test Hulstijn's hypothesis, we will synthesize correlations between L2 vocabulary and L2 writing performance derived from past studies and compare them with other integrated correlations between L2 writing and various predictor variables.



### 2.1.3 *L2 reading*

The relationship between L2 reading and writing skills has been studied extensively (e.g., Csapó & Nikolov, 2009; Schoonen, 2019). According to Fitzgerald and Shanahan (2000), readers and writers rely on four common knowledge bases, namely: (1) domain or content knowledge, (2) metaknowledge about written language, including pragmatics, and (3) knowledge about text attributes, including grapheme and morphological knowledge, lexical and syntactic knowledge, and (4) procedural knowledge (i.e., how to access, use and generate knowledge) in any of the previous three areas. These shared resources can be potential sources of overlap between reading and writing proficiency. Furthermore, the writing models of Hayes (1996, 2012) suggest that writers read the texts they have written, evaluate their appropriateness, and revise them if necessary. In the case of a writing assignment, it is obvious that writers have to read the assigned texts. In other words, reading itself can be a part of the overall writing process.

However, there are some differences between the two literacy skills. Shanahan (2016) points out that asymmetry exists between reading and writing, and they are not just inverse versions of each other. This means that the teaching of either one on its own will be insufficient because of their unique natures. One popular area of research interest has therefore been whether writing is improved by reading or reading is improved by writing, or whether there are bidirectional influences (Hirvela & Belcher, 2016; Shanahan, 2016). Shanahan (2016) argues that various connections between reading and writing appear to be bidirectional. For example, the word recognition abilities of reading can influence the spelling skills of writing, and learning to spell will influence the word recognition skills of reading. Thus, a mutual relationship is supposed to exist between the two literacy skills.

The present study focuses on how L2 reading-writing relationships differ across different developmental stages and different learning circumstances. We will discuss these issues in our review of the moderator variables.

### 2.1.4 *L2 speaking*

As writing is a form of language production, some of its cognitive processes are considered to be parallel to speaking, such as message generation, linguistic formulation, and monitoring (Hayes, 1996, 2012; Levelt, 1989). Various studies have reported positive correlations between L2 writing and speaking performance (e.g., Harrison, Gogean, Jalbert, McManus, Sinclair, & Spurling, 2016; Sparks, Humbach, Patton, & Ganschow, 2011).

However, there are some differences between the two. One major difference is the means of the physical production of formulated language, namely transcription for writing and articulation for speaking. If we adopt a simple view of writing (Juel, 1988), writing skills will be predicted by the combination of the writer's

speaking skills and transcription skills. However, the simple view of writing seems to be too simple because there are other differences between speaking and writing processes. For example, the pressure to produce language instantaneously is generally lower when writing than when speaking, and a writer usually has more time for planning, idea generation, language formulation (e.g., choosing words and syntactic structures), and revision than a speaker (Schoonen et al., 2009; Weigle, 2002). Furthermore, planning for writing is undertaken at multiple levels for longer stretches of text than planning for speaking, which is more locally determined (Schoonen, this volume). Revisions of written texts are also typically more elaborate. As a result, written texts are expected to be more grammatically correct, and pragmatically coherent and adequate, than spoken texts (Schoonen et al., 2009; Weigle, 2002).

One of our research interests is to compare the relative importance of L2 speaking and L1 writing to predict L2 writing skills. L1 and L2 writing share the same modality (written) but L2 writing and speaking do not. By comparing the two predictor variables, we can examine which is more important for L2 writing proficiency, writing skills or more general L2 proficiency. Another focus of the present study is to explore the moderating effects of certain variables that have been frequently focused on in the literature. We will discuss these issues in our review of moderator variables.

### 2.1.5 *L1 writing*

There has been cumulative evidence to suggest that L2 writing ability could be predicted by L1 writing ability (Hirose, 2003; Hirose & Sasaki, 1994; Kamimura, 1996; Leki et al., 2008). Cross-linguistic similarities between L1 and L2 writing within writers have been reported, for example, in strategies for planning, problem solving, and revision (Manchón & Roca de Larios, 2007; Stevenson, Schoonen, & De Glopper, 2006; Whalen & Ménard, 1995). As Cummins and Swain (2016) proposed, there seems to be a common underlying proficiency for L1 and L2 literacy skills.

One of the focuses of our research is to compare the relative importance of L1 writing skills on L2 writing performance with other L2 specific variables such as grammar and vocabulary knowledge, transcription, decoding skills, and L2 reading and speaking skills. Such a comparison will answer the question as to whether L2 writing is a writing problem or a language problem, which is akin to the well-known question originally presented by Alderson (1984) in L2 reading research. If the effect of L1 writing is stronger, writing problems will have a stronger effect on L2 writing ability, but if L2 specific variables have a stronger effect, language problems will be more serious. However, this relationship may not be so straightforward and may be affected by various variables including L2 proficiency. We will discuss these issues in our review of the moderator variables.

### 2.1.6 *Motivational constructs*

L2 writing is a demanding task, and learning to write in L2 can be frustrating and difficult for many L2 learners. Thus, motivation has been recognized as an important internal cause of variability in L2 writing success. Various writing models both in L1 and L2 include motivation as a component (Grabe & Kaplan, 1996; Hayes, 1996, 2012; Zimmerman & Risemberg, 1997). Hayes (1996, 2012) added motivation to his older model (Hayes & Flower, 1980) based on the observation that “whether people write, how long they write, and how much they attend to the quality of what they write will depend on their motivation” (Hayes, 2012, p. 373).

Motivation is a complex concept that has been studied from multiple theoretical perspectives (Dörnyei, 2003; Murphy & Alexander, 2000). Dörnyei (2000) proposed the process model of L2 motivation, which divides L2 motivation into pre-actional, actional, and post-actional phases, each shaped by various internal and contextual motivational influences including goal properties, attitudes towards the L2 and its speakers, expectancy of success, learner beliefs, and attributional factors. Graham (2006) reviewed studies on the relationship between L1 writing and motivation and stated that they were limited mostly to attitudes about writing, self-efficacy, interest, writing apprehension, and attributions for writing success. His data supported the view that motivational differences exist between more and less skilled writers, and that motivation helps shape writing development.

These motivational traits in quantitative studies are usually examined using questionnaires that ask individuals to rate their specific tendencies and thoughts, but we need to remember that motivation can be more unstable and dynamic than other L2 writing sub-skills such as linguistic knowledge (Ushioda & Dörnyei, 2012). For example, a person can be highly motivated on one occasion but not on others. Furthermore, various factors including social, cultural, and situational ones, have been reported to affect L2 writers' motivation (Nicolas-Conesa, Roca de Larios, & Coyle, 2014; Sasaki, 2009). In contrast, an individual's linguistic knowledge is unlikely to fluctuate drastically due to the external or internal circumstances of the L2 writer. We tried to capture these dynamic aspects of motivation in our moderator analyses (see the Review of moderator variables section).

The literature review by Graham (2006) suggests that motivation helps shape L1 writing development, but its effect was weaker than the forces of other writing skills, knowledge, and self-regulation. The present meta-analysis examines the relative effect of motivation on L2 writing performance by comparing other writing sub-skills and knowledge.

## 2.2 Review of low-evidence correlates

### 2.2.1 *L2 transcription*

According to the simple view of writing (Juel, 1988), writing is composed of two basic factors, transcription and ideation. Ideation is the ability to generate and organize ideas, and transcription is comprised of orthographic knowledge and motor skills, such as handwriting and keyboarding. Transcription is also considered as a writing component in the models proposed by Hayes (1996, 2012). Hayes added transcription to his older model (Hayes & Flower, 1980), because several studies revealed that it played a critical role in the writing processes of both L1 children and adults. For example, when children write, if most of their attention has to be focused on spelling, fewer cognitive resources will be available for higher-level processes, such as planning and evaluation. Hayes' argument has been supported by several studies (Christensen, 2004; Jones & Christensen, 1999) showing that handwriting and typing practice improved the quality of L1 children's written texts.

Spelling and orthography are also considered to be important sub-skills of L2 writing. The taxonomy of L2 writing component skills presented by Grabe and Kaplan (1996) also includes them. Spelling accuracy is associated with grapho-phonemic awareness, orthographic and phonological coding, which are integrated parts of lexical and syntactic knowledge. However, Schoonen et al. (2003) showed that orthographic knowledge made a unique contribution to L2 writing performance among other forms of linguistic knowledge, including lexical and syntactic one in their SEM analysis. Tests of L2 spelling, handwriting fluency, and accuracy have been employed to tap into L2 transcription skills. They tend to have a moderate correlation with L2 writing performance (e.g., Harrison et al., 2016; Sparks et al., 2011). A focus of this study is to examine the magnitude of the integrated correlation between L2 transcription and L2 writing performance and compare it with those derived from other predictor variables.

### 2.2.2 *L2 decoding*

L2 decoding is often referred to as a component of L2 reading (Jeon & Yamashita, 2014). However, decoding has also been examined in studies of early L1 and L2 writing development, based on an assumption that decoding predicts spelling ability, which is crucial for writing (e.g., Harrison et al., 2016). The writing models of Hayes (1996, 2012) include reading processes because a writer reads the text he or she has written, and then evaluates and revises it. Thus, decoding is also a sub-process of writing. In this study, we will examine the relative importance of L2 decoding on L2 writing performance and compare its gravity with other predictor variables.

### 2.2.3 *Language aptitude*

Language aptitude is considered to be one of the factors that affect L2 writing development (Kormos, 2012; Leki et al., 2008). According to Carroll (1981), foreign language aptitude is distinct from other cognitive abilities, including intelligence, and has four factors: phonemic coding ability, inductive learning ability, grammatical sensitivity, and rote learning ability. Kormos (2012) made some hypothetical assumptions about the role of aptitude in L2 writing processes and speculated that high aptitude L2 writers would be capable of more efficient and accurate spelling and grammatical encoding, and would have a wider L2 vocabulary repertoire because of their high levels of phonological and grammatical sensitivity, metalinguistic awareness, and good rote learning ability compared to L2 writers with lower levels of aptitude. Thus, higher aptitude will contribute to better writing.

The most frequently used aptitude test is the Modern Language Aptitude Test (MLAT) (Carroll & Sapon, 1959), while another test is the Pimsleur Language Aptitude Battery (PLAB) (Pimsleur, 1966). These tests consist of sub-tests, and a composite score of the sub-tests has been used to represent a unitary concept of language aptitude. Our meta-analysis only accepted those composite scores for inclusion.

Li (2016) conducted a meta-analysis aggregating correlations reported in the existing research between language aptitude and several individual difference variables, as well as L2 achievement. His results showed that the effect sizes for writing were mostly small and insignificant, whereas L2 reading, listening, and speaking skills had significant correlations with several aptitude tests. These results might be attributable to his definition of L2 writing proficiency; he accepted various measures as a criterion variable including discrete scores of writing performance such as language accuracy and composite scores of text-based computational indices, as well as subjective holistic scores. The present study accepted holistic scores of overall writing quality, or composite scores of analytic rating dimensions, including both linguistic and rhetorical ones, or writing scores of standardized proficiency tests only, and compared the relative importance of aptitude with those predictor variables of L2 writing performance.

### 2.2.4 *Working memory*

Working memory is considered to play a central role in writing (Hayes, 1996) and the most widely accepted model of working memory was developed by Baddeley (1986, 2007). The model consists of the central executive, which controls and regulates the three subsystems: the phonological loop, the visuo-spatial sketchpad, and the episodic buffer. According to Kellogg (1996), a writer draws ideas from long-term memory, selects appropriate ones, and organizes them using working memory, especially the visuo-spatial sketchpad. When translating ideas into

sentences and evaluating and revising them, a writer also uses working memory, especially the phonological loop. Several studies have supported Kellogg's model, such as those of Olive (2004, 2012). Limited working memory is considered to be one reason why L1 children follow the simple knowledge-telling model, which lacks planning and reviewing processes (Bereiter & Scardamalia, 1987). Thus, working memory is considered to play an important role in writing development and individual differences in writing (MacArthur & Graham, 2016).

Although an increasing number of L1 writing studies assess the relative importance of the subsystems of working memory using various kinds of tests (e.g., Vanderberg & Swanson, 2007), L2 writing studies have mainly employed phonological short-term memory measures (e.g., word/nonword repetition tests, digit span tests) and listening and reading span tests. Only the latter requires test takers to store information and process it at the same time. Those studies reported small to medium correlations between working memory and L2 writing performance (e.g., Service & Kohonen, 1995; Speciale, Ellis, & Bywater, 2004). The present study examines the integrated average correlations between them and compares the importance of working memory on L2 writing with other predictor variables, especially L2 linguistic knowledge variables.

### 2.2.5 *Metacognitive knowledge*

Metacognitive knowledge such as knowledge about audiences, writing tasks, and writing strategies has been recognized as a component of writing processes (Grabe & Kaplan, 1996; Hayes, 1996, 2012; Zimmerman & Risemberg, 1997). L2 writing processes place high demands on writers' cognitive resources, which have to be managed by metacognition or more specifically, by writing strategies (Roca de Larios, Nicolas-Conesa, & Coyle, 2016). Studies have identified various self-regulation strategies that writers use (Graham, 2006; Manchón, 2001; Zimmerman & Risemberg, 1997). Writing strategies have been frequently investigated and discussed as a subtopic of self-regulation and of metacognitive knowledge. Self-regulation of writing refers to self-initiated thoughts, feelings, and actions that writers use to attain various writing goals (Zimmerman & Risemberg, 1997). Manchón (2001) points out the conceptual diversity of writing strategies and suggests a distinction between a broad and a narrow characterization of writing strategies. Along these lines, writing strategies have recently been narrowly conceptualized as goal-oriented and problem-solving mental actions that may be effective in a specific writing task (Roca de Larios, et al., 2016). Good writers have been observed to apply appropriate strategies to their writing (Cumming, 1989; Pennington & So, 1993).

Metacognitive knowledge, including strategic knowledge, has been investigated utilizing writers' think-aloud protocol during a writing activity, and retrospective

interviews after a writing task (e.g., Roca de Larios, Marín, & Murphy, 2001), but the present study analyzed data derived from quantitative questionnaire data obtained independently from a writing task in order to examine writers' metacognitive knowledge separately from writing performance itself.

In the L2 proficiency model of Hulstijn (2015), metacognitive knowledge is conceptualized as a peripheral component. A focus of this study is to examine the magnitude of the association between metacognitive knowledge and L2 writing performance and compare its importance in L2 writing with other predictor variables, especially L2 linguistic variables.

### 2.3 Review of moderator variables

The present meta-analysis examined the effect of certain moderator variables on L2 writing performance theoretically predicted to be relevant. In this section, we present the rationales for adopting chosen moderators to the moderator analyses for each L2 writing correlate.

Regarding the relationship between L2 grammar and L2 writing performance, we investigated the effects of three potential moderators. First, L2 proficiency was chosen. As we mentioned, inefficient and poor L2 grammar knowledge can be a big obstacle, especially for low-level L2 writers (Roca de Larios et al., 2008; Schoonen et al., 2003). Thus, we hypothesized that L2 grammar knowledge might play a greater role for lower level L2 writers than for more advanced ones. Second, age was chosen. It was assumed that the strength of the correlation between L2 grammar and L2 writing would be smaller among older L2 writers than younger ones because the former usually have richer L1 literacy experience and thus can compensate for poor L2 grammar knowledge with L1 writing experience and strategies. Third, L1–L2 language distance was chosen. Although it is generally believed that certain parsing mechanisms are shared across languages, several researchers have argued that the major aspects of parsing differ across typologically different languages (Koda, 2005). Based on this view, we hypothesized that if L1–L2 language distance is short, cross-linguistic transfer of grammar knowledge will be facilitated, resulting in smaller individual differences among L2 writers and reduced correlational strength between L2 grammar knowledge and L2 writing performance than in opposite cases.

In looking at the relationship between L2 vocabulary and L2 writing performance, we chose five moderators. The first was L2 proficiency. As the studies of Roca de Larios et al. (2008) and the inhibition hypothesis of Schoonen et al. (2003, 2009) suggest, for low level L2 writers, poor L2 vocabulary can be a big obstacle to writing fluently and can inhibit them from paying more attention to higher-level conceptual aspects of L2 writing, such as planning and revision. Thus, the effect of



L2 vocabulary on L2 writing performance might be stronger among lower-level L2 writers than their more advanced counterparts. Second, we chose age as a moderator based on an assumption that adult L2 users have richer experiences in L1 literacy than younger L2 users and can compensate for a lack of L2 vocabulary with various writing strategies. As a result, the relationship between L2 vocabulary and writing performance will be weaker among adult L2 writers. Third, we assumed that L1–L2 language distance may have moderating effects because if the two languages are close, they will have more cognates, and the burden of L2 vocabulary processing and individual differences would be smaller among L2 learners with a shorter L1–L2 distance than in opposite cases.

The fourth and fifth chosen moderators were L2 vocabulary measurement methods. There are basically two issues related to vocabulary measurement methods: assessing vocabulary knowledge breadth vs. depth, and eliciting productive responses vs. receptive ones. Vocabulary breadth or size refers to the number of words a person knows (Nation, 2013; Read, 2000; Schmitt, 2010). There are receptive (selective) and productive (recall) dimensions of vocabulary size tests (Schmitt, 2010). The former usually require test takers to provide the meaning or synonym of an L2 target word either in L1 or L2, or to select a synonym or a definition that best matches the L2 target word presented either in a sentence or in isolation. The latter require participants to produce an L2 word that matches a context or a given L2 synonym or definition, or to translate an L1 word into an L2 target word, in context or in isolation. In writing, productive vocabulary knowledge is required, but receptive vocabulary tests are more frequently employed, probably because it is easy to score them. Depth or quality of vocabulary knowledge refers to how well words are known (Schmitt, 2010). An example of a depth-of-knowledge test tracks the developmental stage of L2 learners' vocabulary knowledge (e.g., Weshe & Paribakht, 1996). Another type of scale measures word associations, such as paradigmatic, syntagmatic, or collocational and analytic links between L2 target words and other L2 words (Read, 2000). These association links are tested receptively (i.e., options are given to choose) or productively (i.e., learners have to produce answers without options). Word derivation tests are also considered to assess knowledge of vocabulary depth (Schmitt, 2010).

The strength of the correlational relationship between L2 reading and L2 writing performance was posited to be potentially moderated by four variables. The first moderator was L2 proficiency. Several studies have suggested that L2 reading-writing relationships differ across different developmental stages. For example, Schoonen (2019) examined Dutch secondary school students performing reading and writing tasks in both L1 (Dutch) and L2 (English), along with various tests assessing sub-skills. Their cross-sectional results demonstrated that among the eighth graders, reading and writing in L2 were more strongly correlated than they



were in L1 and that the correlation gradually dropped to the level of the L1 correlations among the tenth graders, while L1 correlations were relatively stable across the grades. Schoonen stated that this higher correlation in L2 among eighth graders was largely attributable to linguistic knowledge resources. The results of a study by Csapó and Nikolov (2009) also showed a decline in the correlations between L2 reading and writing from grades six to 12. It can be hypothesized that among less proficient L2 learners, the influence of linguistic knowledge is stronger on both reading and writing, causing higher correlation between the two L2 skills than that of more proficient L2 learners. Second, the moderating effects of L2 learners' ages were considered. The results of Schoonen (2019) and Csapó and Nikolov (2009) can be interpreted to mean that age had moderating effects on L2 reading-writing relationships. More mature L2 learners can compensate for their lack of L2 linguistic knowledge with various strategies and topic knowledge, and thus, shared variance of L2 reading and writing caused by L2 linguistic knowledge will be smaller among them. Third, L1–L2 language distance was chosen as a moderator. In line with Koda (2005), we hypothesized that if the distance between L1 and L2 is short, cross-linguistic transfer will be facilitated. This would result in smaller individual differences in L2 linguistic knowledge among writers and reduce the correlation between L2 reading and writing caused by L2 linguistic knowledge. Fourth, learning context was chosen as a moderator. In some foreign language contexts such as in Japan, classroom activities focus a lot more on receptive skills such as reading and listening than on productive skills such as speaking and writing. In those cases, we anticipated that learners may have unevenly developed competencies depending on the modality of communication, resulting in smaller correlations between L2 reading and writing performance than in the case of second language environments.

Examining the relationship between L2 speaking and L2 writing performance, we explored the effects of four moderators. These were L2 proficiency, age, L1–L2 language distance, and learning environment. As we mentioned, we can expect a higher correlation between L2 writing and speaking caused by L2 linguistic knowledge as L2 proficiency of learners is lower, they are younger, and L1–L2 language distance is longer than in opposite cases. In cases of foreign language contexts where there is limited contact with native speakers, learners may have very limited opportunities to speak L2. Thus, unbalanced L2 language development depending on L2 skill areas might be more common in foreign language contexts than in second language contexts resulting in smaller correlations between L2 writing and speaking performance for the former than for the latter.

As for the relationship between L1–L2 writing ability, the first moderator was L2 proficiency. Some scholars have suggested that there is an L2 proficiency threshold to be attained in order to successfully apply L1 writing experiences and metacognitive knowledge to L2 writing (Sasaki & Hirose, 1996; Schoonen et al.,

2009). This threshold hypothesis was initially proposed by Alderson (1984) in L2 reading research, but it has also been applied to L2 writing. Before attaining this threshold, correlations between L1 and L2 writing performances would be weaker than after attaining it. The second moderator was age. Adult L2 learners usually have rich L1 experiences and high L1 writing skills, but that does not always entail high L2 writing skills, especially when in foreign language settings and when their L2 proficiency is low. Hulstijn (2015) also points out that studies using samples of secondary school students tend to be more heterogeneous than those with college student samples, because the latter participants are likely to have higher L1 and L2 literacy skills and be more homogeneous. In these respects, correlations between L1 and L2 skills among adult learners are expected to be weaker than those among younger learners. Third, learning context may also play a role. In a second language context, proficiency gain in L2 writing is sometimes associated with the attrition of L1 writing ability, especially when writers are young and have had less L1 education, but more L2 educational experience (Carson & Kuehn, 1992). Fourth, the strength of L1–L2 writing correlations can be moderated by L1–L2 language distance. For example, Sasaki and Hirose (1996) stated that the relatively weaker correlations between their participants' L1–L2 writing skills, compared to previous studies, might be attributable to the different rhetorical conventions between L1 and L2 (i.e., Japanese and English). Thus, L1–L2 language distance can be a moderator.

Regarding the relationship between motivational constructs and L2 writing performance, we chose five moderators. The first was learning context. Several studies have suggested that motivation for L2 writing differs significantly among learners who have lived or not lived in a place where the L2 is the dominant language and the medium of instruction (Nicolas-Conesa et al., 2014; Sasaki, 2009). The learners in those second language learning contexts generally had higher motivation to write better compositions. Based on these observations, we hypothesized that in a foreign language context, it will be difficult to maintain a high level of motivation, resulting in the greater importance of motivation than in a second language context. The second moderator was L2 proficiency. To be proficient in L2, learners need to be highly motivated and devote continuous effort to learning an L2. Conversely, learners with lower proficiency are likely to be more heterogeneous in terms of their L2 learning motivation, resulting in a higher mean correlation between motivation and L2 writing performance than for more proficient learners. The third moderator was age. Secondary school samples are in general more varied than university student samples in terms of their scholastic ability and motivation (Hulstijn, 2015), and thus the magnitude of the association between motivation and L2 writing performance will be greater for the former than for the latter. Fourth, we assumed that L1–L2 language distance may play a role. If the two languages differ greatly, more time and continuous effort is needed to be proficient in the L2.

In those cases, the importance of L2 learning motivation will be greater than in opposite cases.

The fifth moderator was the type of motivational constructs. Motivation is a complex construct but not all prominent theories of motivation have been studied in the relation to L2 writing proficiency. A preview of the data of this meta-analysis indicated that L2 writing researchers have mainly focused on intrinsic/integrative motivation, attitudes toward L2 writing or learning, goal orientation, self-efficacy, and anxiety or apprehension. According to Masgoret and Gardner (2003), motivation refers to goal-directed behavior, and attitude in L2 acquisition studies refers to at least two concepts: (1) an openness towards an L2 community, and (2) the individual's reaction to L2 learning situations, including classrooms. Goal orientation refers to reasons for L2 learning including integrative orientation and instrumental orientation. Masgoret and Gardner argue that motivation is responsible for L2 achievement, whereas attitude is indirectly related to L2 achievement through motivation. They also state that goal orientation does not necessarily reflect motivation. The results of their meta-analysis supported their argument demonstrating that the correlations between L2 achievement and motivation are uniformly higher than those between L2 achievement and attitude or goal orientation. Self-efficacy is probably the most investigated motivational construct in L2 writing research (e.g., Oh et al., 2015; Sarkhoush, 2013). Self-efficacy is defined as individuals' judgments "of their capability to organize and execute the courses of action required to attain designated types of performances" (Bandura, 1986, p. 391). In the writing model of Zimmerman and Risemberg (1997), a writer's self-efficacy is posited to influence intrinsic motivation for writing, writing goal setting, self-regulatory processes during writing, and eventual writing competence achievement. In other words, self-efficacy is supposed to be indirectly related to L2 writing proficiency through intrinsic motivation. Apprehension or anxiety of L2 writers has also been frequently studied (e.g., Sarkhoush, 2013; Sparks et al., 2011). Generally, less skilled L2 writers are observed to feel more writing apprehension than more skilled writers (Leki et al., 2008). However, Lee (2005) reported that the writing apprehension of her participants was not associated with their L2 writing performance. She discussed the reasons for this, arguing that if a participant's ability to control their anxiety was high enough, anxiety would not affect their writing performance. Based on these previous studies, the present meta-analysis hypothesizes that intrinsic/integrative motivation will be more strongly related to L2 writing performance than other motivational constructs.

### 3. Research questions for the meta-analysis

The following research questions were investigated in the present study:

1. What are the relative strengths of association between L2 writing performance and the following 11 external correlates: L2 grammar knowledge, L2 vocabulary knowledge, L2 reading, L2 speaking, L1 writing, motivational constructs, L2 transcription, L2 decoding, language aptitude, working memory, and metacognitive knowledge?
2. Do theoretically motivated moderators such as age, L2 proficiency, L1–L2 language distance, language learning context, and measurement characteristics systematically affect the relationship between L2 writing performance and its high-evidence correlates? If so, in what way?

### 4. Method

#### 4.1 Literature search and inclusion criteria

First, we electronically and manually examined the 24 most relevant journals in the field of applied linguistics, education, L1 and L2 writing, and literacy studies (see Chapter 5). During this stage, we collected studies that reported on the correlations between L2 writing performance and at least one internal or external variable. Based on this initial search and the theoretical foundation mentioned above, we targeted 11 external correlates with L2 writing performance and located relevant studies using six electronic databases (ERIC, LLBA, ProQuest Central, ProQuest D&T, Web of Science, MLA International Bibliography). We included both refereed and non-refereed (e.g., research reports, conference papers) articles. Studies published before August 2018 were searched through using various combinations of key terms related to L2 writing and 11 predictor variables using their synonyms and variations. These terms were chosen based on the initial search of the 24 journals, the thesauruses supplied in the databases, books, meta-analyses, narrative reviews in the field, and the authors' experiences. Through the journal and database search, when abstracts indicated that the articles might be relevant, full texts were retrieved for further examination. As a result, 565 full study reports were obtained and were further examined to check their eligibility. To be included in the present meta-analysis, a study had to (1) report correlations (Pearson or Spearman) between L2 writing performance and at least one of the 11 predictor variables, and (2) target non-clinical L2 learners. After finalizing the initial list of eligible studies, additional studies were also found through citation chasing of these studies, books,

meta-analyses, and narrative reviews in the field. During this stage, non-electronic papers, such as empirical studies published as book chapters and monographs, were also included. As a result, an additional 82 studies, both refereed and non-refereed, were retrieved and assessed for eligibility. After screening, the sample ultimately consisted of 103 studies (see Appendix A) with 112,475 independent participants and 377 correlations.

#### 4.2 Acceptable measures of L2 writing performance and external correlates

As we have already elaborated on the construction of writing performance and the external correlates to be examined in the present study, we only briefly describe their acceptable measures in Appendix B.

#### 4.3 Coding the primary studies

The first and third author, both of whom are experienced L2 writing researchers, coded 20 studies (19.4% of the total number of studies) independently and the agreement ratio ranged from 81% to 100% across all coded variables. Five trained research assistants then coded 74 studies, and both the first and third author double checked all of the coding and made necessary corrections. The third author coded the remaining nine studies and the first author checked his coding. Any disagreement was discussed and resolved.

As previously discussed, moderator analyses were conducted on only six high-evidence correlates in order to maintain adequate power for the intended moderator analyses. For specific features of coding moderators (age, L2 proficiency, L1–L2 distance, and language learning context), see Chapter 5. Following Jeon and Yamashita (2014), L1–L2 distance was operationalized as having two levels: Indo-European languages (coded as Shorter) and combinations of an Indo-European and a non-Indo-European language (coded as Longer). Regarding measurement characteristics, vocabulary knowledge tests were coded for two types (size vs. depth tests; production vs. recognition tests). Motivational constructs were classified into five types: anxiety, attitude, goal-orientation, intrinsic/integrative motivation, and self-efficacy. We tried to find studies on extrinsic/instrumental motivation as a correlate with L2 writing performance but were unsuccessful. For details regarding these measurement characteristics, see our review of moderator variables.

#### 4.4 Research synthesis

Initially, we coded the intact correlation coefficient values as reported by the primary researchers. Among them, some values were negative due to the nature of the measures (i.e., writing anxiety, apprehension, examining processing time instead of speed, error rates instead of accuracy rates, negative attitude instead of positive attitude). In such cases, inversed values were calculated before aggregation in a meta-analysis. For a longitudinal study that reported data collected at multiple time points, only the data from the first time points were included in the meta-analysis in order to avoid the effect of time-varying irrelevant variables. Some studies reported partly duplicative data. For example, in some cases, data for all of the participants as well as for each sub-group were reported. In other cases, data based on various measures of a construct as well as their composite scores were provided. In those cases, we used the data derived from the whole group or the composite scores only, unless the sub-groups or sub-tests were classified by our hypothesized moderators. For example, if a study reported correlations between L2 writing scores and both vocabulary size and depth scores as well as their composite scores, we used the data based on the size and depth tests only because they were our hypothesized moderators.

When a primary study reported correlations between L2 writing and various sub-tests of a construct without compositing the scores, we did not average the correlations but treated them as multilevel data and conducted meta-analyses via multilevel linear mixed-effects models using R (version 3.6.3) and its metafor package (version 2.1–0) (Viechtbauer, 2010). We treated each effect size nested within a sample, which was in turn nested within a study. When the same participants were involved in different studies, they were given the same study ID and the sample ID in the analyses (i.e., Schoonen, 2019 and Schoonen et al., 2011; Service, 1992 and Service & Kohonen, 1995; Sparks et al., 2011 and Sparks, Patton, Ganschow, & Humbach, 2009 and 2012), and duplicated correlations were removed. In this way, we addressed the issues of effect size dependency (multiple effect sizes from a single study) more appropriately than a fixed-effects or random-effects model. The statistical analyses included effect size aggregation for the association between L2 writing performance and each of the 11 constructs. Fisher's  $Z$  values converted from the correlation coefficients retrieved from the primary studies were weighted by the inverse-variance method using the metafor package. The weighted  $Z$  values were then utilized to estimate an average weighted effect size and its 95% confidence interval (CI) in each aggregation. Publication or availability bias was examined using a funnel plot which displays the distribution of effect sizes obtained from primary studies. In order to test the existence of publication bias, we employed

fail-safe  $N$  (Rosenthal, 1979). Fail-safe  $N$  suggests the number of non-significant studies necessary to make the result non-significant. A  $Q$  test was employed to determine heterogeneity for each aggregation. If the  $Q$  value was significant, it indicated that possible moderators were not considered in the model. Then, we ran a series of moderator analyses, entering a hypothesized moderator in turn into the initial model. A  $Q_M$  test was used to probe whether each hypothesized moderator was statistically significant. Subset differences in a moderator variable with three levels or more were examined by changing the reference level in turn. Significance level was specified as  $p < .05$  throughout the analyses. Each integrated  $Z$  value and its 95% CI were transformed back to  $r$  when reporting the result. When interpreting the results, we followed Plonsky and Oswald (2014) and considered correlations .25, .40, and .60 for small, medium, and large effects, respectively.

## 5. Results

Table 1 shows the results for high-evidence and low-evidence correlates. Before we discuss them in turn, note that the fail-safe  $N$  was larger than the number of included correlations in most cases. For example, in the case of grammar, the fail-safe  $N$  of 74,835 was considerably larger than the number of correlations (71). This suggests that publication bias may not be a severe threat. In contrast, in the case of aptitude, although the fail-safe  $N$  was larger than the number of included correlations, it was much smaller than that of grammar, suggesting potential publication bias for analysis of the relationship between aptitude and L2 writing. All the  $Q$  tests for high-evidence correlates were significant, suggesting that possible moderators were not considered in the models. Thus, we conducted a series of moderator analyses.

Table 1. L2 writing and its external correlates: Overall, component, and subcomponent analyses

Component	Subcomponent	No. of studies	No. of independent participants	No. of dependent participants	No. of rs	R	95% CI	Min r	Max r	Test for residual heterogeneity (Q <sub>E</sub> )	Test for moderators (Q <sub>M</sub> )	Fail-safe N <sup>a</sup>
<i>High-evidence correlates</i>												
Grammar	–	20	93,900	97,658	71	.532***	.465 .592	.000	.781			74,835
	Age	18	1,958	5,716	69			.000	.781	267.057***	2.070	27,848
	Age: Adult	13	1,230	3,954	58	.515***	.433 .589	.000	.740			
	Age: Child/Adolescent	5	728	1,762	11	.616***	.498 .711	.430	.781			
	L1–L2 distance	13	1,497	3,396	28			.020	.781	159.476***	0.911	11,522
	L1–L2: Longer	8	689	1,814	15	.498***	.364 .612	.248	.710			
	L1–L2: Shorter	5	808	1,582	13	.584***	.447 .695	.020	.781			
	Proficiency	12	1,090	3,983	57			.000	.781	134.365***	4.466*	13,975
	Proficiency: Low/Intermediate	8	879	2,082	19	.579***	.485 .659	.020	.781			
Vocabulary	Proficiency: Advanced	4	211	1,901	38	.392***	.223 .538	.000	.650			
	–	20	2,473	9,233	53	.489***	.427 .546	–.101	.759			46,317
	Age	18	1,871	5,012	44			–.101	.759	221.629***	0.051	18,857
	Age: Adult	8	549	972	14	.466***	.351 .566	.110	.650			
	Age: Child/Adolescent	10	1,322	4,040	30	.482***	.392 .562	–.101	.759			
	L1–L2 distance	15	2,143	7,776	38			.110	.759	126.279***	0.002	31,050
	L1–L2: Longer	11	1,328	5,799	30	.515***	.437 .586	.110	.759			
	L1–L2: Shorter	4	815	1,977	8	.512***	.387 .619	.346	.590			
	Proficiency	12	1,383	2,683	20			.040	.629	43.037***	0.619	3,828
	Proficiency: Low/Intermediate	9	1,136	2,436	17	.443***	.362 .518	.040	.617			
	Proficiency: Advanced	3	247	247	3	.512***	.350 .645	.431	.629			
	Test method 1	19	2,389	8,189	47			–.101	.759	228.747***	0.764	38,670
	Test method 1: Size	15	2,054	6,120	26	.481***	.411 .546	.040	.650			
	Test method 1: Depth	8	579	2,069	21	.525***	.433 .607	–.101	.759			
	Test method 2	19	2,389	8,189	47			–.101	.759	204.152***	13.194***	38,670



Table 1. (continued)

Component	Subcomponent	No. of studies	No. of independent participants	No. of dependent participants	No. of rs	R	95%	CI	Min r	Max r	Test for residual heterogeneity (Q <sub>E</sub> )	Test for moderators (Q <sub>M</sub> )	Fail-safe N <sup>a</sup>
	Test method 2: Production	10	1,222	3,664	20	.539***	.478	.595	.320	.733			
	Test method 2: Recognition	16	2,010	4,525	27	.470***	.406	.529	-.101	.759			
L2 reading	–	27	98,722	114,897	41	.588***	.516	.651	-.140	.830			160,833
	Age	25	12,817	22,992	39				-.140	.830	548.819***	0.002	82,757
	Age: Adult	14	2,548	3,454	20	.592***	.485	.681	.350	.830			
	Age: Child/Adolescent	11	10,269	19,538	19	.595***	.476	.693	-.140	.812			
	L1–L2 distance	17	11,040	20,609	28				-.140	.830	431.427***	0.363	54,692
	L1–L2: Longer	11	9,892	19,378	21	.577***	.433	.692	-.140	.830			
	L1–L2: Shorter	6	1,148	1,231	7	.640***	.460	.769	.314	.812			
	Proficiency	18	10,848	20,723	29				-.140	.812	421.263***	0.274	50,458
	Proficiency: Low/Intermediate	12	10,341	19,610	20	.586***	.466	.684	-.140	.812			
	Proficiency: Advanced	6	507	1,113	9	.531***	.332	.685	.350	.640			
	Context	25	12,323	22,498	38				-.140	.830	518.657***	0.130	77,136
	Context: Foreign language	18	11,625	21,194	29	.593***	.501	.671	-.140	.830			
	Context: Second language	7	698	1,304	9	.561***	.393	.693	.351	.650			
L2 speaking	–	16	2,804	3,328	24	.605***	.533	.668	.300	.945			9,173
	Age	14	1,953	2,455	22				.300	.945	108.660***	0.695	7,070
	Age: Adult	10	1,628	2,130	18	.641***	.549	.719	.395	.945			
	Age: Child/Adolescent	4	325	325	4	.567***	.387	.706	.300	.667			
	L1–L2 distance	11	973	1,273	18				.444	.945	53.693***	0.065	3,910
	L1–L2: Longer	6	442	742	9	.653***	.547	.738	.491	.945			
	L1–L2: Shorter	5	531	531	9	.670***	.565	.754	.444	.917			
	Proficiency	6	567	769	11				.300	.917	17.817*	0.0002	753
	Proficiency: Low/Intermediate	5	357	357	8	.563***	.387	.699	.300	.803			
	Proficiency: Advanced	2	210	412	3	.561**	.226	.777	.395	.917			
	Context	15	2,001	2,503	23				.300	.945	93.145***	1.069	7,542

(continued)

Table 1. (continued)

Component	Subcomponent	No. of studies	No. of independent participants	No. of dependent participants	No. of rs	R	95% CI	Min r	Max r	Test for residual heterogeneity (Q <sub>E</sub> )	Test for moderators (Q <sub>M</sub> )	Fail-safe N <sup>a</sup>
L1 writing	Context: Foreign language	11	1,548	1,848	18	.636***	.557 .703	.444	.917			
	Context: Second language	4	453	655	5	.551***	.383 .684	.300	.945			
	–	20	1,531	2,164	40	.426***	.330 .514	–.210	.890			6,053
	Age	19	1,607	2,116	39			–.210	.890	80.613***	4.422*	6,041
	Age: Adult	16	1,272	1,301	21	.397***	.306 .481	–.210	.890			
	Age: Child/Adolescent	3	335	815	18	.570***	.432 .681	.260	.800			
	L1–L2 distance	20	1,655	2,164	40			–.210	.890	105.700***	1.746	6,053
	L1–L2: Longer	15	1,490	1,490	18	.390***	.276 .494	–.210	.780			
	L1–L2: Shorter	5	165	674	22	.531***	.344 .677	.256	.890			
	Proficiency	12	1,279	1,308	15			–.019	.890	33.104***	0.132	1,159
	Proficiency: Low/Intermediate	7	1,099	1,107	9	.465***	.346 .570	–.019	.890			
	Proficiency: Advanced	5	180	201	6	.425***	.225 .591	.150	.608			
	Context	20	1,655	2,164	40			–.210	.890	115.537***	0.272	6,053
Motivational Constructs	Context: Foreign language	15	1,436	1,465	18	.441***	.326 .543	.020	.890			
	Context: Second language	5	219	699	22	.381***	.165 .562	–.210	.800			
	–	29	3,623	10,757	73	.338**	.225 .442	–.397	.960			19,971
	Age	29	3,910	10,757	73			–.397	.960	804.855***	0.664	19,971
	Age: Adult	24	2,424	7,455	63	.316***	.189 .433	–.397	.960			
	Age: Child/Adolescent	5	1,486	3,302	10	.432**	.167 .638	.240	.641			
	L1–L2 distance	29	3,910	10,757	73			–.397	.960	812.686***	0.359	19,971
	L1–L2: Longer	16	1,738	6,174	48	.367***	.215 .503	–.397	.850			
	L1–L2: Shorter	13	2,172	4,583	25	.299**	.122 .459	–.210	.960			
	Proficiency	14	2,203	5,545	32			–.397	.960	325.783***	1.136	5,686
	Proficiency: Low/Intermediate	11	2,027	5,255	27	.329***	.163 .478	–.397	.641			

(continued)

Table 1. (continued)

Component	Subcomponent	No. of studies	No. of independent participants	No. of dependent participants	No. of rs	R	95% CI	Min <i>r</i>	Max <i>r</i>	Test for residual heterogeneity ( <i>Q<sub>E</sub></i> )	Test for moderators ( <i>Q<sub>M</sub></i> )	Fail-safe <i>N</i> <sup>a</sup>
	Proficiency: Advanced Context	3	176	290	5	.506**	.201 .722	-.160	.960			
	Context: Foreign language	29	3,910	10,757	73	.339***	.225 .444	-.397	.960	850.818***	0.082	19,971
	Context: Second language	28	3,671	10,518	71	.293	-.058 .579	-.397	.960			
	Measure	2	239	239	2			.240	.510			
	Measure: Anxiety	27	3,657	9,517	64			-.397	.960	691.765***	41.593***	17,314
	Measure: Attitude	12	1,314	2,398	17	.357b***	.217 .483	-.040	.760			
	Measure: Goal orientation	5	765	2,284	12	.384b***	.242 .510	.020	.623			
	Measure: Intrinsic/integrative motivation	4	977	2,640	7	-.232	-.450 .013	-.397	.440			
	Measure: Self-efficacy	4	239	371	9	.436***	.260 .584	.051	.463			
		13	1,539	1,824	19	.468***	.343 .578	-.210	.960			
<i>Low-evidence correlates</i>												
L2 transcription	-	10	1,152	1,918	17	.535***	.409 .641	.149	.781			4,595
L2 decoding	-	6	556	556	6	.526***	.444 .600	.340	.620			381
Aptitude	-	4	214	214	4	.281*	.028 .500	.040	.580			21
Working memory	-	9	670	1,772	23	.340***	.181 .482	-.090	.660			2,520
Metacognitive knowledge	-	10	1,164	3,347	25	.189	-.089 .439	-.560	.693			1,002

*Note*\*  $p < .05$ .\*\*  $p < .01$ .\*\*\*  $p < .001$ .a. Fail-safe *N* (Rosenthal, 1979).

b. The sign of these values was flipped if necessary, so that positive correlations suggested that learners with less anxiety and positive attitudes scored well in writing tasks.

## 6. Discussion

The primary aim of this study, as formulated in the first research question, was to examine the magnitude of the association between L2 writing performance and its 11 external correlates. By doing so, we aimed to examine the various L2 proficiency models and L1/L2 writing models proposed so far, and compare the relative importance of those 11 predictor variables. To this end, we integrated correlations between L2 writing performance and each of the 11 correlates taken from primary studies. The results revealed that L2 reading comprehension and L2 speaking performance had a large mean correlation with L2 writing performance ( $r = .588$  and  $.605$ , respectively), whereas L1 writing performance had a medium mean correlation with it ( $r = .426$ ). Furthermore, the aggregated mean correlations for L2 linguistic variables, namely L2 grammar knowledge, L2 vocabulary knowledge, L2 transcription, and L2 decoding had medium correlations with L2 writing performance ( $r = .489$  to  $.535$ ), whereas more language-general cognitive and metacognitive skills (i.e., motivational constructs, meta-cognitive knowledge, aptitude, and working memory) had only small mean correlations with it ( $r = .189$  to  $.340$ ). These figures were smaller than those reported in the meta-analysis of L2 reading and its correlates by Jeon and Yamashita (2014), but it must be noted that in our study, attenuation of correlations caused by low reliability of measurements was not corrected. Thus, the true correlations can be larger than ours.

Overall, these results supported the core-periphery model of L2 proficiency (Hulstijn, 2015), which posits that linguistic knowledge and processing speed in morpho-phonological, morpho-syntactic, and lexical domains are the core components of L2 proficiency, whereas more language-general strategic competence and metacognitive knowledge are classified as peripheral components. We will further discuss the results for each correlate below, including the results of moderator analyses as formulated in the second research question.

### 6.1 High-evidence correlates

#### 6.1.1 *L2 grammar knowledge*

The average correlation between L2 writing and grammar was significant and moderate ( $r = .532$  [95% confidence interval:  $.465, .592$ ],  $p < .001$ ). This result is consistent with various L2 proficiency models (Bachman & Palmer, 1996; Canale & Swain, 1980; Carroll, 1972; Hulstijn, 2015) and the taxonomy of L2 writing skills and processes proposed by Grabe and Kaplan (1996), which include L2 grammar knowledge as a component skill.

The result of the moderator analysis revealed that L2 proficiency was a significant moderator ( $Q_M = 4.466$ ,  $p = .035$ ) and the correlation between L2 writing performance and L2 grammar knowledge was stronger for low/intermediate learners ( $r = .579$  [.485, .659]) than for advanced learners ( $r = .392$  [.223, .538]). This result supported our hypothesis that L2 grammar knowledge plays a greater role for lower level L2 learners than for more advanced learners. It appears that the grammar knowledge of low/intermediate learners is still developing, and their individual variability is relatively large and explains the substantial amount of L2 writing proficiency. This result supports the observations of primary studies (e.g., Roca de Larios et al., 2008; Schoonen et al., 2003), which suggested that inefficient and poor L2 grammar knowledge can be a big obstacle, especially for low-level L2 writers. In contrast, the L2 grammar knowledge of advanced learners seems to be more stable and less heterogeneous, resulting in a smaller mean correlation with L2 writing performance.

In contrast, age was not a significant moderator. We assumed that the influence of L2 grammar knowledge for adult L2 writers would be smaller because they have richer L1 literacy experience, and thus can compensate for poor L2 grammar knowledge with L1 writing experience and strategies compared to younger writers. In fact, the integrated mean correlation between L2 writing and L2 grammar knowledge was larger for child/adolescent learners ( $r = .616$  for child/adolescent and .515 for adult learners), but the difference was not statistically significant. However, this result seems to be consistent with our other results, which revealed that L2 reading and speaking skills correlate with L2 writing performance more strongly than L1 writing performance does. In fact, our adult samples included low and intermediate level L2 learners, typically learning L2 in foreign language contexts. The effect of L1 literacy experience may not have been strong enough to compensate for their poor L2 grammar knowledge.

L1–L2 language distance was not a significant moderator either. We hypothesized that if L1–L2 distance was short, cross-linguistic transfer of grammar knowledge would be facilitated resulting in smaller individual differences and a mean correlation between L2 grammar knowledge and L2 writing performance. Contrary to our expectations, the integrated mean correlation was higher among L2 learners with closer L1 ( $r = .584$ ) than in opposite cases ( $r = .498$ ), although the difference was statistically insignificant. As Jeon and Yamashita (2014) pointed out in their meta-analysis of L2 reading and its correlates, even within the same language family, morphosyntactic systems of different language can vary greatly and can cause individual variability of L2 grammar knowledge.

### 6.1.2 L2 vocabulary knowledge

The aggregated mean correlation between L2 vocabulary knowledge and L2 writing performance was significant and moderate ( $r = .489$  [.427, .546],  $p < .001$ ). This result is in accordance with various L2 proficiency models (Bachman & Palmer, 1996; Canale & Swain, 1980; Carroll, 1972; Hulstijn, 2015) and the taxonomy of L2 writing skills proposed by Grabe and Kaplan (1996). Along with L2 grammar knowledge, L2 vocabulary knowledge seems to be an important component of L2 writing proficiency. Schoonen et al. (2009) summarized the results of the NELSON project and stated that L2 grammatical knowledge and processing speed were more strongly associated with L2 writing proficiency than lexical knowledge and processing speed. Our results also demonstrated that the mean correlation for L2 grammar knowledge, including speed measures, was slightly stronger than that of L2 lexical knowledge and speed measures. However, because their 95% CIs largely overlapped, there were no significant differences between the two. Thus, this study concludes that they are equally important components of L2 writing proficiency.

We also conducted moderator analyses for L2 proficiency, age, L1–L2 language distance, and measurement characteristics. First, L2 proficiency was not a significant moderator. On the basis of several studies (e.g., Roca de Larios et al., 2008; Schoonen et al., 2003), we assumed that vocabulary can be a big obstacle to writing efficiently for low/intermediate L2 learners, making the role of vocabulary more significant for them than for advanced learners. Contrary to our expectations, the integrated mean effect size was slightly higher for the advanced learners ( $r = .512$ ) than for the low/intermediate ones ( $r = .443$ ), although the difference was statistically insignificant. This result suggests that vocabulary is also important for advanced learners whose writing tasks were typically academic ones requiring low-frequency, sophisticated vocabulary. As this type of vocabulary knowledge can vary greatly across individuals, even after reaching advanced proficiency, L2 vocabulary development seems to continue and correlate with writing performance moderately. Our results indicated that the extent of L2 vocabulary can predict L2 writing performance as much among advanced L2 learners as among low/intermediate learners.

Age was not a significant moderator either ( $r = .466$  for adult writers and  $r = .482$  for child/adolescent writers). We hypothesized that adults can compensate for poor L2 vocabulary with rich L1 literacy experience and strategies, making the effect of L2 vocabulary on writing smaller for them. However, as our results showed that the effect of L2 proficiency (i.e., L2 reading and speaking) on L2 writing ability was greater than that of L1 writing, the influence of L1 literacy experience seems not to be as large as we had expected.

L1–L2 distance was not a significant moderator, and thus the hypothesized cognate effect was not observed in our study ( $r = .512$  and  $.515$  for each group). In fact, this null result is in accordance with the meta-analyses of Jeon and Yamashita

(2014) and Melby-Lervåg and Lervåg (2011), who aggregated correlations between L2 reading and various variables including L2 vocabulary knowledge. They concluded that cross-linguistic effects are more likely captured in simpler language processes such as decoding, but not in more complex language variables such as vocabulary. Although we did not investigate the moderating effect of L1–L2 language distance on the relationship between L2 decoding and writing performance because of the small number of primary studies, their insight might be applied to this study too.

As for the measurement characteristics, the integrated mean correlation was significantly higher ( $Q_M = 13.194$ ,  $p < .001$ ) for productive vocabulary measures ( $r = .539$  [.478, .595]) than for receptive tests ( $r = .470$  [.406, .529]). This result is not surprising considering that writing is a productive task and that productive vocabulary measures should be closer to real lexical processing in writing and thus should better predict L2 writing performance than receptive vocabulary measures. However, the number of primary studies which employed receptive vocabulary tests was greater (16 studies) than that of productive vocabulary tests (10 studies), despite the importance of productive vocabulary knowledge in writing. This is probably because receptive vocabulary tests are easier to administer and score than productive tests. However, our results warn researchers to choose an appropriate measurement in accordance with their study purposes.

Regarding the measurement types of size and depth, there was no significant difference between them ( $r = .481$  and  $.525$ , respectively). This result might not be surprising. Written texts, especially academic texts usually contain a wide variety of words including low-frequency words (Weigle, 2002). Thus, L2 vocabulary size is crucial for proficiency in L2 writing. However, in order to choose an appropriate word in a particular context, a writer needs to know its part of speech, its register, and the collocational behavior relating to it. Therefore, depth of vocabulary knowledge is also important and seems to predict individual differences in L2 writing performance as much as vocabulary size does.

### 6.1.3 *L2 reading*

L2 reading comprehension had a relatively large mean correlation with L2 writing performance ( $r = .588$  [.516, .651],  $p < .001$ ). This result supports various studies which assume that reading and writing share various common knowledge bases including linguistic knowledge and orthographic processing, and that they have mutual relationships with each other (Fitzgerald & Shanahan, 2000; Schoonen, 2019; Shanahan, 2016). As the writing model of Hayes (2012) suggests, reading skills can be a premise of writing proficiency.

Regarding the results of moderator analyses, L2 proficiency did not have a significant effect on the association between L2 reading and writing performance.

Based on the results of Schoonen (2019) and Csapó and Nikolov (2009), the present study hypothesized that for lower proficiency L2 learners, the influence of L2 linguistic knowledge would be greater, causing a higher correlation between L2 reading and writing proficiency. In fact, the correlation of the two L2 skill areas was stronger for the low/intermediate writers ( $r = .586$ ) than for the advanced L2 writers ( $r = .531$ ), but there was no statistically significant difference between the two. However, this result seems to be in accordance with our results of moderator analysis for L2 grammar and vocabulary. L2 grammar knowledge was a stronger predictor of L2 writing performance for low/intermediate learners than for advanced ones, whereas L2 vocabulary had a larger mean correlation with L2 writing for advanced learners than for low/intermediate ones, though the latter difference was statistically insignificant. The results of Jeon and Yamashita (2014) also showed that L2 vocabulary knowledge was a significantly stronger predictor of L2 reading comprehension for their adolescent/adult samples than for their child samples. Considering that their adolescent/adult learners mainly consisted of college and postgraduate students, they were more advanced L2 learners than the child learners. Therefore, these results suggest that some proportion of shared variance between L2 reading and writing performance for the advanced learners seems to be caused by L2 vocabulary knowledge, whereas for the low/intermediate learners, the main cause of the shared variance seems to be L2 grammar knowledge. On the whole, the size of the correlation between L2 reading and writing was not significantly different for the two groups of L2 learners, but the causes might have been different.

Age was not a significant moderator either ( $r = .592$  and  $.595$  for each group). We hypothesized that importance of L2 linguistic knowledge in writing would decrease as L2 learners get older, because they can compensate for their lack of L2 linguistic knowledge with various strategies and L1 writing experience. However, as our results revealed, the effect of metacognitive strategies on L2 writing was trivial, and the influence of L2 proficiency was greater than the effect of L1 writing skills. Even if adult L2 writers have a slight advantage in terms of L1 literacy experience and strategies, it seems that they were not strong enough to compensate for their lack of L2 linguistic knowledge.

Regarding L1–L2 language distance, there was no significant difference between the two conditions ( $r = .577$  and  $.640$ , respectively). Even within the same language family, processing of different languages may vary greatly, causing individual variability and shared variance between L2 reading comprehension and writing performance. Language learning environment (i.e., foreign language vs. second language) was not a significant moderator either ( $r = .593$  and  $.561$ , respectively). L2 literacy skills might be closely related to each other regardless of the language learning environment.



#### 6.1.4 L2 speaking

The effect of L2 speaking ability on L2 writing performance has not been investigated as much as that of L2 reading (i.e., 16 studies vs. 27 studies), but the magnitude of its integrated effect size was large ( $r = .605$  [.533, .668],  $p < .001$ ) and equivalent to that of L2 reading. Despite the modality difference, speaking and writing are supposed to share various cognitive processes including message generation and linguistic formulation, as the speaking and writing models of Levelt (1989) and Hayes (2012) suggest. These common processes seem to cause the shared variances between the two L2 skill areas. As the simple view of writing (Juel, 1988) suggests, a considerable amount of L2 writing proficiency seems to be predicted by L2 speaking proficiency and L2 transcription ability, which had strong and medium effects on L2 writing performance, respectively.

As for the results of the moderator analyses, neither L2 proficiency ( $r = .561$  and  $.563$  for each group) or the age of participants ( $r = .567$  and  $.641$  for each group) were significant moderators. We assumed that the dependency on L2 linguistic knowledge in L2 proficiency would be larger, when L2 proficiency of learners is lower and/or they are younger, contributing to higher correlations between L2 speaking and writing performance. However, as our results revealed, while L2 grammar knowledge had a stronger effect on L2 writing for low/intermediate learners than for advanced ones, L2 vocabulary knowledge had a slightly stronger effect for advanced learners. Speaking in general does not require low-frequency advanced vocabulary compared to writing, but a closer look at the speaking tasks revealed that while most of the child/adolescent samples worked on relatively simple tasks such as picture description, the adult samples accepted more complicated academic tasks (e.g., TOEFL and IELTS speaking tasks). These academic speaking tasks generally require elaborate language including low-frequency vocabulary. In fact, the integrated mean correlation between L2 writing and speaking performance was slightly higher for the adult learners ( $r = .641$ ) than for the younger learners ( $r = .567$ ), though the difference was statistically insignificant. Considering that our advanced samples were almost exclusively university or postgraduate students, the same interpretation can be applied to the result of age as a moderator.

L1–L2 language distance was not a significant moderator either ( $r = .653$  and  $.670$  for each condition). As we already mentioned, even within the same language family, processing of different languages can vary greatly, causing individual variability, which may become the source of shared variance between L2 speaking and writing performance. As for language learning context (i.e., foreign language vs. second language), significant moderating effects were also not observed. We hypothesized that foreign language writers would have unbalanced proficiency across different L2 skill areas, resulting in a smaller mean correlation between L2 writing and speaking performance. Contrary to our expectations, the mean correlation

between the two was strong for the foreign language samples ( $r = .636$ ) and slightly higher than that of the second language samples ( $r = .551$ ), although the difference was statistically insignificant. The studies conducted in foreign language contexts involved proportionally more low/intermediate L2 learners than in second language contexts. As our study indicated, more variance in L2 writing performance was explained by L2 grammar knowledge for low/intermediate learners than for advanced learners. Even when foreign language learners do not have many opportunities to speak the target language, their speaking performance might be substantially accounted for by L2 linguistic knowledge, which may have caused the shared variance with L2 writing performance, resulting in the strong mean correlation between the two.

### 6.1.5 *L1 writing*

The integrated mean correlation between L1 and L2 writing performance was significant and moderate ( $r = .426$  [.330, .514],  $p < .001$ ). In other words, a substantial amount of individual variability in L2 writing performance can be explained by L1 writing skills. This result supports Cummins and Swain (2016) who argue that there is a common underlying proficiency for L1 and L2 literacy skills. However, the magnitude of the correlation was significantly smaller than those between L2 writing performance and L2 reading/L2 speaking performance, as its 95% CI did not overlap with theirs. It is worth noting that although L1 and L2 writing share the same modality (written) but L2 speaking and writing do not, the effect of L2 speaking on L2 writing was significantly stronger than that of L1 writing. This result supports the notion that L2 writing ability is essentially determined by L2 language proficiency. Furthermore, the effects of L2 linguistic knowledge (grammar, vocabulary, transcription, and decoding) on L2 writing performance were generally greater than that of L1 writing performance. Specifically, the 95% CI derived from L2 grammar just slightly overlapped with that of L1 writing. Thus, we can conclude that L2 writing is rather an L2 language problem than a writing problem.

Regarding the moderator analyses, age was a significant moderator ( $Q_M = 4.422$ ,  $p = .036$ ), and the mean correlation between L1 and L2 writing was higher for the child/adolescent samples ( $r = .570$  [.432, .681]) than for the adult samples ( $r = .397$  [.306, .481]). This result supports our assumption that the former would be more heterogeneous in terms of their L1 and L2 literacy skills causing a higher mean correlation between the two variables than the latter. Despite the importance of L1 literacy skills in L2 writing proficiency for child/adolescent learners, only three studies that targeted them were available to this effect size aggregation. Most of the studies investigating the association between L1 and L2 writing performance examined adult writers (16 studies). Writing researchers need to pay more attention to the L1 literacy skills of young L2 writers in the course of their writing development.

As for L2 proficiency, based on the threshold hypothesis (Schoonen et al., 2009), we assumed that a higher integrated correlation between L1 and L2 writing performance would be observed among advanced learners than among low/intermediate learners. However, the result showed an insignificant difference between the two groups ( $r = .425$  and  $.465$ , respectively). Our definition of L2 proficiency was based on standardized test scores or descriptions of primary studies, such as the CEFR levels of participants. However, because proficiency information for elementary and secondary school students was scarce in primary studies, we coded them as low/intermediate if they were in foreign language settings. While advanced learners were almost exclusively university or postgraduate students, the low/intermediate learners were more heterogeneous and included L2 learners ranging from young children to postgraduate students. As our results show that an association between L1 and L2 writing skills was significantly stronger for the child/adolescent samples than for the adult sample, age might have affected the result. However, even if we reexamine the primary studies focusing on adult learners and compare the two proficiency groups, there seems to be no significant difference between them. In fact, examining the interactional effect is difficult because of the small number of studies. This is partly because only 37.5% of the primary studies provided L2 proficiency information on the participants. We call for more primary studies reporting the L2 proficiency of their participants.

L1–L2 language distance was not a significant moderator either. Based on the observation of Sasaki and Hirose (1996), we hypothesized that if L1–L2 language distance was short, their rhetorical conventions would be similar, and thus, the cross-linguistic transfer would be facilitated, resulting in a larger mean correlation in those cases. In fact, the integrated mean correlation was larger for the closer L1 and L2 cases ( $r = .531$  [.344, .677]) than for the more different cases ( $r = .390$  [.276, .494]), although the difference was insignificant. This can be attributable to the wide range of their 95% CIs. A closer look at the primary studies indicates that among the low/intermediate samples, the closer L1–L2 group showed higher correlations ( $r = .74$  to  $.89$ ) than in opposite cases ( $r = -.019$  to  $.564$ ), whereas for the advanced samples, the two conditions were equivalent. Cross-linguistic transfer might be facilitated when L1 and L2 are closer and learners' proficiency is lower. However, because of the small number of studies, this insight is speculative and we need to cumulate more primary studies to examine this hypothesis.

As for the language learning context, we assumed that in a second language context, proficiency gain in L2 would sometimes be associated with L1 writing ability loss (e.g., Carson & Kuehn, 1992), resulting in a smaller correlation between L1 and L2 writing performance than in cases of foreign language contexts. In fact, the aggregated mean correlation was larger in foreign language contexts ( $r = .441$  [.326, .543]) than in second language contexts ( $r = .381$  [.165, .562]), although the

difference was insignificant. When we focused on the low/intermediate samples, correlations between L1 and L2 writing skills tended to be higher in foreign language contexts ( $r = .344$  to  $.890$ ) than in second language contexts ( $r = -.019$  to  $.230$ ), whereas for the advanced learners, the two conditions were almost equivalent. If L2 proficiency of learners is lower and they are in a foreign language environment, they may rely on L1 writing experiences more than other cases. Overall, our results suggest that various factors can affect the cross-linguistic transfer between L1 and L2 literacy skills.

#### 6.1.6 *Motivational constructs*

Motivation is considered to be an important component of writing proficiency and is included in various writing models (Grabe & Kaplan, 1996; Hayes, 2012; Zimmerman & Risemberg, 1997). Our results demonstrated that the motivational factors had a small but significant mean correlation with L2 writing performance ( $r = .338$  [.225, .442],  $p < .001$ ), and thus supported these writing models. Motivation seems to be an important internal cause of individual differences in L2 writing success. However, its magnitude was significantly smaller than that of L2 linguistic components, since its 95% CI barely overlapped with theirs. Thus, despite their importance, the effect of motivational factors on L2 writing proficiency seems to be smaller than those of L2 linguistic components.

The hypothesized moderating effect of learning context was not significant. We assumed that a higher mean correlation between motivational constructs and L2 writing would be observed for the foreign language learners than for the second language learners, because it would be more difficult to maintain high-levels of motivation for the former causing more individual variability among them. In fact, the aggregated mean correlation was significant and greater for the former ( $r = .339$  [.225, .444],  $p < .001$ ) than for the latter ( $r = .293$  [-.058, .579],  $p = .101$ ) which was statistically insignificant. However, the difference between them was statistically insignificant, because the 95% CIs largely overlapped. Also, we would like to note that this result is not reliable because of the small number of studies (i.e., two studies) for the second language contexts. Motivation in L2 writing has been a popular research topic in foreign language contexts but not in second language contexts, although motivation might be equally important for L2 writers in both contexts. More studies will be needed to examine the role of motivational constructs for second language writers.

L2 proficiency was not a significant moderator either. Our assumption was that advanced learners would, in general, be highly motivated and their individual differences in terms of motivational constructs would be smaller than those of low/intermediate learners, who would be more heterogeneous. Thus, we had expected that the latter would yield a higher mean correlation between motivational constructs

and L2 writing performance. Contrary to our expectations, the aggregated mean correlation was larger for the advanced samples ( $r = .506$  [.201, .722]) than for the low/intermediate samples ( $r = .329$  [.163, .478]), although the difference was insignificant. This result seems to be attributable to the wide range of the 95% CI for the advanced learners, mainly caused by the small number of studies involved (i.e., three studies). This result demonstrates that motivation research has focused more on low/intermediate writers than advanced writers, although motivation might also be important for advanced learners. More studies will be needed to investigate the role of motivation for advanced L2 learners.

Regarding age, we hypothesized that child/adolescent learners would be more heterogeneous than adult learners who are mainly university students, in terms of their literacy skills and motivation, as suggested by Hulstijn (2015), resulting in a higher mean correlation between motivational constructs and L2 writing performance than for adult learners. In fact, the mean correlation was higher for the child/adolescent samples ( $r = .432$  [.167, .638]) than for the adult samples ( $r = .316$  [.189, .433]), although the difference was statistically insignificant. This result seems to be attributable to the wide range of 95% CI for the younger learners, mainly caused by the small number of studies (i.e., five studies). Motivation research has focused more on mature L2 writers, although motivation might be just as, if not more, important for younger learners. Overall, our results revealed that L2 researchers have investigated the effect of motivational constructs on L2 writing mainly for foreign language university students with low or intermediate proficiency. Motivation studies need to investigate more diverse L2 writers.

The fourth moderator, L1–L2 language distance, also did not have a significant effect. We assumed that if the two languages differed greatly, the importance of motivation would increase because more learning time and continuous effort would be needed to become proficient in a language substantially different from the L1. In fact, the aggregated mean correlation was larger for longer language distance ( $r = .367$  [.215, .503]) than for shorter language distance ( $r = .299$  [.122, .459]), although the difference was insignificant. Motivational constructs seem to be important even for learners with L1s with a shorter language distance from the target language.

Regarding the types of motivational constructs, intrinsic/integrative motivation and self-efficacy had significant and medium mean correlations with L2 writing performance ( $r = .436$  [.260, .584] and  $r = .468$  [.343, .578], respectively), whereas attitude and anxiety had small but significant mean correlations with it ( $r = .384$  [.242, .510] and  $r = .357$  [.217, .483], respectively). These statistically significant results supported previous studies and writing models, which claimed that writers with higher intrinsic/integrative motivation, higher self-efficacy, better attitudes toward L2 learning, and less anxiety generally demonstrate better writing

performance (Oh et al., 2015; Sarkhoush, 2013; Sparks et al., 2011). In contrast, goal-orientation had an insignificant mean correlation with L2 writing performance ( $r = -.232 [-.450, .013]$ ,  $p = .064$ ). The present study included various kinds of goal orientation, but because of the small number of primary studies, we did not conduct separate analyses for them. A closer look at the primary studies shows that even the mastery-approach to goal-orientation had a negative correlation with L2 writing performance ( $r = -.166$  in Farsani, Beikmohammadi, & Mohebbi, 2014). More studies need to be considered in order to examine the role of goal-orientation in L2 writing.

Comparing the effects of these motivational constructs on L2 writing, self-efficacy had a significantly greater effect than anxiety and goal orientation. In contrast, goal orientation had a significantly smaller correlation with L2 writing than all other constructs. These results partly supported our hypothesis which posited that intrinsic/integrative motivation would have a strongest effect on L2 writing achievement among other motivational constructs. Intrinsic/integrative motivation had a stronger effect than anxiety, attitude, and goal-orientation, but the difference was significant only in respect of goal-orientation. These insignificant results seem to be attributable to the small number of studies on intrinsic/integrative motivation (i.e., four studies) causing the wide range of its 95% CI. Motivation research has been conducted qualitatively, but more quantitative studies will be required in order to examine the hypothesis meta-analytically. However, self-efficacy might be as important as intrinsic/integrative motivation, as the gravity of their mean correlations with L2 writing performance was equivalent and their 95% CIs largely overlapped. Although Zimmerman and Risemberg (1997) posited that self-efficacy has an indirect effect on writing achievement through intrinsic motivation, our results suggest that self-efficacy may have a more direct and unique effect on L2 writing proficiency.

## 6.2 Low-evidence correlates

### 6.2.1 *L2 transcription*

The aggregated mean correlation between L2 transcription and L2 writing performance was significant and moderate ( $r = .535 [.409, .641]$ ,  $p < .001$ ) and as large as those derived from L2 grammar and vocabulary. This result supports the writing models of Hayes (2012) and Juel (1988), in which transcription is described as a crucial part of the writing process. However, it must be noted that most of the primary studies included in this analysis targeted primary and secondary school students. Due to the small number of studies, we did not conduct a moderator analysis, but L2 decoding skills might be less important for more mature L2 writers.

### 6.2.2 *L2 decoding*

This had a significant and medium mean correlation with L2 writing performance ( $r = .526$  [.444, .600],  $p < .001$ ). Although L2 decoding is usually considered a component of L2 reading, and not commonly examined in the course of L2 writing research, our results indicated that its impact on L2 writing was as large as that of L2 grammar and vocabulary knowledge. As the writing model of Hayes (2012) suggests, reading skills, including decoding, seem to be a premise of writing proficiency. However, we need to note that most of the primary studies involved in this analysis examined child/adolescent writers, and it is not clear whether the same can be said for more mature L2 learners. Further primary research is needed to examine the role of decoding in adult L2 learners during their writing development.

### 6.2.3 *Language aptitude*

Language aptitude had a small but significant effect on L2 writing performance, as the aggregated mean correlation was  $r = .281$  [.028, .500],  $p = .030$ . The difference between this study and the meta-analysis of Li (2016), which reported an insignificant integrated correlation between L2 writing performance and language aptitude ( $r = .34$ ,  $p = \text{n.s.}$ ), was that while the writing measures of Li (2016) included discrete measures of writing performance such as language accuracy and composite scores of text-based computational indices, we did not accept such measures as a criterion of L2 writing performance. Our result is consistent with the assumptions of Kormos (2012), who posited that high aptitude L2 writers have advantages in terms of language sensitivity and good rote learning ability which will contribute to efficient language processing and a wide repertoire of L2 vocabulary. However, we need to be cautious as there might be a publication bias in our data suggested by the small number of the fail-safe  $N$  (i.e., 21 for four integrated correlations). In either way, the effect of aptitude on L2 writing seems to be significantly smaller than that of L2 linguistic knowledge. This result supports the core-periphery model of L2 proficiency (Hulstijn, 2015), which posits that linguistic knowledge and processing speed are the core components of L2 proficiency, whereas language-general cognitive skills and metacognitive knowledge are classified into peripheral components.

### 6.2.4 *Working memory*

The integrated mean correlation between L2 writing performance and working memory was small but significant ( $r = .340$  [.181, .482],  $p < .001$ ). This result is consistent with various writing models (Hayes, 1996, 2012; Kellogg, 1996), which assume that working memory plays a central role when selecting and organizing ideas and translating ideas into sentences and evaluating them. However, despite its importance, the gravity of the effect is generally smaller than that of L2 linguistic components. This result is another evidence to support the core-periphery model of L2 proficiency (Hulstijn, 2015).



### 6.2.5 *Metacognitive knowledge*

The aggregated mean correlation for metacognitive knowledge, such as knowledge about audience, writing tasks, and writing strategies was small and insignificant ( $r = .189 [-.089, .439]$ ,  $p = .181$ ). This result is rather surprising because various writing models (Grabe & Kaplan, 1996; Hayes, 1996, 2012; Hayes & Flower, 1980; Zimmerman & Risemberg, 1997) as well as L2 proficiency models (Bachman & Palmer, 1996; Canale & Swain, 1980) include metacognitive components. However, this result seems to support the core-periphery model of Hulstijn (2015), which classifies metacognitive, strategic components into peripheral components. A re-examination of primary studies indicated that some writing strategies such as “thinking in L1 while writing in L2” and “use of simpler words and phrases in L2” had negative correlations with L2 writing performance (Cohen & Brooks-Carson, 2001). A more detailed study is required to examine the role of metacognitive knowledge in L2 writing proficiency.

## 7. Conclusion

The present study systematically reviewed past studies to investigate the relationship between L2 writing and its external correlates and the effects of certain hypothesized moderators. By doing so, this meta-analysis aimed to examine certain important models of L1 and L2 writing and of L2 proficiency, including the core-periphery model of Hulstijn (2015). Overall, L2 writing performance was more strongly correlated with L2 specific linguistic knowledge than more language-general cognitive skills, metacognitive knowledge, and motivational constructs. Furthermore, L2 writing proficiency correlated more strongly with L2 reading and speaking skills than L1 writing skills, although L1 and L2 writing share the same output modality. These results supported the core-periphery model of Hulstijn (2015), which argued that core components such as linguistic knowledge and its processing speed assume a central role in L2 performance. Furthermore, L2 proficiency models of Bachman and Palmer (1996) and Canale and Swain (1980), and the taxonomy of L2 writing skills proposed by Grabe and Kaplan (1996), seem to be correct to include components other than purely linguistic ones such as metacognitive knowledge, but our results indicated that those were peripheral components of L2 writing proficiency, as predicted by Hulstijn (2015). To summarize, the present study supports the notion that L2 writing performance poses a language problem rather than a writing problem.

The second aim of this meta-analysis was to identify moderator variables which systematically affect the strength of the association between L2 writing performance and the 11 targeted correlates. Our results demonstrated that proficiency was a significant moderator, and individual differences of L2 writing performance were



predicted by L2 grammar knowledge for low/intermediate learners more strongly than for advanced learners. On the contrary, the importance of L2 vocabulary knowledge seemed to be more consistent from the low to advanced stages of L2 writing development. Age was also a significant moderator, and the younger learners had a stronger average correlation between L1 and L2 writing performance than the older learners did. The threshold hypothesis of Schoonen et al. (2009) was not supported by the present study, as not only proficiency but also various factors seem to affect the association between L1 and L2 writing achievement. The problem was that an equivalent comparison could not be made in order to test the hypothesis, partly because of the lack of information on L2 proficiency of participants in primary studies. A future study needs to consider interactional effects of variables on the relationship between L1 and L2 writing achievement. To do so, more studies are called for to include information on L2 proficiency of participants. Some measurement characteristics were also significant moderators, which suggests that researchers need to carefully consider and select appropriate measurements in accordance with their research purposes.

This study also revealed some relatively under-investigated but potentially important components of L2 writing proficiency. L2 transcription and L2 decoding skills were two such components. Although they have been less frequently investigated than L2 grammar and vocabulary knowledge, their mean correlations with L2 writing performance were as large as those derived from L2 grammar and vocabulary knowledge. As the primary study of Schoonen et al. (2003) also suggests, transcription and decoding skills seem to make unique contributions to L2 writing proficiency. However, because most of the primary studies that examined the effects of L2 transcription and decoding skills targeted child/adolescent L2 learners, we do not know whether the same can be said for older learners. Therefore, we would like to encourage L2 writing researchers to devote more effort to studies on these under-investigated writing components targeting various participants.

An implication of this study for teachers and instructors is that L2 writing ability is strongly connected with L2 reading and speaking ability and their development seems to take place hand-in-hand. Pre-writing reading and speaking activities may enhance the effect of L2 writing instruction. An integrated teaching approach, such as content-based instruction (Stoller, 2004), and an integrated communication skills approach (Koda & Yamashita, 2019) will be effective. Another implication is that teachers should reconsider the importance of L2 linguistic knowledge, such as L2 grammar and vocabulary knowledge, and transcription and decoding skills, as the basis of writing proficiency. As our results indicate, their effect on writing performance is stronger than language aptitude, working memory, motivational constructs, and metacognitive knowledge including strategies. Teachers need to make continuous efforts to improve the linguistic knowledge of L2 writers.

Despite these useful implications for researchers and practitioners, the present study is not without limitations. First, the dichotomous categories of language learning contexts (i.e., second language vs. foreign language) and L1–L2 distance (i.e., both Indo-European vs. a combination of Indo-European and non-Indo-European) may have been too crude. The second language group included minority language speakers with primary or secondary education, as well as international students with tertiary education or who were preparing for it, although the majority fell into the latter category. These learners might have differed substantially from one another in terms of their L1 literacy ability and motivational conditions. A future study will need to consider this point. Regarding L1–L2 language distance, even when both L1 and L2 are Indo-European, the processing of different languages can vary greatly and can cause individual variability in L2 writing proficiency and its sub-skills. Differences in rhetorical conventions across L1 and L2 may play a role too. Future research will need to improve the operationalization of L1–L2 distance to suitably address these issues. Second, as noted earlier, equivalent comparisons were at times difficult for some hypothesized moderators. For example, advanced learners were almost exclusively university or postgraduate students, whereas low/intermediate learners were more heterogeneous, ranging from primary school pupils to postgraduate students. Future research should consider the interactional effect of potential moderators (e.g., proficiency  $\times$  school level). To do so, we call for more primary studies with diverse participants and study features. Including objective information of L2 proficiency is also strongly encouraged. Third, the correlations between L2 writing and its components integrated by this study were not corrected for attenuation. Thus, the true mean correlations could have been larger than reported here, and this possibility must be taken into account when comparing the present study results with those of other meta-analyses that used corrected correlations. Fourth, some measurement characteristics were not differentiated. For example, L2 writing performance was mainly evaluated using either holistic scores or the composite scores of analytic ratings, but separate statistical analyses were not conducted for them. In addition, writing tasks also varied from simple tasks such as e-mail writing and picture description to more advanced academic writing tasks. L2 reading and speaking and L1 writing tasks also varied. Different linguistic, cognitive, and metacognitive skills might be needed for different kinds of tasks. Those task effects should be explored further. Grammar knowledge tests also included various test formats (e.g., fill-in-the-blank tests, grammaticality judgement tests, error detection and correction tests). Considering the significant effects of measurement methods as demonstrated by the present study, the effects of test formats should be explored further. It should be noted that very few studies assessed the processing speed and efficiency of writing components such as grammar knowledge, although its importance has been recognized in the literature (Schoonen et al., 2003, 2011).

Therefore, we call for more L2 writing studies incorporating speed and processing measures of writing components in the future.

Despite these limitations, we believe that our study offers useful insights into the relationship between L2 writing proficiency and its components to understand individual differences in L2 writing ability. As a future direction towards a comprehensive model of L2 writing proficiency, meta-analytic structural equation modeling (SEM) (Jak, 2015) of L2 writing components should be pursued. This technique can combine the meta-analytic power of study integration and the modeling power of SEM. This method allows researchers to test the empirical validity of various hypothesized L2 writing models incorporating direct and indirect effects of L2 writing components. In this way, a comprehensive picture of L2 writing proficiency and individual differences will emerge. To do so, primary studies which report correlations not only between L2 writing performance and its various components, but also between each pair of components, need to be considered. This is a goal worth pursuing.

## References

- Alderson, J. C. (1984). Reading in a foreign language: A reading problem or a language problem? In J. C. Alderson & A. H. Urquhart (Eds.), *Reading in a foreign language* (pp. 1–24). Longman.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.
- Baddeley, A. D. (1986). *Working memory*. Oxford University Press.
- Baddeley, A. D. (2007). *Working memory, thought, and action*. Oxford University Press.  
<https://doi.org/10.1093/acprof:oso/9780198528012.001.0001>
- Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Prentice Hall.
- Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Lawrence Erlbaum Associates.
- Bulté, B., & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing*, 26, 42–65.  
<https://doi.org/10.1016/j.jslw.2014.09.005>
- Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1(1), 1–47. <https://doi.org/10.1093/applin/1.1.1>
- Carroll, J. B. (1972). Fundamental considerations in testing for English language proficiency of foreign students. In H. B. Allen & R. N. Campbell (Eds.), *Teaching English as a second language: A book of readings* (pp. 313–320). McGraw-Hill.
- Carroll, J. B. (1981). Twenty-five years of research on foreign language aptitude. In K. C. Diller (Ed.), *Individual differences and universals in language learning aptitude* (pp. 119–154). Newbury House.
- Carroll, J. B., & Sapon, S. M. (1959). *The Modern Language Aptitude Test*. Psychological Corporation.

- Carson, J. E., & Kuehn, P. A. (1992). Evidence of transfer and loss in developing second language writers. *Language Learning*, 42(2), 157–182. <https://doi.org/10.1111/j.1467-1770.1992.tb00706.x>
- Christensen, C. A. (2004). Relationship between orthographic-motor integration and computer use for the production of creative and well-structured text. *British Journal of Educational Psychology*, 74(4), 551–564. <https://doi.org/10.1348/0007099042376373>
- Cohen, A. D., & Brooks-Carson, A. (2001). Research on direct versus translated writing: Students' strategies and their results. *The Modern Language Journal*, 85(2), 169–188. <https://doi.org/10.1111/0026-7902.00103>
- Csapó, B., & Nikolov, M. (2009). The cognitive contribution to the development of proficiency in a foreign language. *Learning and Individual Differences*, 19(2), 209–218. <https://doi.org/10.1016/j.lindif.2009.01.002>
- Cumming, A. (1989). Writing expertise and second-language proficiency. *Language learning*, 39(1), 81–135. <https://doi.org/10.1111/j.1467-1770.1989.tb00592.x>
- Cummins, J., & Swain, M. (2016). *Bilingualism in education: Aspects of theory, research and practice*. Routledge.
- De Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition*, 34(1), 5–34. <https://doi.org/10.1017/S0272263111000489>
- Dörnyei, Z. (2000). Motivation in action: Towards a process-oriented conceptualisation of student motivation. *British Journal of Educational Psychology*, 70(4), 519–538. <https://doi.org/10.1348/000709900158281>
- Dörnyei, Z. (2003). Attitudes, orientations and motivations in language learning: Advances in theory, research and applications. *Language Learning*, 53(S1), 3–32. <https://doi.org/10.1111/1467-9922.53222>
- Farsani, M. A., Beikmohammadi, M., & Mohebbi, A. (2014). Self-regulated learning, goal-oriented learning, and academic writing performance of undergraduate Iranian EFL learners. *TESL-EJ*, 18(2), 1–19.
- Fitzgerald, J., & Shanahan, T. (2000). Reading and writing relations and their development. *Educational Psychologist*, 35(1), 39–50. [https://doi.org/10.1207/S15326985EP3501\\_5](https://doi.org/10.1207/S15326985EP3501_5)
- Flahive, D. E., & Snow, B. G. (1980). Measures of syntactic complexity in evaluating ESL compositions. In J. W. Oiler & K. Perkins (Eds.), *Research and language testing* (pp. 171–176). Newbury House.
- Grabe, W., & Kaplan, R. B. (1996). *Theory and practice of writing: An applied linguistic perspective*. Longman.
- Graham, S. (2006). Writing. In P. Alexander & P. Winne (Eds.), *Handbook of educational psychology* (pp. 457–477). Lawrence Erlbaum Associates.
- Harrison, G. L., Goegan, L. D., Jalbert, R., McManus, K., Sinclair, K., & Spurling, J. (2016). Predictors of spelling and writing skills in first- and second-language learners. *Reading and Writing*, 29(1), 69–89. <https://doi.org/10.1007/s11145-015-9580-1>
- Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. In C. M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences and applications* (pp. 1–27). Lawrence Erlbaum Associates.
- Hayes, J. R. (2012). Modeling and remodeling writing. *Written Communication*, 29(3), 369–388. <https://doi.org/10.1177/0741088312451260>
- Hayes, J. R., & Flower, L. S. (1980). Identifying the organization of writing processes. In L. W. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing* (pp. 3–30). Lawrence Erlbaum Associates.

- Hirose, K. (2003). Comparing L1 and L2 organizational patterns in the argumentative writing of Japanese EFL students. *Journal of Second Language Writing*, 12(2), 181–209. [https://doi.org/10.1016/S1060-3743\(03\)00015-8](https://doi.org/10.1016/S1060-3743(03)00015-8)
- Hirose, K., & Sasaki, M. (1994). Explanatory variables for Japanese students' expository writing in English: An exploratory study. *Journal of Second Language Writing*, 3(3), 203–229. [https://doi.org/10.1016/1060-3743\(94\)90017-5](https://doi.org/10.1016/1060-3743(94)90017-5)
- Hirvela, A., & Belcher, D. (2016). Reading/writing and speaking/writing connections: The advantages of multimodal pedagogy. In R. M. Manchón & P. K. Matsuda (Eds.), *Handbook of second and foreign language writing* (pp. 587–612). Walter de Gruyter. <https://doi.org/10.1515/9781614511335-030>
- Hulstijn, J. H. (2015). *Language proficiency in native and non-native speakers: Theory and research*. John Benjamins. <https://doi.org/10.1075/llt.41>
- Jak, S. (2015). *Meta-analytic structural equation modelling*. Springer. <https://doi.org/10.1007/978-3-319-27174-3>
- Jeon, E.-H., & Yamashita, Y. (2014). L2 reading comprehension and its correlates: A meta-analysis. *Language Learning*, 64(1), 160–212. <https://doi.org/10.1111/lang.12034>
- Jones, D. A., & Christensen, C. A. (1999). Relationship between automaticity in handwriting and students' ability to generate written text. *Journal of Educational Psychology*, 91(1), 1–6. <https://doi.org/10.1037/0022-0663.91.1.44>
- Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grades. *Journal of Educational Psychology*, 80(4), 437–447. <https://doi.org/10.1037/0022-0663.80.4.437>
- Kamimura, T. (1996). Composing in Japanese as a first language and English as a foreign language: A study of narrative writing. *RELC journal*, 27(1), 47–69. <https://doi.org/10.1177/003368829602700103>
- Kellogg, R. T. (1996). A model of working memory in writing. In C. M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences and applications* (pp. 57–71). Lawrence Erlbaum Associates.
- Koda, K. (2005). *Insights into second language reading: A cross-linguistic approach*. Cambridge University Press. <https://doi.org/10.1017/CBO9781139524841>
- Koda, K., & Yamashita, J. (2019). *Reading to learn in a foreign language: An integrated approach to foreign language instruction and assessment*. Routledge.
- Kormos, J. (2012). The role of individual differences in L2 writing. *Journal of Second Language Writing*, 21(4), 390–403. <https://doi.org/10.1016/j.jslw.2012.09.003>
- Kyle, K., & Crossley, S. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing*, 34(1), 12–24. <https://doi.org/10.1016/j.jslw.2016.10.003>
- Laufer, B. (1998). The development of passive and active vocabulary in a second language: same or different? *Applied Linguistics*, 19(2), 255–271. <https://doi.org/10.1093/applin/19.2.255>
- Lee, S. Y. (2005). Facilitating and inhibiting factors in English as a foreign language writing performance: A model testing with structural equation modeling. *Language Learning*, 55(2), 335–374. <https://doi.org/10.1111/j.0023-8333.2005.00306.x>
- Leki, I., Cumming, A., & Silva, T. (2008). *A synthesis of research on second language writing in English*. Routledge.
- Levelt, W. J. M. (1989). *Speaking: Form intention to articulation*. The MIT Press.
- Li, S. (2016). The construct validity of language aptitude: A meta-analysis. *Studies in Second Language Acquisition*, 38(4), 801–842. <https://doi.org/10.1017/S027226311500042X>

- MacArthur, C. A., & Graham, S. (2016). Writing research from a cognitive perspective. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 24–40). The Guilford Press.
- Manchón, R. M. (2001). Trends in the conceptualizations of second language composing strategies: A critical analysis. *International Journal of English Studies*, 1(2), 47–70.
- Manchón, R. M., & Roca de Larios, J. (2007). On the temporal nature of planning in L1 and L2 composing. *Language Learning*, 57(4), 549–593.  
<https://doi.org/10.1111/j.1467-9922.2007.00428.x>
- Masgoret, A. M., & Gardner, R. C. (2003). Attitudes, motivation, and second language learning: A meta-analysis of studies conducted by Gardner and associates. *Language Learning*, 53(1), 123–163. <https://doi.org/10.1111/1467-9922.00212>
- Melby-Lervåg, M., & Lervåg, A. (2011). Cross-linguistic transfer of oral language, decoding, phonological awareness and reading comprehension: A meta-analysis of the correlational evidence. *Journal of Research in Reading*, 34(1), 114–135.  
<https://doi.org/10.1111/j.1467-9817.2010.01477.x>
- Murphy, P. K., & Alexander, P. A. (2000). A motivated exploration of motivation terminology. *Contemporary Educational Psychology*, 25(1), 3–53. <https://doi.org/10.1006/ceps.1999.1019>
- Nation, P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press.  
<https://doi.org/10.1017/CBO9781139858656>
- Nicolas-Conesa, F., Roca de Larios, J. R., & Coyle, Y. (2014). Development of EFL students' mental models of writing and their effects on performance. *Journal of Second Language Writing*, 24, 1–19. <https://doi.org/10.1016/j.jslw.2014.02.004>
- Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, 50(3), 417–528.  
<https://doi.org/10.1111/0023-8333.00136>
- Oh, E., Lee, C. M., & Moon, Y. I. (2015). The contributions of planning, L2 linguistic knowledge and individual differences to L2 writing. *The Journal of Asia TEFL* 12(2), 45–85.
- Olive, T. (2004). Working memory in writing: Empirical evidence from the dual-task technique. *European Psychologist*, 9(1), 32–42. <https://doi.org/10.1027/1016-9040.9.1.32>
- Olive, T. (2012). Working memory in writing. In V. W. Berninger (Ed.), *Past, present, and future contributions of cognitive writing research to cognitive psychology* (pp. 485–503). Psychology Press.
- Pennington, M. C., & So, S. (1993). Comparing writing process and product across two languages: A study of 6 Singaporean university student writers. *Journal of Second Language Writing*, 2(1), 41–63. [https://doi.org/10.1016/1060-3743\(93\)90005-N](https://doi.org/10.1016/1060-3743(93)90005-N)
- Pimsleur, P. (1966). *Pimsleur Language Aptitude Battery*. Harcourt Brace Jovanovich.
- Plonsky, L., & Oswald, F. L. (2014). How big is “big”? Interpreting effect sizes in L2 research. *Language Learning*, 64(4), 878–912. <https://doi.org/10.1111/lang.12079>
- Purpura, J. E. (2004). *Assessing grammar*. Cambridge University Press.  
<https://doi.org/10.1017/CBO9780511733086>
- Read, J. (2000). *Assessing vocabulary*. Cambridge University Press.  
<https://doi.org/10.1017/CBO9780511732942>
- Roca de Larios, J., Manchón, R., Murphy, L., & Marín, J. (2008). The foreign language writer's strategic behaviour in the allocation of time to writing processes. *Journal of Second Language Writing*, 17(1), 30–47. <https://doi.org/10.1016/j.jslw.2007.08.005>
- Roca de Larios, J., Marín, J., & Murphy, L. (2001). A temporal analysis of formulation processes in L1 and L2 writing. *Language Learning*, 51(3), 497–538.  
<https://doi.org/10.1111/0023-8333.00163>



- Roca de Larios, J., Nicolas-Conesa, F., & Coyle, Y. (2016). Focus on writers: Processes and strategies. In R. M. Manchón & P. K. Matsuda (Eds.), *Handbook of second and foreign language writing* (pp. 267–286). Walter de Gruyter. <https://doi.org/10.1515/9781614511335-015>
- Rosenthal, R. (1979). The “file drawer problem” and tolerance for null results. *Psychological Bulletin*, 86(3), 638–641. <https://doi.org/10.1037/0033-2909.86.3.638>
- Sarkhoush, H. (2013). Relationship among Iranian EFL learners’ self-efficacy in writing, attitude towards writing, writing apprehension and writing performance. *Journal of Language Teaching and Research*, 4(5), 1126–1132. <https://doi.org/10.4304/jltr.4.5.1126-1132>
- Sasaki, M. (2009). Changes in English as a foreign language students’ writing over 3.5 years: A sociocognitive account. In R. M. Manchón (Ed.), *Writing in foreign language contexts: Learning, teaching, and research* (pp. 49–76). Multilingual Matters. <https://doi.org/10.21832/9781847691859-006>
- Sasaki, M., & Hirose, K. (1996). Explanatory variables for EFL students’ expository writing. *Language learning*, 46(1), 137–168. <https://doi.org/10.1111/j.1467-1770.1996.tb00643.x>
- Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Palgrave MacMillan. <https://doi.org/10.1057/9780230293977>
- Schoonen, R. (2019). Are reading and writing building on the same skills? The relationship between reading and writing in L1 and EFL. *Reading and Writing*, 32(3), 511–535. <https://doi.org/10.1007/s11145-018-9874-1>
- Schoonen, R., Snellings, P., Stevenson, M., & Van Gelderen, A. (2009). Towards a blueprint of the foreign language writer: The linguistic and cognitive demands of foreign language writing. In R. M. Manchón (Ed.), *Writing in foreign language contexts: Learning, teaching, and research* (pp. 77–101). Multilingual Matters. <https://doi.org/10.21832/9781847691859-007>
- Schoonen, R., Van Gelderen, A., De Glopper, K., Hulstijn, J., Simis, A., Snellings, P., & Stevenson, M. (2003). First language and second language writing: The role of linguistic knowledge, speed of processing, and metacognitive knowledge. *Language Learning*, 53(1), 165–202. <https://doi.org/10.1111/1467-9922.00213>
- Schoonen, R., Van Gelderen, A., Stoel, R. D., Hulstijn, J., & De Glopper, K. (2011). Modeling the development of L1 and EFL writing proficiency of secondary school students. *Language learning*, 61(1), 31–79. <https://doi.org/10.1111/j.1467-9922.2010.00590.x>
- Service, E. (1992). Phonology, working memory, and foreign-language learning. *The Quarterly Journal of Experimental Psychology Section A*, 45(1), 21–50. <https://doi.org/10.1080/14640749208401314>
- Service, E., & Kohonen, V. (1995). Is the relation between phonological memory and foreign language learning accounted for by vocabulary acquisition? *Applied Psycholinguistics*, 16(2), 155–172. <https://doi.org/10.1017/S0142716400007062>
- Shanahan, T. (2016). Relationships between reading and writing development. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (2nd ed., pp. 194–207). The Guilford Press.
- Sparks, R. L., Humbach, N., Patton, J., & Ganschow, L. (2011). Subcomponents of second-language aptitude and second-language proficiency. *The Modern Language Journal*, 95(2), 253–273. <https://doi.org/10.1111/j.1540-4781.2011.01176.x>
- Sparks, R. L., Patton, J., Ganschow, L., & Humbach, N. (2009). Long-term relationships among early first language skills, second language aptitude, second language affect, and later second language proficiency. *Applied Psycholinguistics*, 30(4), 725–755. <https://doi.org/10.1017/S0142716409990099>

- Sparks, R. L., Patton, J., Ganschow, L., & Humbach, N. (2012). Do L1 reading achievement and L1 print exposure contribute to the prediction of L2 proficiency? *Language Learning*, 62(2), 473–505. <https://doi.org/10.1111/j.1467-9922.2012.00694.x>
- Speciale, G., Ellis, N. C., & Bywater, T. (2004). Phonological sequence learning and short-term store capacity determine second language vocabulary acquisition. *Applied Psycholinguistics*, 25(2), 293–321. <https://doi.org/10.1017/S0142716404001146>
- Stevenson, M., Schoonen, R., & De Glopper, K. (2006). Revising in two languages: A multi-dimensional comparison of online writing revisions in L1 and FL. *Journal of Second Language Writing*, 15(3), 201–233. <https://doi.org/10.1016/j.jslw.2006.06.002>
- Stoller, F. L. (2004). Content-based instruction: Perspectives on curriculum planning. *Annual Review of Applied Linguistics*, 24, 261–283. <https://doi.org/10.1017/S0267190504000108>
- Ushioda, E., & Dörnyei, Z. (2012). Motivation. In S. M. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 396–409). Routledge.
- Vanderberg, R., & Swanson, H. L. (2007). Which components of working memory are important in the writing process? *Reading and Writing*, 20(7), 721–752. <https://doi.org/10.1007/s11145-006-9046-6>
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3), 1–48. <https://doi.org/10.18637/jss.v036.i03>
- Weigle, S. C. (2002). *Assessing Writing*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511732997>
- Wesche, M., & Paribakht, T. S. (1996). Assessing L2 vocabulary knowledge: Depth versus breadth. *Canadian Modern Language Review*, 53(1), 13–40. <https://doi.org/10.3138/cmlr.53.1.13>
- Whalen, K., & Ménard, N. (1995). L1 and L2 writers' strategic and linguistic knowledge: A model of multiple-level discourse processing. *Language Learning*, 45(3), 381–418. <https://doi.org/10.1111/j.1467-1770.1995.tb00447.x>
- Zimmerman, B. J., & Risemberg, R. (1997). Becoming a self-regulated writer: A social cognitive perspective. *Contemporary Educational Psychology*, 22(1), 73–101. <https://doi.org/10.1006/ceps.1997.0919>

## Appendix A. 103 studies included in the meta-analysis

- Abu-Akel, A. (1997). On reading-writing relationships in first and foreign languages. *JALT Journal*, 19(2), 199–216.
- Abu-Rabia, S. (2003). The influence of working memory on reading and creative writing processes in a second language. *Educational Psychology*, 23(2), 209–222. <https://doi.org/10.1080/0144341030303227>
- Abu-Rabia, S. (2004). Teachers' role, learners' gender differences, and FL anxiety among seventh-grade students studying English as a FL. *Educational Psychology*, 24(5), 711–721. <https://doi.org/10.1080/0144341042000263006>
- Alderson, J. C., Clapham, C., & Steel, D. (1997). Metalinguistic knowledge, language aptitude and language proficiency. *Language Teaching Research*, 1(2), 93–121. <https://doi.org/10.1177/136216889700100202>
- Babayigit, S. (2014). Contributions of word-level and verbal skills to written expression: Comparison of learners who speak English as a first (L1) and second language (L2). *Reading and Writing*, 27(7), 1207–1229. <https://doi.org/10.1007/s11145-013-9482-z>



- Canale, M., Frennette, N., & Belanger, M. (1987). Evaluation of minority students writing in first and second languages. In J. Fine (Ed.), *Second language discourse: A textbook of current research* (pp. 147–165). Ablex.
- Carlson, S., Bridgeman, B., Camp, R., & Wanders, J. (1985). *Relationship of admission test scores to writing performance of native and non-native speakers of English* (ETS Research Report 85–21). Educational Testing Service. Retrieved from <https://www.ets.org/Media/Research/pdf/RR-85-21-Carlson.pdf>
- Carrell, P. L., & Connor, U. (1991). Reading and writing descriptive and persuasive texts. *The Modern Language Journal*, 75(3), 314–324. <https://doi.org/10.1111/j.1540-4781.1991.tb05361.x>
- Carson, J. E., Carrell, P. L., Silberstein, S., Kroll, B., & Kuehn, P. A. (1990). Reading-writing relationships in first and second language. *TESOL Quarterly*, 24(2), 245–266. <https://doi.org/10.2307/3586901>
- Carson, J. E., & Kuehn, P. A. (1992). Evidence of transfer and loss in developing second language writers. *Language Learning*, 42(2), 157–182. <https://doi.org/10.1111/j.1467-1770.1992.tb00706.x>
- Chao, Y.-C. J. (2003). Contrastive rhetoric, lexico-grammatical knowledge, writing expertise, and metacognitive knowledge: An integrated account of the development of English writing by Taiwanese students (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (3119448)
- Chen, D.-W. (1999). *The connections between L1 and L2 writing performances: From the perspective of writing expertise*. Retrieved from <https://eric-ed-gov/?id=ED464489>
- Chen, M. C., & Lin, H. J. (2009). Self-efficacy, foreign language anxiety as predictors of academic performance among professional program students in a general English proficiency writing test. *Perceptual and Motor Skills*, 109(2), 420–430. <https://doi.org/10.2466/pms.109.2.420-430>
- Cohen, A. D., & Brooks-Carson, A. (2001). Research on direct versus translated writing: Students' strategies and their results. *The Modern Language Journal*, 85(2), 169–188. <https://doi.org/10.1111/0026-7902.00103>
- Cook, M. L. (1988). The validity of the contrastive rhetoric hypothesis as it relates to Spanish-speaking advanced ESL students (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (8826125)
- Cox, T., & Davies, R. S. (2012). Using automatic speech recognition technology with elicited oral response testing. *CALICO Journal*, 29(4), 601–618. <https://doi.org/10.11139/cj.29.4.601-618>
- Csapó, B., & Nikolov, M. (2009). The cognitive contribution to the development of proficiency in a foreign language. *Learning and Individual Differences*, 19(2), 209–218. <https://doi.org/10.1016/j.lindif.2009.01.002>
- DeMauro, G. (1992). Examination of the relationships among TSE, TWE and TOEFL scores. *Language Testing*, 9(2), 149–161. <https://doi.org/10.1177/026553229200900203>
- Elder, C. (2009). Validating a test of metalinguistic knowledge. In R. Ellis, S. Loewen, C. Elder, R. Erlam, J. Philp, & H. Reinders (Eds.), *Implicit and explicit knowledge in second language learning, testing and teaching* (pp. 113–138). Multilingual Matters. <https://doi.org/10.21832/9781847691767-007>
- Elder, C., & Ellis, R. (2009). Implicit and explicit knowledge of an L2 and language proficiency. In R. Ellis, S. Loewen, C. Elder, R. Erlam, J. Philp, & H. Reinders (Eds.), *Implicit and explicit knowledge in second language learning, testing and teaching* (pp. 167–193). Multilingual Matters. <https://doi.org/10.21832/9781847691767-009>
- Erkan, D. Y., & Saban, A. (2011). Writing performance relative to writing apprehension, self-efficacy in writing, and attitudes towards writing: A correlational study in Turkish tertiary-level EFL. *The Asian EFL Journal Quarterly*, 13(1), 164–192.

- Evans, M., & Fisher, L. (2005). Measuring gains in pupils' foreign language competence as a result of participation in a school exchange visit: The case of Y9 pupils at three comprehensive schools in the UK. *Language Teaching Research*, 9(2), 173–192.  
<https://doi.org/10.1191/1362168805lr1620a>
- Farsani, M. A., Beikmohammadi, M., & Mohebbi, A. (2014). Self-regulated learning, goal-oriented Learning, and academic writing performance of undergraduate Iranian EFL learners. *TESL-EJ*, 18(2), 1–19.
- Fatemi, A. H., & Vahidnia, F. (2013). An investigation into Iranian EFL learners' level of writing self-efficacy. *Theory and Practice in Language Studies*, 3(9), 1698–1704.  
<https://doi.org/10.4304/tpls.3.9.1698-1704>
- Flahive, D., & Bailey, N. (1993). Exploring reading/writing relationships in adult second language learners. In J. Carson & I. Leki (Eds.), *Reading in the composition class: Second language perspectives* (pp. 128–140). Heinle and Heinle
- Ghaderi, M., & Nikou, F. R. (2016). The relationship between Krashen's affective filter hypothesis and Iranian EFL learners' writing skill. *Modern Journal of Language Teaching Methods*, 6(1), 264–276.
- Gustilo, L. E. (2013). An analysis of writer's performance, resources, and idea generation processes: The case of Filipino engineering students. *Language Testing in Asia*, 3(1), 1–14.  
<https://doi.org/10.1186/2229-0443-3-2>
- Gutiérrez, X. (2012). Implicit knowledge, explicit knowledge, and achievement in second language (L2) Spanish. *Canadian Journal of Applied Linguistics*, 15, 20–41.
- Gutiérrez, X. (2016). Analyzed knowledge, metalanguage, and second language proficiency. *System*, 60, 42–54. <https://doi.org/10.1016/j.system.2016.06.003>
- Harris, S. N. (2014). Multicompetence in foreign language writing: Function, effects, and beliefs (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (3620879)
- Harrison, G. L., Goegan, L. D., Jalbert, R., McManus, K., Sinclair, K., & Spurling, J. (2016). Predictors of spelling and writing skills in first- and second-language learners. *Reading and Writing*, 29(1), 69–89. <https://doi.org/10.1007/s11145-015-9580-1>
- Hashemnejad, F., Zoghi, M., & Amini, D. (2014). The relationship between self-efficacy and writing performance across genders. *Theory and Practice in Language Studies*, 4(5), 1045–1052.  
<https://doi.org/10.4304/tpls.4.5.1045-1052>
- Hassan, B. A. (2001). *The relationship of writing apprehension and self-esteem to the writing quality and quantity of EFL university students*. Retrieved from <https://eric-ed-gov/?id=ED459671>
- Heththong, R., & Teo, A. (2013). Does writing self-efficacy correlate with and predict writing performance? *International Journal of Applied Linguistics & English Literature*, 2(1), 157–167.  
<https://doi.org/10.7575/ijalel.v.2n.1p.157>
- Hirose, K. (2003). Comparing L1 and L2 organizational patterns in the argumentative writing of Japanese EFL students. *Journal of Second Language Writing*, 12, 181–209.  
[https://doi.org/10.1016/S1060-3743\(03\)00015-8](https://doi.org/10.1016/S1060-3743(03)00015-8)
- Hirose, K., & Sasaki, M. (1994). Explanatory variables for Japanese students' expository writing in English: An exploratory study. *Journal of Second Language Writing*, 3(3), 203–229.  
[https://doi.org/10.1016/1060-3743\(94\)90017-5](https://doi.org/10.1016/1060-3743(94)90017-5)
- Hubert, M. D. (2008). The relationship between writing and speaking in the U.S. university Spanish language classroom (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (3330267)

- In'nami, Y., Koizumi, R., & Nakamura, K. (2016). Factor structure of the Test of English for Academic Purposes (TEAP) test in relation to the TOEFL iBT test. *Language Testing in Asia*, 6(1), 1–23. <https://doi.org/10.1186/s40468-016-0025-9>
- Ito, F. (2004). The interrelationship among first language writing skills, second language writing skills, and second language proficiency of EFL university students. *JACET Bulletin*, 39, 43–58.
- Ito, F. (2011). L2 reading–writing correlation in Japanese EFL high school students. *The Language Teacher*, 35(5), 23–29. <https://doi.org/10.37546/JALTTLT35.5-2>
- Jang, Y., & Lee, J. (2019). The effects of ideal and ought-to L2 selves on Korean EFL learners' writing strategy use and writing quality. *Reading and Writing*, 32(5), 1129–1148. <https://doi.org/10.1007/s11145-018-9903-0>
- Kamimura, T. (1996). Composing in Japanese as a first language and English as a foreign language: A study of narrative writing. *RELC Journal*, 27(1), 47–69. <https://doi.org/10.1177/003368829602700103>
- Kamiya, N. (2017). Can the National Center Test in Japan be replaced by commercially available private English tests of four skills? In the case of TOEFL Junior Comprehensive. *Language Testing in Asia*, 7(1), 1–22. <https://doi.org/10.1186/s40468-017-0046-z>
- Kang, H. S. (2011). The relationship between different dimensions of lexical proficiency and writing quality of Korean EFL learners. *응용언어학*, 27(3), 81–104.
- Karakoç, D., & Köse, G. D. (2017). The impact of vocabulary knowledge on reading, writing and proficiency scores of EFL learners. *Dil ve Dilbilimi Çalışmaları Dergisi*, 13(1), 352–378.
- Kashef, S. (2016). Self-efficacy and Iranian female EFL learners' writing performance: An investigation into their relationship. *Modern Journal of Language Teaching Methods*, 6(6), 233–240.
- Kobayashi, H., & Rinnert, C. (1992). Effects of 1st language on 2nd language writing: Translation versus direct composition. *Language Learning*, 42(2), 183–215. <https://doi.org/10.1111/j.1467-1770.1992.tb00707.x>
- Kormos, J., & Sáfár, A. (2008). Phonological short-term memory, working memory and foreign language performance in intensive language learning. *Bilingualism*, 11(2), 261–271. <https://doi.org/10.1017/S1366728908003416>
- Lee, H. K. (2012). Exploring the relationship among L1 writing, L2 writing, and L2 linguistic proficiency depending on L2 topic difficulty. *Asia-Pacific Education Researcher*, 21(3), 576–586.
- Lee, S. Y. (2005). Facilitating and inhibiting factors in English as a foreign language writing performance: A model testing with structural equation modeling. *Language Learning*, 55(2), 335–374. <https://doi.org/10.1111/j.0023-8333.2005.00306.x>
- Leong, C. K., Shum, M. S. K., Tai, C. P., Ki, W. W., & Zhang, D. (2019). Differential contribution of psycholinguistic and cognitive skills to written composition in Chinese as a second language. *Reading and Writing*, 32(2), 439–466. <https://doi.org/10.1007/s11145-018-9873-2>
- Leontjev, D., Huhta, A., & Mantyla, K. (2016). Word derivational knowledge and writing proficiency: How do they link? *System*, 59, 73–89. <https://doi.org/10.1016/j.system.2016.03.013>
- Li, Z. (2015). An argument-based validation study of the English Placement Test (EPT): Focusing on the inferences of extrapolation and ramification (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (3728798)
- Lin, M.-C., Cheng, Y.-S., Lin, S.-H., & Hsieh, P.-J. (2015). The role of research-article writing motivation and self-regulatory strategies in explaining research-article abstract writing ability. *Perceptual and Motor Skills*, 120(2), 397–415. <https://doi.org/10.2466/50.PMS.120v17x9>

- Llach, M. P. A. (2010). Examining the role of L2 proficiency in L2 reading-writing relationships. *Estudios Ingleses de la Universidad Complutense*, 18, 35–52. Retrieved from [https://www.researchgate.net/publication/277268940\\_Examining\\_the\\_role\\_of\\_L2\\_proficiency\\_in\\_the\\_relationship\\_between\\_L2\\_reading\\_and\\_writing](https://www.researchgate.net/publication/277268940_Examining_the_role_of_L2_proficiency_in_the_relationship_between_L2_reading_and_writing)
- Llach, M. P. A., & Gallego, M. T. (2009). Examining the relationship between receptive vocabulary size and written skills of primary school learners. *Journal of the Spanish Association of Anglo-American Studies*, 31(1), 129–147.
- Lopez, E. M. (2004). Re-education for the reading and writing processes: An exploration of the teaching and learning of undergraduate reading and writing in Colombia (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (U185458)
- MacIntyre, P. D., Noels, K. A., & Clément, R. (1997). Biases in self-ratings of second language proficiency: The role of language anxiety. *Language learning*, 47(2), 265–287. <https://doi.org/10.1111/0023-8333.81997008>
- Mallahi, O., Amirian, S. M. R., Zareian, G. R., & Adel, S. M. R. (2016). An investigation into the individual differences correlates of Iranian undergraduate EFL learners' writing competence: A mixed methods approach. *Iranian Journal of Applied Linguistics*, 19(1), 99–140. <https://doi.org/10.18869/acadpub.ijal.19.1.99>
- Mantyla, K., & Huhta, A. (2014). Knowledge of word parts. In J. Milton & T. Fitzpatrick (Eds.), *Dimensions of vocabulary knowledge* (pp. 45–59). Palgrave Macmillan. [https://doi.org/10.1007/978-1-137-36831-7\\_4](https://doi.org/10.1007/978-1-137-36831-7_4)
- Manyike, T. V. (2007). The acquisition of English academic language proficiency among grade 7 learners in South African schools (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (0821155)
- Miralpeix, I., & Muñoz, C. (2018). Receptive vocabulary size and its relationship to EFL language skills. *International Review of Applied Linguistics in Language Teaching*, 56(1), 1–24. <https://doi.org/10.1515/iral-2017-0016>
- Moeller, A. J., Theiler, J. M., & Wu, C. (2012). Goal setting and student achievement: A longitudinal study. *The Modern Language Journal*, 96(2), 153–169. <https://doi.org/10.1111/j.1540-4781.2011.01231.x>
- Ndlovu, K. E. D. (2010). Story-writing development from grades 4 to 6: Do language status and reading profile matter? (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (NR73167)
- Neugebauer, S. R., & Howard, E. R. (2015). Exploring associations among writing self-perceptions, writing abilities, and native language of English-Spanish two-way immersion students. *Bilingual Research Journal*, 38(3), 313–335. <https://doi.org/10.1080/15235882.2015.1093039>
- Nicolás-Conesa, F., de Larios, J. R., & Coyle, Y. (2014). Development of EFL students' mental models of writing and their effects on performance. *Journal of Second Language Writing*, 24, 1–19. <https://doi.org/10.1016/j.jslw.2014.02.004>
- Oh, E., Lee, C. M., & Moon, Y. I. (2015). The contributions of planning, L2 linguistic knowledge and individual differences to L2 writing. *The Journal of Asia TEFL* 12(2), 45–85.
- Pae, H. K., & Greenberg, D. (2014). The relationship between receptive and expressive subskills of Academic L2 proficiency in nonnative speakers of English: A multigroup approach. *Reading Psychology*, 35(3), 221–259. <https://doi.org/10.1080/02702711.2012.684425>
- Pae, H. K., & O'Brien, B. (2018). Overlap and uniqueness: Linguistic componential traits contributing to expressive skills in English as a foreign language. *Reading Psychology*, 39(4), 384–412. <https://doi.org/10.1080/02702711.2018.1443298>

- Pimsarn, P. (1986). The reading and writing relationship: A correlational study of English as a second language learners at the collegiate level (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (8626048)
- Rifkin, B. (2005). A ceiling effect in traditional classroom foreign language instruction: Data from Russian. *The Modern Language Journal*, 89(1), 3–18. <https://doi.org/10.1111/j.0026-7902.2005.00262.x>
- Roche, T., & Harrington, M. (2013). Recognition vocabulary knowledge as a predictor of academic performance in an English as a foreign language setting. *Language Testing in Asia*, 3(1), 1–13. <https://doi.org/10.1186/2229-0443-3-12>
- Saadat, M., & Dastgerdi, M. F. (2014). Correlates of L2 writing ability of Iranian students majoring in English. *Procedia – Social and Behavioral Sciences*, 98, 1572–1579. <https://doi.org/10.1016/j.sbspro.2014.03.580>
- Sáfar, A., & Kormos, J. (2008). Revisiting problems with foreign language aptitude. *International Review of Applied Linguistics in Language Teaching*, 46, 113–136. <https://doi.org/10.1515/IRAL.2008.005>
- Salmani-Nodoushan, M. A. (2015). Anxiety as it pertains to EFL writing ability and performance. *i-Manager's Journal on Educational Psychology*, 9(1), 1–12. <https://doi.org/10.26634/jpsy.9.1.3521>
- Sarkhoush, H. (2013). Relationship among Iranian EFL learners' self-efficacy in writing, attitude towards writing, writing apprehension and writing performance. *Journal of Language Teaching and Research*, 4(5), 1126–1132. <https://doi.org/10.4304/jltr.4.5.1126-1132>
- Sasaki, M., & Hirose, K. (1996). Explanatory variables for EFL students' expository writing. *Language Learning*, 46(1), 137–174. <https://doi.org/10.1111/j.1467-1770.1996.tb00643.x>
- Schoonen, R. (2019). Are reading and writing building on the same skills? The relationship between reading and writing in L1 and EFL. *Reading and Writing*, 32(3), 511–535. <https://doi.org/10.1007/s11145-018-9874-1>
- Schoonen, R., Van Gelderen, A., Stoel, R. D., Hulstijn, J., & De Glopper, K. (2011). Modeling the development of L1 and EFL writing proficiency of secondary school students. *Language Learning*, 61(1), 31–79. <https://doi.org/10.1111/j.1467-9922.2010.00590.x>
- Service, E. (1992). Phonology, working memory, and foreign-language learning. *The Quarterly Journal of Experimental Psychology Section A*, 45(1), 21–50. <https://doi.org/10.1080/14640749208401314>
- Service, E., & Kohonen, V. (1995). Is the relation between phonological memory and foreign language learning accounted for by vocabulary acquisition? *Applied Psycholinguistics*, 16(2), 155–172. <https://doi.org/10.1017/S0142716400007062>
- Shah, P. M., Mahmud, W. H., Din, R., Yusof, A., & Pardi, K. M. (2011). Self-efficacy in the writing of Malaysian ESL learners. *World Applied Sciences Journal (Innovation and Pedagogy for Lifelong Learning)*, 15, 8–11 Retrieved from [http://www.idosi.org/wasj/wasj15\(IPLL\)11/2.pdf](http://www.idosi.org/wasj/wasj15(IPLL)11/2.pdf)
- Shi, L., & Qian, D. (2012). How does vocabulary knowledge affect Chinese EFL learners' writing quality in web-based settings? Evaluating the relationships among three dimensions of vocabulary knowledge and writing quality. *Chinese Journal of Applied Linguistics*, 35(1), 117–127. <https://doi.org/10.1515/cjal-2012-0009>
- Shin, S.-Y., & Ewert, D. (2015). What accounts for integrated reading-to-write task scores? *Language Testing*, 32(2), 259–281. <https://doi.org/10.1177/0265532214560257>
- Skibniewski, L., & Skibniewska, M. (1986). Experimental study: The writing processes of intermediate/advanced foreign language learners in their foreign and native languages. *Studia Anglica Posnaniensia*, 19, 142–163.

- Sparks, R. L., Humbach, N., Patton, J., & Ganschow, L. (2011). Subcomponents of second-language aptitude and second-language proficiency. *The Modern Language Journal*, 95(2), 253–273. <https://doi.org/10.1111/j.1540-4781.2011.01176.x>
- Sparks, R. L., Patton, J., Ganschow, L., & Humbach, N. (2009). Long-term relationships among early first language skills, second language aptitude, second language affect, and later second language proficiency. *Applied Psycholinguistics*, 30(4), 725–755. <https://doi.org/10.1017/S0142716409990099>
- Sparks, R. L., Patton, J., Ganschow, L., & Humbach, N. (2012). Do L1 reading achievement and L1 print exposure contribute to the prediction of L2 proficiency? *Language Learning*, 62(2), 473–505. <https://doi.org/10.1111/j.1467-9922.2012.00694.x>
- Speciale, G., Ellis, N. C., & Bywater, T. (2004). Phonological sequence learning and short-term store capacity determine second language vocabulary acquisition. *Applied Psycholinguistics*, 25(2), 293–321. <https://doi.org/10.1017/S0142716404001146>
- Tanyer, S. (2015). The role of writing and reading self-efficacy in first-year preservice EFL teachers' writing performance. *Proceedings of the 1st GlobELT Conference on Teaching and Learning English as an Additional Language*, 199, 38–43. <https://doi.org/10.1016/j.sbspro.2015.07.484>
- Tracy, G. E. (1989). The effects of sentence-combining practice on syntactic maturity and writing quality in ESL students in freshman composition (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (9004042)
- Varnaseri, M., & Farvardin, M. T. (2016). The relationship between depth and breadth of vocabulary knowledge and Iranian EFL learners' listening comprehension. *Modern Journal of Language Teaching Methods*, 6(2), 544–554. <https://doi.org/10.5861/ijrsl.2016.1438>
- Wagnild, J. V. (1988). The relationship between years of high school French study and performance on proficiency tests in listening, reading, speaking, and writing for entering freshmen at the University of Minnesota (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (8911033)
- Wakabayashi, T. (2002). Bilingualism as a future investment: The case of Japanese high school students at an international school in Japan. *Bilingual Research Journal*, 26(3), 631–658. <https://doi.org/10.1080/15235882.2002.10162582>
- Winke, P. M. (2005). Individual differences in adult Chinese second language acquisition: The relationships among aptitude, memory and strategies for learning (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (3175844)
- Wistner, B. (2014). Effects of metalinguistic knowledge and language aptitude on second language learning (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (3611192)
- Wu, Y. (1992). First and second language writing relationship: Chinese and English (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (9300521)
- Xinhua, Z. (2008). Is syntactic maturity a reliable measurement to investigate the relationship between English speaking and writing? *The Asian EFL Journal Quarterly*, 10(1), 133–153.
- Yang, Y. (2017). An empirical study on correlation between second language writing anxiety and CET-4 writing score. *Proceedings of the 2017 3rd International Conference on Social Science and Higher Education*, 99, 387–390. <https://doi.org/10.2991/icsshe-17.2017.97>
- Yun, Y. (2005). Factors explaining EFL learners' performance in a timed essay writing test: A structural equation modeling approach (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (3199191)



- Zhang, J., McBride-Chang, C., Wagner, R. K., & Chan, S. (2014). Uniqueness and overlap: Characteristics and longitudinal correlates of native Chinese children's writing in English as a foreign language. *Bilingualism: Language and Cognition*, 17(2), 347–363.  
<https://doi.org/10.1017/S1366728913000163>
- Zhang, X. (1993). English collocations and their effect on the writing of native and non-native college freshmen (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (9319454)
- Zhang, Y., & Guo, H. (2012). A study of English writing and domain-specific motivation and self-efficacy of Chinese EFL learners. *Journal of Pan-Pacific Association of Applied Linguistics*, 16(2), 101–121.

## **Appendix B. Acceptable measures of included writing correlates and writing performance**

### *L2 writing performance*

In order to be deemed as a measure of L2 writing performance, the participants had to engage in an L2 writing task as an act of communication, whether writing a narrative or academic essay, describing a picture, responding to a film, or writing a letter or email. Integrated tasks, such as reading-listening-writing tasks and cooperative writing tasks were excluded, because we focused on individual differences of L2 writing and its component skills. As for an index of L2 writing performance, holistic scores of overall writing quality, composite scores of analytic (multi-trait) rating dimensions, or writing scores of standardized proficiency tests for non-native speakers were accepted (e.g., Pearson Test of English Academic). Those holistic or analytic ratings had to be conducted by human raters. Composite scores of analytic ratings had to include both linguistic and rhetorical dimensions of writing as components.

### *L2 grammar knowledge*

An acceptable grammar test had to assess some form of L2 morphosyntactic or syntactic knowledge. An acceptable test format included fill-in-the-blank tests with or without multiple choice, timed or untimed grammaticality judgement tests, sentence combining tests, error detection (and correction) tests, and meta-language tests (e.g., tests eliciting explanations of grammar rules from participants).

### *L2 vocabulary knowledge*

To be accepted as a measure of L2 vocabulary knowledge, the test had to assess at least one dimension of vocabulary knowledge, including the breadth (size), depth, receptive, or productive knowledge of vocabulary.

### *L2 reading comprehension*

To be considered as a measure of L2 reading comprehension, the measure had to involve reading an L2 passage and answering comprehension questions either in their L1 or L2. Reading scores of standardized proficiency tests for L2 users were also used as an index of L2 reading comprehension.

*L2 speaking performance*

To be deemed as a measure of L2 speaking performance, participants had to engage in a communicative speaking task, such as an interview, retelling a story, or describing a picture. Speaking scores of standardized L2 proficiency tests were also accepted.

*L1 writing performance*

As an acceptable L1 writing measure, participants had to engage in an L1 writing task, which was comparable to the L2 writing task that they also worked on. These tasks included narrative or academic essay writing, describing a picture, and letter or email writing. Holistic scores of overall L1 writing quality, or composite scores of analytic rating dimensions were accepted as an index of L1 writing performance.

*Motivational constructs*

As motivation is a complex concept, a wide range of operational definitions such as intrinsic and integrative motivation; attitudes about L2 community, L2 learning, and/or L2 writing; self-efficacy in L2 learning, in L2 writing, and/or more generally; goal orientation on L2 learning and/or L2 writing; and apprehension or anxiety on L2 learning, L2 writing, and/or L2 classrooms were deemed relevant in the present study. These traits had to be quantitatively scored using questionnaires.

*L2 transcription*

To be considered as a measure of L2 transcription, the measure had to assess handwriting fluency (speed and accuracy), or knowledge of L2 spelling rules at a word or sentence level. These assessments had to be conducted independently from L2 writing performance itself.

*L2 decoding*

An acceptable L2 decoding measure had to assess oral L2 reading fluency, speed, and accuracy of pseudo word or real word reading, or L2 letter naming.

*Language aptitude*

Acceptable measures of language aptitude had to be the composite scores of widely used foreign language aptitude tests, such as the MLAT and PLAB, including their versions translated into other languages.

*Working memory*

As for working memory measures, traditional working memory tests such as word or non-word repetition tests, digit span tests, and listening and reading span tests were accepted in the present study.

*Metacognitive knowledge*

An acceptable metacognitive knowledge test or questionnaire had to operationally assess an L2 writer's perceived and actual use of writing strategies, self-regulative strategies, self-assessment of L2 proficiency or L2 writing, or knowledge of L2 written texts and/or their characteristics. An acceptable test or questionnaire had to be conducted independently from the L2 writing task itself.





## L2 listening comprehension

### Theory and research

Elvis Wagner

Temple University

This chapter examines how L2 listening ability has been modeled and operationalized in the research literature, and provides a critical overview of the dominant models. It also describes how researchers have used both taxonomies of listening skills as well as data-driven approaches to creating models of listening ability. The chapter then provides a critical discussion of how the constructs of L2 listening ability have been operationalized and measured by empirical researchers. The chapter concludes with an analysis of why models and operationalizations of L2 listening ability have often neglected to include or focus on those aspects of language that are unique to listening ability.

#### 1. **Background: A conceptual introduction to the key constructs and models of L2 listening**

Second language (L2) listening ability has been the focus of a robust amount of research in the last thirty years. This chapter will examine and critique how the key constructs of L2 listening ability have been modeled and operationalized in the research literature.

When trying to formulate a model of L2 listening ability, most researchers have based their model at least in part on models of L1 listening ability. This certainly makes sense, as the processes are the same physiologically, and are very similar in many other ways as well. Therefore, before describing the different models of L2 listening ability, it is necessary to briefly describe some models of L1 listening ability that have been influential for L2 models.

Anderson's cognitive information processing model (1985, 1990, 1995) has been hugely influential for L2 listening researchers, and continues to be the cognitive model used as the basis for most models of L2 listening ability. The model has evolved over time, but his 1995 model conceptualizes human information processing as a series of internal states that are constantly being revised and updated

as new information (the input) enters and is acted upon. In the model, there are three distinct yet interdependent stages: the selective perception stage, the parsing stage, and the utilization stage. In the selective perception stage, the input (in this case, spoken input) is registered in sensory memory, with the listener attending to at least some of the input, and this attended-to input (sometimes referred to as intake) is transferred to working memory. In the parsing stage, the attended-to input stored in working memory is acted upon, integrating this new information with previously known information, and moved to long-term memory. In the third stage, the information is retrieved from long-term memory and is used for a particular purpose. These three stages are partially ordered in the manner in which the input is transmitted and received, yet each stage overlaps with the other stages and influences the processing of the information at each stage. Anderson's model has evolved over time, but this conceptualization seems to be the dominant one for most models of L2 listening ability.

Another important and relevant model of memory and cognition that has been applied to (some) models of L2 listening ability is Paivio's dual coding theory (1971, 1986, 1991). In this theory, verbal systems represent the properties of language, and nonverbal systems represent the nonlinguistic world. Both of these classes exist together in different modalities. Visual linguistic input is prototypically printed words, while visual nonlinguistic input would include visual objects; auditory linguistic input is prototypically spoken words, while auditory nonlinguistic input would include environmental sounds. An important component of the theory is the idea of functional independence of subsystems. That is, the verbal and nonverbal systems are obviously interrelated, yet still independent, as evidenced by the fact that one system can be active while the other is inactive, or they can be active in parallel (Paivio, 1991). In a listening context at a coffee shop, for example, a listener might be simultaneously processing aural linguistic information (e.g., spoken words) spoken by a conversation partner, as well as aural non-linguistic information (e.g., background coffee shop noises). That listener might also be processing visual linguistic information (e.g., signs in the coffee shop), and visual non-linguistic information (e.g., the gestures and appearance of the speaker, objects and people in the coffee shop).

### 1.1 Taxonomies and data-driven examinations of L2 listening ability

Instead of creating formal models of L2 listening ability, a number of researchers elected to create taxonomies of the skills or sub-skills that are involved in the L2 listening process (e.g., Aitken, 1978; Lund, 1991; Petersen, 1991; Richards, 1983; Weir, 1993). Perhaps the most influential of these taxonomies was Richards (1983). Richards created a taxonomy of micro-skills for both academic listening and

conversational listening. For academic listening, Richards listed 18 micro-skills, including the “ability to identify the purpose and scope of lecture”, the “ability to identify topic of lecture and follow topic development”, and the “ability to identify relationships among units within discourse” (p. 229). Similarly, Richards listed 33 micro-skills that make up conversational listening ability, including the “ability to retain chunks of language of different lengths for short periods”, “the ability to discriminate among the distinctive sounds of the target language”, and “the ability to recognize the functions of stress patterns of words” (p. 228).

In almost direct contrast to the taxonomy listing/creation are data-driven examinations of listening ability. These data-driven examinations have focused on L2 listening test performance, and how the characteristics of the texts and tasks used on those test affect test-taker performance. Two examples are Freedle and Kostin (1999) and Nissan et al. (1996). These two studies are similar, in that they examined test-taker performance on the TOEFL listening section in the hopes of identifying variables that would account for the variance in item difficulty on the tests. Freedle and Kostin examined three different item types: main idea, inference-application (in which test-takers must use their background knowledge to make the appropriate inferences), and inferencing (the test-taker must make appropriate inferences based on information that is available in the spoken text). Using a regression analysis, they found that main idea items and inference-application items were significantly easier than inferencing items, and identified a number of text variables (e.g., negatives, referentials, rhetorical organizers, fronted structures, concrete texts, multisyllabic words, lexical overlap, subject matter, pauses and fillers) that affected the difficulty of each of these item types. Similarly, using a regression analysis, Nissan et al. (1996) found five variables that contributed to the variance in item difficulty, although the three best predictor variables (i.e., lower frequency vocabulary, implicit information in the text, dialogue finishing with a statement and not a question) only accounted for 12% of the variance.

While these theory-based taxonomies and data-driven lists of variables that affect listening performance are useful in many ways (i.e., for curriculum developers, test writers, and teachers), they are less useful for research purposes. The taxonomies do not “provide clear definitions or non-redundant orderings of components in any systematic graded hierarchy” (Dunkel, Henning, & Chaudron, 1993, p. 182), nor do they usually provide a hierarchy of which of the sub-skills are more important or fundamental to the listening process. Indeed, these taxonomies are generally hypothetical listings of things involved in listening ability, and they usually have little or no empirical research validating them. In contrast, while the data-driven variables are useful in identifying what aspects of listening are difficult, they tend to be focused on the lower-order components of listening, and are less able to provide a bigger picture definition or overall model of listening ability.

Indeed, it is this identification of the shortcomings of previous taxonomic papers and data-driven listings of listening variables that motivated Buck and Tatsuoka's (1998) empirical exploration of the knowledge, skills and abilities that are part of L2 listening ability. They used a rule-space procedure to identify the knowledge and cognitive processing skills required on their particular L2 listening test, and then examined how each of these attributes affected the listeners' performance on that test (rule-space methodology is a type of statistical pattern recognition technique that provides information about how individual test-takers perform on each of the attributes). 412 Japanese test-takers completed an open-ended, short-answer listening comprehension test, and rule-space methodology was used to provide information about the cognitive attributes ("anything that affects performance on a task: either a task characteristic, or any of the knowledge, skills or abilities necessary to complete the task" [p. 121]) that contributed to test performance. They identified 15 primary attributes and 14 interaction attributes that explained 96% of the variance for 96% of the test-takers. The 15 primary attributes included things like "the ability to scan fast spoken text, automatically and in real time", "The ability to use previous items to help information location", and "The ability to understand and utilize heavy stress", and "The ability to make text-based inferences" (pp. 141–142). Their list of 15 primary attributes mirrors skills or sub-skills identified in many of the listening taxonomies described above. The authors acknowledge that test-taker attributes interact with the attributes of the test itself (the spoken text and items), and thus they avoid some of the pitfalls of the taxonomies that focused solely on the listeners' ability, and the data-driven lists, that focused solely on the characteristics of the test. While Buck and Tatsuoka's study was informative, it was based on a single set of test-takers taking one particular test, and the "model" that resulted from it was never really developed beyond the results of the study.

## 1.2 Models of second language (L2) listening ability

Many of the models of L2 listening ability described here are focused on assessment. This is not surprising, because principled assessment entails choosing or creating a construct definition of the ability to be assessed, and then an operationalization of that model.

A common theme in many models of L2 listening is the idea that listening involves both bottom-up processing and top-down processing. Bottom-up processing occurs when the listener perceives the aural input and tries to interpret and build up the meaning bit by bit, word by word (Kelly, 1991). It involves speech perception and word recognition, which provide the data for the listener in attempting to decode the utterance in trying to comprehend it (Rost, 2002). Top-down processing requires the listener to use their background, contextual, and world knowledge to

set up expectations and create a conceptualization of what the utterance means. Kelly (1991) describes it as “...the application of cognitive faculties in the attempt to give the sound input meaning. The mind sets up the expectations and the sound provides confirmation” (p. 135). The two processes work in tandem and are constantly influencing each other. At times one process might be dominant, and at other times the other process. This idea of the two different types of processing occurring simultaneously and integratively is appealing both logically and intuitively, and it is also helpful pedagogically, but in practice it is difficult to operationalize them separately, since there are so tightly interrelated.

Perhaps the most influential and useful model of L2 listening ability was presented by Buck (2001) in his book, *Assessing listening*. Based on his previous empirical research (e.g., Buck, 1991, 1994; Buck & Tatsuoaka, 1998; Buck, Tatsuoaka, Kostin, & Phelps, 1997), Buck provided what he called a “default listening construct” (p. 113). In presenting this construct definition, he provided a number of recommendations to consider when creating listening tests, including focusing on those aspects of language that are unique to listening; requiring the listener to understand basic linguistic information on different topics and texts; going beyond the literal meaning and also assessing inferred meanings; considering how to deal with knowledge-dependent interpretations of the text; and avoid assessments that assess general cognitive abilities (Buck, 2001). His formal definition of the default listening construct is the ability:

- to process extended samples of realistic spoken language, automatically and in real time,
- to understand the linguistic information that is unequivocally included in the text, and
- to make whatever inferences are unambiguously implicated by the content of the passage. (Buck, 2001, p. 114)

The first bulleted point addresses a number of issues that other models bypass or miss altogether. According to Buck, L2 listening ability involves the ability to process “extended” samples of spoken language. It involves the ability to process “realistic” spoken language (and in fact, many assessments purporting to assess a listener’s communicative competence often do not use spoken texts with “realistic” spoken language, a point that will be expanded on below). Listening requires the listener to process the spoken language “automatically and in real time”, again, addressing how listening actually occurs in real-world situations. The second bulleted point is relatively straightforward, involving the ability to understand explicitly stated information. The third bulleted point describes the ability to make appropriate inferences from the spoken input.

This “default listening construct” is useful because it is simple and straightforward, with just three bulleted points, yet these three points address the construct broadly and thoroughly, and it can be adapted to different listening situations to create a more contextualized definition of L2 listening ability.

Wagner (2002) reviewed the literature and posited a model of L2 listening ability based on previous L2 taxonomies and models, including Buck’s (2001) model. He posited a two-factor model of L2 listening ability in an academic listening domain. The two factors were “the ability to perform bottom-up processing” and the ability to “perform top-down processing”. The bottom-up processing factor was operationalized through two sub-skills “the ability to identify details and facts” in the text, and the ability to “recognize supporting ideas” in the spoken texts. The top-down processing factor was operationalized through four sub-skills: the ability to identify the controlling idea/gist; ability to make text-based inferences, the ability to make inferences about speakers’ attitudes and pragmatic meaning, and the ability to deduce vocabulary through context (Wagner, 2002, p. 12). Wagner operationalized this model of L2 listening ability and created a 20-item (multiple choice and limited production items) video listening test that he administered to 85 high school ESL students in the U.S. He then used the results of this test to perform a series of exploratory factor analyses (EFAs) to examine the validity of his theorized model of L2 listening. The results of these EFAs provided little evidence in support of the original model, but based on these results, Wagner re-interpreted the model as a two-factor model of L2 listening ability: the ability to listen for explicitly-stated information, and the ability to listen for implicit information. This revised model corresponds very closely to Buck’s (2001) model of L2 listening ability.

Wagner’s (2004) study built on his (revised) 2002 model. He posited a two-factor model of L2 listening ability: the ability to comprehend “explicitly stated spoken information”, and the ability to comprehend “implicit spoken information” (p. 9). He then tried to validate this model based on the results of test-takers’ performance on the listening sections of the MELAB ( $n = 823$ ) and the ECPE ( $n = 11,212$ ). The results of the EFAs showed limited support based on the MELAB data, but little evidence based on the ECPE data.

Rost’s (2011) book *Teaching and research listening* (2nd edition), is a very important book in the L2 listening literature, even though Rost did not explicitly define a model of listening ability. Instead, the first section of the book attempts to define listening by describing the various processing that listening entails, including neurological processing (i.e., consciousness, hearing, and attention); linguistic processing (i.e., perceiving speech, segmenting, grouping, and parsing); semantic processing (i.e., constructing meaning by integrating memory and prior experience); and pragmatic processing (i.e., constructing meaning by inferring speaker

intention). This exploration of the various processes involved in L2 listening is a useful resource for attempts at operationalizing models of L2 listening ability.

Vandergrift and Goh (2012) developed their working cognitive model for L2 listening comprehension based on Levelt's (1983, 1989, 1995) model of speech production. Although technically a model of speaking, Levelt's model of producing and monitoring speech is relevant because by integrating the speaking and comprehension components together in one model, it allows a conceptualization of both one-way and two-way (interactive) listening. In Vandergrift and Goh's (2012) model, listening begins with the perception of sound signals by an acoustic-phonetic processor. This information is stored very briefly in working memory while processed for meaning, and then the sounds are displaced by the subsequent incoming sounds. Analysis of the speech sounds then begins, involving both bottom-up and top-down processing. The next stage involves parsing, in which the phonetic representations of the perceived sounds are parsed for meaning, by segmenting the utterance. This segmenting is done using both syntactic and semantic cues. Again, the processing activity is done not linearly but in a parallel fashion, in which the bottom-up processing informs the top-down processing (and vice versa), until a reasonable mental representation of the meaning of the spoken utterance is created. This is done by the listener by linking the mental representation of the spoken sounds with the listener's existing knowledge from long-term memory. The listener creates a mental representation of their interpretation of what they have heard, and stores this in long-term memory.

Vandergrift and Goh stress that the aspect of their model that has not been adequately addressed in previous cognitive models of listening ability is the idea of metacognition, the idea that the listener can have conscious control of the listening process, including planning, monitoring, problem-solving, and evaluating. They argue that listeners with more metacognitive awareness are better able to control and regulate the cognitive processes involved in listening. Vandergrift and Goh stressed that it was a working model, because there is no widely accepted comprehensive theory that adequately explains the comprehension process. They also stressed that theirs was not a comprehensive model, acknowledging the fact that listening occurs in various social contexts, and thus a comprehensive model would have to include the affective components of listening.



### 1.3 L2 listening, L1 listening, and reading

Models of L1 reading and L1 listening have been influential in modeling and operationalizing L2 listening ability. Obviously, listening in an L2 is very similar to listening in one's L1. Indeed, virtually all of the models of L2 listening ability described here are based almost entirely on conceptualizations of L1 listening ability. Yet it also seems obvious that there are differences between L1 and L2 listening. Buck (2001) argues that they are fundamentally similar, yet differ in "emphasis" (p. 48), suggesting that L2 listeners are affected by an incomplete knowledge of the linguistic system, content or textual schemata, background knowledge and cultural knowledge.

Vandergrift (2006) examined whether L2 listening proficiency was more aligned with overall L2 language proficiency, or L1 listening ability. He wanted to examine empirically the linguistic interdependence hypothesis, which predicts that L2 listening performance is largely predicted by L1 listening ability. The idea is that "learners do not relearn a language skill; rather the skill is available to them as needed within another language context" (Vandergrift, 2006, p. 6), and Vandergrift argues that this hypothesis has been researched extensively in the L2 reading literature (e.g., Bernhardt & Kamil, 1995; Lee & Schallert, 1997). Because this literature suggests that both L2 proficiency and L1 reading proficiency are predictors of an individual's L2 reading proficiency, Vandergrift wanted to see if the same applies for L2 listening. He divided 75 L1 English speaking 8th graders in Canada who were learning French into a high ability and a low ability group. The 75 participants took both a French (L2) listening test and an English (L1) listening comprehension test. He found that while both L2 proficiency and L1 listening ability predicted group membership (high ability or low ability), and that L2 proficiency accounted for about 25% of the variance, L1 listening ability accounted for about 14% of the variance. The results from Vandergrift (2006) seem to be in line with Hulstijn's (2015) BLC-HLC model of language ability, which posits that higher language cognition (HLC) is a complement or extension of basic language cognition (BLC – implicit and unconscious phonetic, prosodic, phonological, morphological, and syntactic knowledge, in conjunction with explicit and conscious lexical knowledge). According to this theory, linguistic knowledge explains most variance in language performance, but as proficiency increases, other peripheral factors such as general cognitive ability, L1 ability, metacognition, etc., are able to explain at least some of the variance in language performance.

Just as there are many similarities between L1 and L2 listening, there are also a number of similarities between listening and reading (including L2 listening and L2 reading), and conceptualizations and models of reading have been influential on conceptualizing and modeling L2 listening ability. Indeed, the two processing models that began this chapter are prime examples of this phenomenon. Anderson's

cognitive processing model is applicable for both reading and listening, although Paivio's dual coding theory goes much farther in acknowledging the different modalities involved.

Aotani (2011) reviewed the research related to the similarities and differences between L2 reading and listening. He described two opposing views in the literature. The unitary process view is based on the idea that L2 listening and L2 reading are more or less the same, except that the mode of the input is different (spoken versus written). In contrast, the dual process view holds that there are real and measurable differences between L2 reading and listening. Aotani argued that the dual process view is now more accepted in the field. As evidence for this, numerous researchers (e.g., Buck, 2001; McCarthy & Carter, 1995; Vandergrift, 2007; Wagner 2014a, 2014b, 2018) have described how L2 reading and listening differ, especially due to the online nature of listening, which prevents listeners from having the chance to go back and re-process the input. In addition, the non-verbal and paralinguistic (i.e., stress, rhythm, and intonation) components of spoken language are important parts of listening, and do not have equivalents in reading. Finally, the nature of written texts differs fundamentally from the nature of spontaneous spoken texts, as will be discussed below.

## **2. Measurement practices: A critical discussion of how the constructs and their components have been operationalized and measured by empirical researchers**

L2 listening ability is more difficult to operationalize and measure than the other language skills (i.e., reading, writing, speaking) for a number of reasons. Like reading, it is an internal process. Listening is also difficult to measure because it requires the selection or creation of spoken texts to present to the listeners, and choosing or creating the appropriate texts is surprisingly more difficult than many researchers anticipate. Presenting the spoken texts is also challenging, as the researcher must decide whether to do it "live" (with a human speaking the text), or utilize technology such as audio or audio-visual recordings. Numerous researchers have argued that when testing listening ability, it is necessary to include and even focus on those aspects that are unique to listening ability (e.g., Buck, 2001; Rost, 2011; Wagner, 2014b), yet many of the characteristics that are unique to listening have not been included in the operationalization and measurement of listening ability. What follows is a critical discussion of the way L2 listening ability has been operationalized and measured in the field, focusing on: listening task response, differentiating between reading and listening, the length of the spoken text used as the input, assessing interactive speaking/listening ability, using audio-only versus audio-visual input, the

accent/dialect variety of the spoken input, and the use of real-world spoken input. This critical discussion is grounded in the idea of the need to include those aspects of listening that are unique to listening when operationalizing and measuring L2 listening ability.

## 2.1 Interactive speaking/listening ability

Listening has traditionally been operationalized and measured as a one-way activity. Obviously, one-way listening is common, and part of many domains of interest for L2 researchers and testers. Academic listening (including listening to lectures) often involves one-way listening. But listening ability is often intertwined with interactive speaking and listening, in which the participant is both a listener and speaker. There seems to be growing recognition that the operationalization of listening ability should also include listening ability as part of an interactive speaking/listening ability (Ockey & Wagner, 2018). There also seems to be growing recognition that while isolating a single skill might be appropriate in some cases, the overarching idea of interaction and communicative competence in second language acquisition research necessitates broadening the construct.

Nevertheless, even with this increased awareness of listening being part of a broader communicative construct, many researchers have focused on the one-way, listening-only aspect of listening ability. There are probably many reasons for this, including the fact that listening researchers want to focus on listening only, and not on speaking. In addition, assessing listening ability as part of interactive speaking/listening ability certainly presents definite measurement challenges (Nakatsuhara, 2018; Ockey, 2018). Nevertheless, these are challenges that L2 listening researchers will have to address, because conceptualizing and operationalizing the incredibly broad and diverse construct of listening ability without including interactive speaking/listening ability results in an obviously impoverished construct.

## 2.2 Differentiating listening from reading

As described above, numerous researchers have described the many differences between reading and listening, and there is an increasing realization that it is necessary to think of them as different processes. Field (2008) describes how L2 listening ability has frequently been conceptualized as a set of sub-skills, and that these sub-skills combine and interact in the overall process of listening. This idea of identifiable and distinct sub-skills is common in reading, and thus has been applied to listening, but Buck (2001) criticizes the notion that the reading “subskills” can be

directly applied to listening. He argues that what these sub-skills are in listening is unclear, and also describes the difficulty in operationalizing and measuring them, and differentiating between them.

Yet many models and operationalizations of listening ability seem to continue to depict listening as essentially the same as reading, except for the input being verbal rather than written. A good example of this are studies that have examined the amount of shared variance between L2 reading and L2 listening ability (e.g., Aotani, 2011; Song, 2008). These studies have found that listening and reading do correlate quite highly. However, the tests of listening ability in these studies seem to be essentially L2 reading tests that have written input that is read aloud, rather than including real-world spoken input, and so it is possible that the two are overly conflated. Again, Buck (2001) and Rost (2011) argue that when operationalizing and assessing listening ability, it is important to focus on those aspects that are unique to listening ability. Indeed, many L2 listening tests continue to operationalize listening ability as if it were the same as reading ability.

In his (2001) default listening construct, Buck very specifically described how proficient L2 listeners had to be able to do automatic processing, and that listeners did not have the luxury of utilizing controlled processing in most listening contexts. Because of the nature of spoken texts, the rate of the input is not controlled by the listener, but by the speaker. This is in great contrast to reading, in which the reader controls the rate of input, and thus is more able to utilize controlled processing in processing the input. This leads to two natural critiques of how L2 listening has been operationalized. The first critique has to do with the number of times a spoken text is presented to test-takers. Buck (2001) argues that it is artificial and problematic to present the spoken text multiple times to test-takers. Doing so allows them to utilize controlled processing, and is inauthentic in that, in most real-life listening contexts, the listener does not have the opportunity to listen to the text multiple times. The second critique has to do with the speech rate of the spoken input. The research is clear that speech rate affects L2 listening comprehension (e.g., East & King, 2012; Griffiths, 1992; McBride, 2011). Basically, there is an inverse relationship between the rate of the input and comprehension – the higher the speech rate, the lower the comprehension (there are caveats, of course, but this is the general rule). Again, according to Buck (2001), using artificially low rates of spoken input allows L2 listeners to utilize controlled processing, when in real-life situations, this would not be possible. Similarly, Wagner (2016); Wagner and Wagner (2016), and Wagner and Ockey (2018) have criticized the use of carefully scripted and overly-enunciated spoken texts in L2 listening assessments, in part because they have lower speech rates than real-world speaking and listening contexts, and thus are not representative of the ability needed in most listening contexts.

### 2.3 Length of the spoken text

Buck's (2001) default listening construct also explicitly included the importance of listening assessments including "extended samples" of spoken language (p. 114). This seemed to be in response to earlier operationalizations and assessments that only included listening to very short (word or sentence level) spoken texts. Obviously, listening ability involves the ability to comprehend spoken texts of varying lengths, from word and sentence level utterances to very long texts (e.g., academic lectures). Using only very short oral texts is attractive from a practicality standpoint, as shorter texts are easier to create and easier to administer, and their use allows for better control of other variables (i.e., memory). But to accurately assess listening ability, spoken texts of varying lengths are needed, in order to adequately represent the types of listening that would be expected in the particular domain of interest. Somewhat surprisingly, there does not seem to be any recent empirical research investigating how the length of a spoken text can affect L2 listening performance, although obviously listeners need to be able to understand spoken texts of varying lengths.

### 2.4 Listening task response

Because it is not possible to "get inside the listener's head", listening tests usually require test-takers to listen to spoken input, and then do something with that input to demonstrate comprehension (i.e., answer comprehension questions, respond orally, fill out a chart, do a dictation, etc.). Based on their responses, the researchers make inferences about their actual listening ability. Thus, the type of task response that is utilized is very important in operationalizing and measuring the construct.

One seemingly obvious task type would be a written recall or dictation. But a dictation task presents problems when administering (when to insert pauses for writing, how many times to play the text), and with scoring because scoring a dictation reliably is surprisingly difficult (Buck, 2001). Another concern is that dictation includes much more than just listening ability (i.e., writing ability, memory, etc.). And of even more concern, when it comes to dictation, it is unclear to what extent a dictation task actually assesses listening comprehension. Requiring listeners to transcribe a spoken text word for word is very unlike real-world listening, where the listener has to segment the input and ascertain the semantic and syntactic meaning of this input, and compare this information with the listener's background knowledge (and contextual and co-textual knowledge) (Wagner, 2014a). In addition, dictation probably does not assess listeners' ability to make inferences (Buck, 2001), which would seem to be a vital component of listening ability. Even with

these limitations, dictations can be useful for L2 listening researchers as a listener's response can be analyzed to see what aspect of listening to the spoken text the listener had difficulty with (e.g., vocabulary, a particular grammatical or syntactic structure, etc.), although again listeners tend to focus on semantic processing, while writing the dictation would seemingly require syntactic processing. Listening cloze tasks are similar to dictations tasks, and have many of the same shortcomings as dictation tasks. Oral repetition tasks, in which the test-taker has to listen to a spoken text (usually at the word, phrase, or sentence level) and then orally repeat the input, also have much in common with dictation.

Listening summarization tasks are those in which the test-taker listens to a spoken text, and then has to write or say a summary of the text, or recall as much of the text as they can. They differ from dictations in that the listener is not expected to provide a word-for-word recall of the spoken text. Instead, the listener summarizes in their own words what they have heard. Such a task is more similar to real life listening contexts than dictation tasks, which require the word-for-word recall. Some disadvantages of these types of tasks include scoring challenges, due to the fact that writing (or speaking) ability is also being assessed. Also being assessed is summarization ability as well as memory, and it is unclear the extent to which these two abilities are part of the construct of L2 listening.

In contrast to these dictation-type tasks, comprehension tasks are probably more commonly used in many testing situations, as well as in research areas in which listening ability is operationalized and measured. Perhaps the most common type of comprehension task involves comprehension questions that the listener must answer based on what s/he heard. Most listening tests employ some variety of listening tasks involving discrete point comprehension questions, in which the listener listens to a spoken text and then chooses the correct answer (i.e., multiple-choice) or writes or states the correct answer (i.e., short answer constructed response). These types of tasks have many advantages, in that they are relatively easy to create and score reliably, and they seem to be able to tap into the listening comprehension process. Researchers have used these types of comprehension tasks extensively, in part because they have the advantage of allowing the researcher to assess the different components of their particular model of listening ability. Thus, a particular comprehension question (or questions) can be created that purportedly tap into a particular ability, skill, or sub-skill of listening. Within a single testing situation, multiple questions can be used to assess each of the particular abilities/skills of interest, thus resulting in a more reliable and valid assessment.

## 2.5 Audio-only versus audio-visual input

Whether listening ability involves the ability to comprehend nonverbal information is a question that has vexed L2 listening researchers for decades. On the one hand, listening ability has traditionally been focused on a person's ability to process and comprehend verbal information, and there is the sentiment that the non-verbal components of spoken language were somehow extraneous to actual listening ability (Buck, 2001; Coniam, 2001). In contrast, Wagner (2008, 2010a, 2010b, 2013) has argued that in most domains, the ability to process the visual, nonverbal information that accompanies a speaker's verbal output should be considered part of the construct of listening ability. In most real-world listening contexts, the listener can see the speaker, and the visual, nonverbal information provided by the speaker's body language, lip movements, and gestures, as well as the contextual information provided by the setting (i.e., the speaker's physical appearance, the physical background setting, etc.). The listener processes this visual information while simultaneously processing the verbal information, and uses the different input sources to try and make sense of the message (Paivio, 1971, 1986, 1991). Similarly, Ockey (2007) has argued that an "expansion of the construct" is necessary, in order to go beyond this traditional notion of listening ability including only aural input. Yet many models and operationalizations of listening ability include only aural input, or include nonverbal information only as an afterthought.

## 2.6 Accent and dialect variety

When operationalizing models of L2 listening ability, a researcher must identify the accent variety or varieties of the spoken texts used for the aural/audio-visual input. Traditionally, the perceived standard variety of the language in that research context has been used. Using English as an example, if a researcher was conducting research in the UK, then standard British English was used as the language variety of the input, and seemingly "accented" speech was avoided. One of the problems with this approach, of course, is identifying the "standard" variety, as well as who gets to determine what the standard variety is (or even if a "standard" variety actually exists). Widespread languages will have a variety of accents and dialects, based on national, regional, cultural, and other differences. In addition, many speakers of a language are non-native speakers of that language. Indeed, English is spoken by more nonnative speakers of English than by native speakers.

A highly proficient listener is able to comprehend multiple varieties of that language, not just the standard variety. Two areas of research are relevant here. One, multiple studies have found that familiarity with an accent leads to increased



comprehension of that accent (e.g., Gass & Varonis, 1984). Second, it is widely accepted in L2 pronunciation research that a speaker can be accented, but still be comprehensible. Accentedness and comprehensibility are correlated, but still separate constructs (Isaacs, 2008). For many languages (and certainly English), a proficient listener must have multidialectal listening skills (be able to understand a number of different speech varieties), as well as be able to accommodate or adapt to unfamiliar varieties of spoken language (Canagarajah, 2006).

## 2.7 Real-world spoken input

Perhaps the most problematic way that L2 listening ability has been modeled and operationalized involves the inauthentic nature of the spoken texts used in these operationalizations. Depending on the context of the speaking situation, a speaker might be able to plan exactly what s/he is going to say, or s/he might have virtually no chance to plan the speech. Wagner (2013, 2014a, 2014b, 2018) has described how there is a “continuum of scriptedness” with spoken language. At one end of the continuum is speech that is totally planned. This might involve writing a text, editing and polishing it, and then speaking it aloud, with careful enunciation. At the other end of the continuum is speech where the speaker composes the message and verbally utters it at virtually the same time, with no planning. This level of scriptedness affects the characteristics of the speech. As has been demonstrated repeatedly in the literature, unplanned and unscripted spoken language is “messy”, and is filled with connected speech, filled and unfilled pauses, repetitions, false starts, etc. In contrast, texts that are scripted and then read aloud might have very few of these phenomena.

That spoken texts vary in the extent to which they have these different phenomena is not in doubt or controversial. There is a long history of applied linguistics researchers investigating these phenomena in speech. It seems obvious that a proficient listener would be able to listen to and comprehend spoken texts with differing levels of scriptedness (from different points on the continuum of scriptedness). Yet when listening ability is operationalized through assessments that involve listening to spoken language, the spoken language almost invariably comes from the scripted end of the continuum (Field, 2013; Wagner 2016; Wagner & Wagner, 2016; Yanagawa, 2016). This is problematic for two main reasons. First, the unscripted spoken language is much more prevalent in most real world speaking and listening contexts. In most cases, the speaker does not have time to plan, and thus language that has connected speech, hesitation phenomena and non-linear organizational patterns is the norm, not the exception. Second, an increasing amount of research (e.g., Carney, 2018; Wagner, 2018; Wagner & Toth, 2014) demonstrates that many



L2 listeners have more difficulty comprehending unscripted spoken language (that has these characteristics of real-world, unplanned spoken language) compared to scripted spoken language. Researchers such as Gilmore (2007) and Wagner (2014a) have argued that many classroom L2 learners are exposed primarily to “textbook texts”, texts that are scripted and created especially for L2 learners, and that lack the characteristics of real-world, unplanned spoken language. This lack of exposure to and teaching about real-world spoken language can result in L2 learners that are proficient in understanding scripted spoken language, yet are unable to understand real-world spoken language (Brown & Trace, 2018; Wagner 2018; Wagner & Ockey, 2018). Again, a proficient listener needs to be able to comprehend different varieties of spoken texts, from a variety of genres, and from a variety of formality levels. As Wagner has repeatedly argued (Ockey & Wagner, 2018; Wagner, 2014b, 2016; Wagner & Wagner, 2016), to assess L2 listeners’ listening proficiency using only spoken texts that are scripted and read aloud results in an overly narrow operationalization of the construct of L2 listening ability. This critique applies not only to L2 listening assessment, but also to SLA research purporting to investigate L2 listening within a communicative competence framework.

### 3. Conclusion

It is easy to critique the way L2 listening ability has been modeled and operationalized in previous research, as none of these critiques are new, and are found widely in the literature. Numerous researchers have argued that if the goal is to measure listening ability, then the measurement must include and even focus on those aspects that are unique to listening ability (e.g., Buck, 2001; Rost, 2011; Wagner, 2014b). But as the critique above demonstrates, many of the characteristics that are unique to listening have been consciously or unconsciously avoided in the operationalization and measurement of listening ability. The question then becomes “Why have these aspects unique to listening been avoided in its modeling and operationalization?” Although it is probably impossible to fully answer this question, what follows is an (speculative, yet informed) examination of some of the probable reasons.

The first reason is tradition. L2 listening ability has traditionally been operationalized (and tested, and taught) as one-way, non-interactive listening, using scripted, audio-only texts, that are revised and edited and then spoken aloud, often with a speaker especially trained to enunciate clearly. In relation to the spoken texts used in L2 textbooks and assessments, Gilmore (2007) and Wagner (2014) have argued that these industries are conservative, and reluctant to change the status quo. Related to this idea of conservatism, Wagner (Wagner, 2014b, 2018; Wagner & Wagner, 2018) has argued that the L2 testing industry might be reluctant to use

unscripted spoken texts because they might sound unprofessional. In other words, using recorded spoken texts that have filled pauses, overlaps, digressions, repetitions, and false starts, and with lots of reduced and connected speech, might cause test-takers (and other test users) to perceive the test as unprofessional and second rate. While it is certainly understandable that test developers might be reluctant to use spoken texts that sound unprofessional, Wagner (2018) argues that test developers can avoid this problem by stressing that they are using unscripted, authentic, spoken texts that have the characteristics of real-world spoken language, and thus their tests better assess the communicative competence of the test-takers.

Secondly, it is more difficult, less efficient, and more expensive to identify and find authentic spoken texts that are suitable for L2 listening tests than it is to create spoken texts that address the particular needs of a researcher or test developer. Wagner (2014b, 2018) describes how in most testing contexts, test developers have a number of pre-determined test specifications (including length/duration, number of questions, skills to be assessed, etc.), and it is much easier to create a spoken text that is specifically scripted to address these test specifications, than it is to find a real-world text that would be suitable. Similarly, using an audio-visual text presents cost and logistical issues for test developers – more expenses related to the creation and delivery of audio-visual texts. But if the goal is to operationalize a model of L2 listening ability that really tries to address communicative competence, and genuinely tries to assess a language user's ability to understand spoken language in real world contexts (rather than in a language laboratory), then the added expense involved in using authentic spoken texts seems necessary to incur.

Thirdly, it seems likely that many of the models of listening are based at least in part on models of reading. As described above, both are internal processes involving the processing of input, and they obviously have many similarities. Anderson's cognitive model and Paivio's dual coding theory are applicable to both reading and listening. It seems likely that research into reading has been more prevalent than listening research, because reading must be taught even in the L1, while listening is rarely taught in the L1. Within the cognitive paradigm, the idea is that the cognitive processing goes on inside the mind of reader/listener, and the focus is then on the input and the processing, and does not acknowledge some of the "messier" aspects of real-life communication. Again, this is more defensible with reading, where the written input is on the printed page, and there is a greater degree of uniformity than with listening, which has a multitude of characteristics that can affect the input (e.g., accentedness, background noise, mumbling, etc.). This reliance on reading as the dominant processing paradigm seems to have led to listening being operationalized as just a variation of reading, involving the processing of oral, rather than written text. The problem is that an unplanned, spontaneous oral text is very different from a planned, scripted, edited, and polished written text that is read aloud slowly with

clear enunciation. Again, listening and reading have many similarities, but rather than operationalizing them as if they are virtually identical (except for the channel of the input) it is necessary to include the characteristics that are unique to listening when operationalizing listening.

This also seems analogous to the competence/performance distinction made by Chomsky and his acolytes in transformational grammar. The way listening has been operationalized focuses on the idea of competence (the internal state of the mind), while real-world listening performance (and listening to spoken input with real-world linguistic characteristics) is neglected. There seems to be the recognition that these real-world characteristics of unplanned spoken language (e.g., hesitation phenomena, oral grammatical norms, connected speech, non-linear discursive organizational patterns) exist, but somehow they are superfluous to the underlying competence involved in listening. The problem with this view, of course, is that listening happens in the real world, where spoken texts are messy, with backtracking and false starts, repetitions, slang, connected speech, and mumbling. Field (2013) criticizes the way L2 listening has been operationalized on many listening tests, and argues for the need that these operationalizations have cognitive validity, which he defines in relation to assessing listening as “the extent to which the tasks employed succeed in eliciting from candidates a set of processes which resemble those employed by a proficient listener in a real-world listening event” (p. 77).

To conclude, when modeling and operationalizing L2 listening ability, it is important that researchers consider the nature of real-world listening. Real-world listening is “messy”, and requires a host of skills beyond the previous idealized, listening-as-reading operationalizations.

## References

- Aitken, K. (1978). Measuring listening comprehension. *English as a second language. TEAL Occasional Papers* (Vol. 2). British Columbia Association of Teachers of English as an Additional Language. (ERIC Document Reproduction Service No. ED155945)
- Anderson, J. (1985). *Cognitive psychology and its implications* (2nd ed.). Freeman.
- Anderson, J. (1990). *Cognitive psychology and its implications* (3rd ed.). Freeman.
- Anderson, J. (1995). *Cognitive psychology and its implications* (4th ed.). Freeman.
- Aotani, M. (2011). Factors affecting the holistic listening of Japanese learners of English (Unpublished doctoral dissertation). Temple University Japan, Tokyo, Japan.
- Bernhardt, E., & Kamil, M. (1995). Interpreting relationships between L1 and L2 reading: Consolidating the linguistic interdependent hypotheses. *Applied Linguistics*, 16(1), 15–34. <https://doi.org/10.1093/applin/16.1.15>
- Brindley, G. (1998). Assessing listening abilities. *Annual Review of Applied Linguistics*, 18, 171–191. <https://doi.org/10.1017/S0267190500003536>

- Brown, J. D., & Trace, J. (2018). In G. Ockey & E. Wagner (Eds.), *Assessing L2 listening: Moving towards authenticity* (pp. 45–63). John Benjamins.
- Buck, G. (1991). The testing of listening comprehension: An introspective study. *Language Testing*, 8(1), 67–91. <https://doi.org/10.1177/026553229100800105>
- Buck, G. (1994). The appropriacy of psychometric measurement models for testing second language listening comprehension. *Language Testing*, 11(2), 145–170. <https://doi.org/10.1177/026553229401100204>
- Buck, G. (2001). *Assessing listening*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511732959>
- Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing*, 15(2), 119–157. <https://doi.org/10.1177/026553229801500201>
- Buck, G., Tatsuoka, K., Kostin, I., & Phelps, M. (1997). The sub-skills of listening: Rule-space analysis of a multiple-choice test of second language listening comprehension. In A. Huhta, V. Kohonen, L. Kurki-Suonio, & S. Luoma (Eds.), *Current developments and alternatives in language assessment*. Universities of Tampere and Jyväskylä.
- Canagarajah, S. (2006). Changing communicative needs, revised assessment objectives: Testing English as an international language. *Language Assessment Quarterly*, 3(3), 229–242. [https://doi.org/10.1207/s15434311laq0303\\_1](https://doi.org/10.1207/s15434311laq0303_1)
- Carney, N. (2018). Diagnosing L2 bottom-up listening abilities of Japanese university EFL learners (Unpublished doctoral dissertation). Temple University Japan, Tokyo, Japan.
- Coniam, D. (2001). The use of audio or video comprehension as an assessment instrument in the certification of English language teachers: A case study. *System*, 29, 1–14. [https://doi.org/10.1016/S0346-251X\(00\)00057-9](https://doi.org/10.1016/S0346-251X(00)00057-9)
- Dunkel, P., Henning, G., & Chaudron, C. (1993). The assessment of an L2 listening comprehension construct: A tentative model for test specification and development. *The Modern Language Journal*, 77(2), 180–191. <https://doi.org/10.1111/j.1540-4781.1993.tb01962.x>
- East, M., & King, C. (2012). L2 learners' engagement with high stakes listening test: Does technology have a beneficial role to play? *CALICO Journal*, 29, 208–223. <https://doi.org/10.11139/cj.29.2.208-223>
- Field, J. (2008). *Listening in the language classroom*. Cambridge University Press.
- Field, J. (2013). Cognitive validity. In A. Geranpayeh & L. Taylor (Eds.), *Examining listening. Research and practice in assessing second language listening* (pp. 77–151). Cambridge University Press.
- Freedle, R., & Kostin, I. (1999). Does the text matter in a multiple-choice test of comprehension? The case for the construct validity of TOEFL's minitalks. *Language Testing*, 16(1), 2–32. <https://doi.org/10.1177/026553229901600102>
- Gass, S., & Varonis, M. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language Learning*, 34(1), 65–89. <https://doi.org/10.1111/j.1467-1770.1984.tb00996.x>
- Gilmore. (2007). Authentic materials and authenticity in foreign language learning. *Language Teaching*, 40(2), 97–118. <https://doi.org/10.1017/S0261444807004144>
- Griffiths, R. (1992). Speech rate and listening comprehension: Further evidence of the relationship. *TESOL Quarterly*, 26(2), 385–391. <https://doi.org/10.2307/3587015>
- Hulstijn, J. H. (2015). *Language proficiency in native and non-native speakers: Theory and research*. John Benjamins. <https://doi.org/10.1075/llt.41>

- Isaacs, T. (2008). Towards defining a valid assessment criterion of pronunciation proficiency in non-native English-speaking graduate students. *Canadian Modern Language Review*, 64(4), 555–580. <https://doi.org/10.3138/cmlr.64.4.555>
- Kelly, P. (1991). Lexical ignorance: The main obstacle to listening comprehension with advanced foreign language learners. *IRAL*, 29, 135–149.
- Lee, J., & Schallert, D. (1997). The relative contribution of L2 language proficiency and L1 reading ability to L2 reading performance: A test of the threshold hypothesis in an EFL context. *TESOL Quarterly*, 31(4), 713–739. <https://doi.org/10.2307/3587757>
- Levelt, W. J. M. (1983). Monitoring and self-repair in speech. *Cognition*, 14(1), 41–104. [https://doi.org/10.1016/0010-0277\(83\)90026-4](https://doi.org/10.1016/0010-0277(83)90026-4)
- Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. The MIT Press.
- Levelt, W. J. M. (1995). The ability to speak: From intentions to spoken words. *European Review*, 3(1), 13–23. <https://doi.org/10.1017/S1062798700001290>
- Lund, R. (1991). A comparison of second language listening and reading comprehension. *The Modern Language Journal*, 75(2), 196–204. <https://doi.org/10.1111/j.1540-4781.1991.tb05350.x>
- McBride, K. (2011). The effect of rate of speech and distributed practice on the development of listening comprehension. *Computer Assisted Language Learning*, 24(2), 131–154. <https://doi.org/10.1080/09588221.2010.528777>
- McCarthy, M., & Carter, R. (1995). Spoken grammar: What is it and how can we teach it? *ELT Journal*, 49(3), 207–218. <https://doi.org/10.1093/elt/49.3.207>
- Nakatsuhara, F. (2018). Investigating examiner interventions in relation to the listening demands they make on candidates in oral interview tests. In G. Ockey & E. Wagner (Eds.), *Assessing L2 listening: Moving towards authenticity* (pp. 205–226). John Benjamins. <https://doi.org/10.1075/llt.50.14nak>
- Nissan, S., DeVenencenzi, F., & Tang, K. (1996). *An analysis of factors affecting the difficulty of dialogue items in TOEFL listening comprehension* (TOEFL Research Report No. 51). Educational Testing Service.
- Ockey, G. (2007). Construct implications of including still image or video in computer-based listening tests. *Language Testing*, 24(4), 517–537. <https://doi.org/10.1177/0265532207080771>
- Ockey, G. (2018). The degree to which it matters if an oral test task requires listening. In G. Ockey & E. Wagner (Eds.), *Assessing L2 listening: Moving towards authenticity* (pp. 193–204). John Benjamins. <https://doi.org/10.1075/llt.50.13ock>
- Ockey, G., & Wagner, E. (2018). *Assessing L2 listening: Moving towards authenticity*. John Benjamins. <https://doi.org/10.1075/llt.50>
- Paivio, A. (1971). *Imagery and verbal processes*. Holt, Rinehart, and Winston.
- Paivio, A. (1986). *Mental representations: A dual-coding approach*. Oxford University Press.
- Paivio, A. (1991). Dual coding theory: retrospect and current status. *Canadian Journal of Psychology*, 45(3), 255–287. <https://doi.org/10.1037/h0084295>
- Peterson, P. (1991). A synthesis of methods for interactive listening. In M. Celce-Murcia (Ed.), *Teaching English as a second or foreign language* (2nd ed., pp. 106–122). Newbury House.
- Richards, J. C. (1983). Listening comprehension: Approach, design, procedure. *TESOL Quarterly*, 17(2), 219–240. <https://doi.org/10.2307/3586651>
- Rost, M. (2002). *Teaching and researching listening*. Pearson Education.
- Rost, M. (2011). *Teaching and researching listening* (2nd ed.). Pearson.
- Song, M. (2008). Do divisible subskills exist in second language (L2) comprehension? A structural equation modeling approach. *Language Testing*, 25(4), 435–464. <https://doi.org/10.1177/0265532208094272>

- Vandergrift, L. (2006). Second language listening: Listening ability or language proficiency. *The Modern Language Journal*, 90(1), 6–18. <https://doi.org/10.1111/j.1540-4781.2006.00381.x>
- Vandergrift, L. (2007). Recent developments in second and foreign language listening comprehension. *Language Teaching*, 40(3), 191–210. <https://doi.org/10.1017/S0261444807004338>
- Vandergrift, L., & Goh, C. (2012). *Teaching and learning second language listening: Metacognition in action*. Routledge.
- Wagner, E. (2002). Video listening tests: A pilot study. *Working Papers in TESOL & Applied Linguistics, Teachers College, Columbia University*, 2(1). Retrieved on 1 May 2019 from <http://www.tc.edu/tesolalwebjournal>
- Wagner, E. (2004). A construct validation study of the extended listening sections of the ECPE and MELAB. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, 2, 1–26.
- Wagner, E. (2008). Video listening tests: What are they measuring? *Language Assessment Quarterly*, 5(3), 218–243. <https://doi.org/10.1080/15434300802213015>
- Wagner, E. (2010a). Test-takers' interaction with an L2 video listening test. *System*, 38(2), 280–291. <https://doi.org/10.1016/j.system.2010.01.003>
- Wagner, E. (2010b). The effect of the use of video texts on ESL listening test-taker performance. *Language Testing*, 27(4), 493–513. <https://doi.org/10.1177/0265532209355668>
- Wagner, E. (2013). An investigation of how the channel of input and access to test questions affect L2 listening test performance. *Language Assessment Quarterly*, 10(2), 178–195. <https://doi.org/10.1080/15434303.2013.769552>
- Wagner, E. (2014a). Using unscripted spoken texts to prepare L2 learners for real world listening. *TESOL Journal*, 5(2), 288–311. <https://doi.org/10.1002/tesj.120>
- Wagner, E. (2014b). Assessing listening. In A. Kunnan (Ed.), *Companion to language assessment* (Vol. 1, pp. 47–63). Wiley-Blackwell.
- Wagner, E. (2016). Authentic texts in the assessment of L2 listening ability. In J. Banarjee & D. Tsagari (Eds.), *Contemporary second language assessment* (pp. 438–463). Continuum.
- Wagner, E. (2018). A comparison of L2 listening performance on tests with scripted or authenticated spoken texts. In G. Ockey & E. Wagner (Eds.), *Assessing L2 listening: Moving towards authenticity* (pp. 29–44). John Benjamins. <https://doi.org/10.1075/llt.50.03wag>
- Wagner, E., & Wagner, S. (2016). Scripted and unscripted spoken texts used in listening tasks on high stakes tests in China, Japan, and Taiwan. In V. Aryadoust & J. Fox (Eds.), *Current trends in language testing in the Pacific Rim and the Middle East: Policies, analyses, and diagnoses* (pp. 103–123). Cambridge Scholars.
- Wagner, E., & Ockey, G. (2018). An overview of the use of authentic, real-world spoken texts on L2 listening tests. In G. Ockey & E. Wagner (Eds.), *Assessing L2 listening: Moving towards authenticity* (pp. 13–28). John Benjamins. <https://doi.org/10.1075/llt.50.c2>
- Wagner, E. & Toth, P. (2014). Teaching and testing L2 Spanish listening using scripted versus unscripted texts. *Foreign Language Annals*, 47(3), 404–422. <https://doi.org/10.1111/flan.12091>
- Weir, C. (1993). *Understanding and developing language tests*. Prentice Hall.
- Yanagawa, K. (2016). Examining the authenticity of the Center Listening Test: Speech rate, reduced forms, hesitation and fillers, and processing levels. *JACET Journal*, 60, 97–115.



## L2 listening and its correlates

### A meta-analysis

Yo In'nami, Rie Koizumi, Eun Hee Jeon and Yuya Arai  
Chuo University / Seisen University / University of North Carolina  
at Pembroke / Waseda University

Second-language (L2) listening ability is considered to consist of a diverse range of components that collectively help one to understand aural messages (e.g., Vandergrift & Goh, 2012). By building on and expanding previous meta-analyses, this chapter reports on a meta-analysis of L2 listening and its components. Two research questions were addressed: (1) what is the overall relationship between L2 listening and all its components collectively? and (2) what is the relationship between L2 listening and each of its components? Regarding (1), results from 118 studies (533 correlations) showed a significant, moderate relationship between L2 listening and its components overall ( $r = .446$ ). Regarding (2), L2 listening was more strongly related to linguistic knowledge (grammar and vocabulary) than cognitive ability and affective features (aptitude, metacognitive awareness, working memory, attitude, and motivation), with anxiety (another affective feature) located in between. These results suggest how various components relate to L2 listening. The results also support the prediction that generally core components (i.e., linguistic knowledge) were expected to be more strongly correlated with L2 listening than were peripheral components (i.e., cognitive ability).

#### 1. Introduction

Listening comprehension involves a wide range of tasks, including everyday conversation, news, and sustained lectures among others. Nevertheless, listening has been an under-researched area, compared to reading, as the cognitive process of listening cannot be directly observed and listening involves spoken texts that adds more complexity (e.g., phonological modification, accent, and speech rate) than reading (e.g., Andringa, Olsthoorn, van Beuningen, Schoonen, & Hulstijn, 2012; Buck, 2001; Vandergrift & Goh, 2012). Further, the improvement in listening skills



was previously assumed to coincide with increased proficiency (Brown, 2011). This could be still true, but an increasing amount of research is now available into the relationship between second-language (L2) listening and various features/variables/components that influence L2 listening. Now that enough empirical studies have accumulated in the domain, we are in a better position to summarize them. Thus, this chapter reports on a meta-analysis of L2 listening and its correlates.

## 2. Literature review

Many components are considered to influence L2 listening. The components examined in the current study are discussed below, divided into linguistic knowledge (i.e., grammar, vocabulary, phonological awareness, and morphological knowledge), cognitive ability (i.e., aptitude, metacognitive awareness, and working memory), and affective features (i.e., attitude, motivation, and anxiety). They were first derived from the literature (e.g., Buck, 2001; Ockey & Wagner, 2018; Vandergrift & Goh, 2012) and then revised, based on the results of our search of previous studies in L2 listening, as reported in the Method section. Thus, the components reviewed below have been those relatively frequently researched; they were not an exhaustive list of variables important for L2 listening. Further, the components are related to Hulstijn's (2015, 2019) core and peripheral model of language proficiency.

There have been four main meta-analyses dealing with the relationships between L2 listening and its components. For example, Karalík and Merç (2019) included various L2 components, including linguistic (grammar, vocabulary size, and vocabulary depth), cognitive (working memory and metacognitive awareness), and affective factors (anxiety). They showed that linguistic factors were correlated more strongly with L2 listening than cognitive ones. Table 1 summarizes the results of previous meta-analyses from four studies: Karalík & Merç (2019); Zhang & Zhang (2022); Li (2016), and Teimouri, Goetze, & Plonsky (2019). Although each meta-analysis has provided essential summaries of accumulated L2 listening components, each focuses only on certain aspects of linguistic, cognitive, and affective factors. Our meta-analysis includes all these factors and constructs a broader overview of the relationships between these components and L2 listening. Below, we will review these components in relation to L2 listening.

**Table 1.** L2 listening and its correlates from meta-analytic studies

Component	No. of studies <sup>a</sup>	No. of rs	R	95%	CI	I <sup>2</sup>
<i>Linguistic knowledge</i>						
Grammar Knowledge <sup>b</sup>	3	3	.69**	.54	.79	71.560
Vocabulary <sup>c</sup>	44	47	.56**	.50	.62	91.79
Vocabulary Size <sup>b</sup>	16	16	.62**	.55	.68	82.682
Vocabulary Depth <sup>b</sup>	3	3	.80**	.57	.91	92.933
Vocabulary: Recognition <sup>c</sup>	20	20	.49* <sup>f</sup>	.39	.58	–
Vocabulary: Production (Recall) <sup>c</sup>	18	18	.61* <sup>f</sup>	.54	.69	–
Vocabulary: Aural (Auditory) <sup>c</sup>	15	15	.60* <sup>f</sup>	.54	.65	–
Vocabulary: Written (Orthographical) <sup>c</sup>	30	30	.52* <sup>f</sup>	.44	.59	–
<i>Cognitive ability</i>						
Aptitude <sup>d</sup>	4	4	.30* <sup>f</sup>	–	–	–
Aptitude: Phonetic coding <sup>d</sup>	8	8	.12* <sup>f</sup>	–	–	–
Aptitude: Language analytic ability <sup>d</sup>	10	10	.25* <sup>f</sup>	–	–	–
Aptitude: Rote memory <sup>d</sup>	8	8	.21* <sup>f</sup>	–	–	–
Metacognitive awareness <sup>b</sup>	7	7	.54**	.29	.72	93.804
Working Memory <sup>b</sup>	10	10	.30**	.24	.35	29.243
<i>Affective feature</i>						
Anxiety <sup>b</sup>	4	4	–.59**	–.73	–.42	86.678
Anxiety <sup>c</sup>	4	4	–.46* <sup>f</sup>	–.65	–.27	–

*Note*

a. When the number of studies was not reported in the original sources, the number of correlations was reported instead in the table.

b. Adapted from Karalík & Merç (2019, Table 3).

c. Adapted from Zhang & Zhang (in press, Table 4).

d. Adapted from Li (2016, Table 6).

e. Adapted from Teimouri et al. (2019, Table 7).

f. The correlations were statistically significant but their significance level was not reported and was written here as

\*  $p < .05$ ;

\*\*  $p < .01$ .

– = Not reported. Meta-analyses on L2 listening (In'nami & Koizumi, 2009; Montero Perez, van den Noortgate, & Desmet, 2013; Shintani & Wallace, 2014) were not focused on these components in Table 1, and therefore, were not reviewed here.

## 2.1 Linguistic knowledge

### 2.1.1 Grammar

Grammatical knowledge refers to the knowledge of grammatical form (or “linguistic forms on the subsentential, sentential, and suprasentential levels”) and grammatical meaning, including both literal and intended meanings (Purpura, 2004, p. 61). It is considered to influence L2 listening comprehension (e.g., Vandergrift

& Goh, 2012) and various studies have been conducted in this area. Andringa et al. (2012) examined cognitive variables in relation to L2 listening and reported a strong correlation of .77 between grammar and L2 listening. This was the strongest correlation among the variables examined in their study and suggests that grammatical knowledge plays an essential role in processing and understanding aural information in L2. Similarly, strong correlations between grammar and L2 listening have been reported overall (e.g.,  $r = .77$  in Vafaei, 2016; .49 in Geva & Farnia, 2012), which is consistent with Karalík and Merç's (2019) meta-analysis as shown in Table 1 ( $r = .69$ ). This suggests that there is an overall high correlation between L2 listening and grammatical knowledge.

### 2.1.2 Vocabulary

Vocabulary has been intensively examined to show how it is related to L2 proficiency, reading, listening, writing, and speaking (Clenton & Booth, 2021; Qian & Lin, 2020). Zhang and Zhang (2022) have reported a moderate relationship between vocabulary and L2 listening, which demonstrates the importance of vocabulary in comprehending aural input.

Vocabulary has been categorized into multiple subcomponents (e.g., Daller, Milton, & Treffers-Daller, 2007; Nation, 2020). Size and depth of vocabulary are a frequently used categorization (e.g., Schmitt, 2014; Yanagisawa & Webb, 2020). Size refers to how many words learners know in terms of a word form and a primary meaning, whereas depth refers to "how well a learner knows individual words or how well words are organized in the learner's mental lexicon" (Stæhr, 2009, p. 579). According to Nation (2020), depth consists of numerous aspects, such as frequency, association, affix knowledge, and syntactic characteristics. Read (2004) proposed including precision of meaning, comprehensive word knowledge, and network knowledge within the depth framework. Both size and depth can be assessed with various tests, but multiple-choice formats are often used. Stæhr (2009) showed that L2 learning was correlated with size and depth to a similar degree ( $r = .70$  and  $.65$ , respectively). Karalík and Merç (2019) found a slightly weaker correlation of listening with size than with depth in their meta-analysis ( $r = .62$  and  $.80$ , respectively).

Another categorization is the recognition and production of vocabulary, which are also termed receptive and productive vocabulary (Nation, 2020; Schmitt, 2010). Recognition concerns whether learners can comprehend a word and is usually assessed using multiple-choice formats that ask learners to retrieve meaning from word form or to retrieve word form from meaning. Production involves writing or speaking a word, which is typically measured using constructed-response formats that require learners to write a word form or meaning. Zhang and Zhang's (in press) meta-analysis reported that L2 listening was correlated with vocabulary recognition less strongly than with vocabulary production ( $r = .49$  and  $.61$ , respectively).

A further classification is aural and written vocabulary. Aural or spoken vocabulary pertains to the sound or pronunciation of a word, whereas written vocabulary concerns the written form or spelling of a word (Nation, 2020; Schmitt, 2010). In testing aural vocabulary, learners listen to word form or meaning. In assessing written vocabulary, they read a word form or meaning. Although both aural and written vocabulary can be measured using various formats, multiple-choice formats are frequently selected. Zhang and Zhang (2022) conducted a meta-analysis of aural and written dimensions and reported that L2 listening was slightly more strongly associated with aural vocabulary than with written vocabulary ( $r = .60$  and  $.52$ , respectively).

Although all aspects of vocabulary knowledge are essential for effective L2 listening, the question of what is the most important in predicting L2 listening has been a driving force in the research (Milton & Masrai, 2021; Qian & Lin, 2020). While previous meta-analyses have shown the relative strengths of synthesized correlations, it should be noted that confidence intervals (which indicate a degree of precision; Borenstein, 2019) overlap with each other (see Table 1). McLean, Stewart, and Batty (2020) using a bootstrapping method, showed that scores on the production formats when asking learners to write word meaning or word form (i.e., meaning-recall and form-recall) generally tend to be correlated to L2 reading more strongly than were scores on the recognition formats when asking learners to select word meaning (i.e., meaning-recognition). However, they also showed that the distributions of the correlations of vocabulary tests in these three formats to L2 reading overlapped substantially. This suggests that similar correlations between L2 reading and vocabulary are expected, regardless of the format of vocabulary tests. This suggestion could hold true for L2 listening to the extent that reading and listening are considered to involve similar cognitive processes (Rost, 2016). This led us to examine the relative strengths of the correlations for each vocabulary dimension with L2 listening.

Further, although grammar and vocabulary have conventionally been conceptualized separately, such conceptualization has been recently questioned in corpus linguistics. Römer (2017) examined how grammar and vocabulary were used in authentic, spoken, general, and academic English. It was found that phrases, rather than individual grammatical items or words, are an important characteristic of oral language, and recommended considering grammar and vocabulary as a single construct. Based on this statement, it would be interesting to combine grammar and vocabulary into one variable to examine if the results would differ from when they were analyzed separately.

### 2.1.3 *Phonological awareness*

Phonological awareness refers to the ability to understand the sound structures of language (e.g., Vandergrift & Goh, 2012). Phonological awareness allows one to recognize sounds, decompose them into units (e.g., phonemes, syllables, and words), remember these units and predict upcoming units (e.g., Cheung, Chen, Lai, Wong, & Hills, 2001). As these are an important part of comprehension, phonological awareness has been primarily researched in the first language (L1) context. There have been studies in L2, but only a limited number have concerned L2 listening. For example, Andringa et al. (2012) examined a series of variables in relation to L2 listening and found that two aspects of phonological awareness (i.e., segmentation accuracy and segmentation speed) were each correlated with L2 listening strongly and weakly ( $r = .64$  and  $-.36$ , respectively). Similar results were obtained from Taguchi (2008) and Geva and Farnia (2012), which reported a weak, yet significant, relationship between phonemic discrimination ability and L2 listening ( $r = .23$  and  $.36$ , respectively). Taken together, the results from previous studies suggest that phonological awareness is correlated with L2 listening.

### 2.1.4 *Morphological knowledge*

Morphological knowledge refers to the degree to which one understands morphemes such as root words and affixes (e.g., impatiently can be divided into three morphemes: im-, patient, -ly). Such knowledge helps one in better text comprehension and has been extensively examined in reading and literacy studies in L1 and L2. For example, Jeon and Yamashita's (2014) meta-analysis of L2 reading and its correlates reported a strong correlation of .61 (95% confidence interval = .52 to .69) between morphological knowledge and L2 reading. This highlights the importance and the consistent effects of morphological knowledge in L2 reading comprehension.

As part of this line of research on the componentiality of L2 reading, morphological knowledge and L2 listening have been included as predictors. Goodwin, August, and Calderon (2015) showed that morphological knowledge and phonological awareness were related to L2 listening moderately and marginally ( $r = .49$  and  $.20$ , respectively). Taken together, the results from previous studies suggest that morphological knowledge is correlated with L2 listening.

## 2.2 Cognitive ability

### 2.2.1 *Aptitude*

Aptitude is a well-researched area in L2 acquisition. This owes much to John Carroll's pioneering studies in this area and his development of a measure of aptitude – the Modern Language Aptitude Test (MLAT) – with Stanley Sapon (Stansfield & Reed,

2019). This was followed by the development of other tests such as the Pimsleur Language Aptitude Battery (PLAB), the LLAMA, and the High-Level Language Aptitude Battery (Hi-LAB). According to Stansfield and Reed (2019), the MLAT was designed to measure phonetic coding (i.e., sound-symbol association), language analytic ability (i.e., grammar learning), rote memory (i.e., vocabulary learning), and inductive language learning (i.e., number learning). Studies using the MLAT reported mixed results – Harley and Hart (1997) reported significant, moderate correlations between L2 listening and language analytic ability for early immersion students ( $r = .49$ ), whereas Sáfár and Kormos (2008) reported no significant relationship of L2 listening to phonetic coding, language analytic ability, and rote learning ( $r = .04, -.20$ , and  $-.07$ , respectively). Li's (2016) meta-analysis showed a weak or marginal correlation between L2 listening and aptitude overall, phonetic coding, language analytic ability, and rote memory ( $r = .30, .12, .25$ , and  $.21$ , respectively; see Table 1 above). Thus, aptitude and its components seem to be at most weakly related to L2 listening.

### 2.2.2 *Metacognitive awareness*

Although metacognitive awareness has been studied in the fields of education and psychology and shown to influence performance, it has only recently been systematically examined in L2 listening research (Vandergrift, Goh, Mareschal, & Tafaghodtari, 2006). Among early studies, Goh (1997) conducted a diary study of 40 learners of English and reported that learners' metacognitive awareness in L2 listening could be classified into person knowledge, task type awareness, and strategic knowledge. The findings indicate that learners have awareness associated with L2 listening and that they attempt to make full use of their knowledge to understand aural input. Graham (2006) examined learners' self-reported listening strategies through a questionnaire and an interview to investigate what views and beliefs (i.e., metacognitive awareness) they have toward L2 listening. Learners were found to have difficulty in processing aural text due to the speed of text delivery, a factor that made it difficult to identify and understand each word.

Findings from previous studies were woven together to develop a measure of metacognitive awareness in L2 listening research: The Metacognitive Awareness Listening Questionnaire (MALQ; Vandergrift et al., 2006). The MALQ aims to measure directed attention (e.g., focus on the text while avoiding distraction), mental translation (e.g., translate [key] words while listening), person knowledge (e.g., listeners' perceptions toward listening comprehension), planning and evaluation (e.g., set a goal and monitor one's progress in understanding the text), and problem-solving (e.g., try to guess the meaning of a word from context). Studies using the MALQ (e.g., Goh & Hu, 2014; Wallace, 2018) have reported correlations between these aspects of metacognitive awareness and L2 listening. Further, as shown

in Table 1; Karalík and Merç (2019) reported a medium degree of a meta-analyzed correlation between metacognitive awareness and L2 listening ( $r = .54$ ). Taken together, the results from previous studies suggest that metacognitive awareness is moderately correlated with L2 listening.

### 2.2.3 *Working memory*

Working memory has been shown to relate to L1 and L2 processing and performance, as it plays an important role in storing and processing various information, including language (e.g., Baddeley & Logie, 1999). For example, in listening, one needs to remember what has been said to integrate it with the upcoming information. It is important to understand not only what the speaker has said but also why they said something in a particular manner. Such information is then combined with what the speaker is saying right now. Thus, it is essential to put previous and current information together to construct a mental representation of the point made by the speaker, and this requires working memory.

This dual functioning of remembering (or storing, holding, maintaining) information temporarily while processing new information is addressed in working memory, where phonological working memory is considered to be associated with storing audio information and where executive working memory is considered to be involved in processing information. Thus, working memory is essential in understanding L1 and L2 listening, but its association with listening has been reported to be weak. For example, Andersson (2010) examined the relationship between working memory and L2 listening in children and showed that both phonological and executive working memories were weakly related to L2 listening ( $r = .25$  to  $.36$  and  $.37$  to  $.40$ , respectively). Karalík and Merç's (2019) meta-analysis also found a weak correlation between working memory and L2 listening, as shown in Table 1 ( $r = .30$ ).

This was partially supported in Brunfaut and Révész (2015), where neither phonological nor executive working memory was related to L2 listening when overall listening comprehension was tested ( $r$ s not reported). In tasks measuring specific details, both memory types were weakly correlated with L2 listening ( $r = .297$  and  $.312$ , respectively). This seems to suggest that measurement of specific details makes tests challenging, taxing learners' working memory when storing and processing information.

Such a relationship between working memory and L2 listening is weak and becomes statistically nonsignificant when compared to other variables. Andringa et al. (2012) found that phonological and executive working memories were each correlated weakly with L2 listening when observing their raw correlation estimates ( $r = .23$  to  $.25$ ;  $.21$ , respectively). When these correlations were analyzed using structural equation modeling, with phonological and executive working memory



components comprising a latent variable of working memory, they were weakly related to L2 listening ( $r = .32$ ) but did not significantly predict L2 listening in the model where vocabulary knowledge and reasoning ability did predict L2 listening. The results were corroborated in Wallace and Lee (2020). In their study, L2 university students were administered measures of auditory vocabulary size, updating and shifting in executive working memory, and L2 listening. The results showed that the two subcomponents of executive working memory did not predict listening ( $\beta = .022$  to  $.167$ ) and that only vocabulary size predicted listening ( $\beta = .410$ ). The results were the same regardless of whether listening was modeled as a single variable, two variables of short and long texts, or two variables of explicit and implicit items. Taken together, both phonological and executive working memories have been shown to weakly relate to L2 listening, and this is particularly true relative to vocabulary. Further, it remains to be examined how the components of executive working memory – updating and shifting – relate to listening. Although this was investigated in Wallace (2018), our study aimed to extend the Wallace study by undertaking a meta-analysis of relevant studies.

## 2.3 Affective features

### 2.3.1 *Attitude and motivation*

Attitude toward the learning situation is one component of Gardner's socio-educational model of L2 acquisition (Gardner, 2000) and has been measured using the Attitude/Motivation Test Battery (AMTB). In this model, attitude and another component – integrativeness – are considered to influence motivation. Motivation is then considered to influence L2 achievement. In other words, motivation is modeled to directly influence L2 achievement, whereas attitude toward the learning situation and integrativeness are modeled to influence L2 achievement through motivation. As this model, along with the AMTB, has been widely studied, Masgoret and Gardner's (2003) meta-analysis of studies using the AMTB is worth reviewing, although they did not focus on L2 listening. They found that the synthesized correlation between motivation and L2 achievement was slightly higher than that of L2 achievement with attitude toward the learning situation and integrativeness (e.g.,  $r = .29$ ,  $.17$ , and  $.21$  in studies using objective measures). This was consistent with Gardner's (2000) socio-educational model, suggesting that motivation is more strongly related to success in L2 learning than are the other variables.

Although motivation has been studied in L2 acquisition overall, there has not been much research into its relationship with L2 listening. Vandergrift (2005) examined the three components of motivation – amotivation, extrinsic motivation, and intrinsic motivation in relation to L2 listening. Unsurprisingly, amotivation



was negatively and weakly correlated with L2 listening ( $r = -.34$ ). This suggests that students who see no value in understanding messages in L2 are likely to perform poorly in L2 listening tasks. In contrast, extrinsic and intrinsic motivation components were not significantly related to L2 listening ( $r = .16$  and  $.12$ , respectively). Similar results were also reported in Tafaghodtari and Vandergrift (2008), with amotivation negatively and weakly related to L2 listening ( $r = -.26$ ) and extrinsic and intrinsic motivations not related to L2 listening ( $r = .003$  and  $.06$ , respectively). Thus, from previous studies, depending on a component of motivation, there would be negative, weak correlations or no correlations with L2 listening.

### 2.3.2 *Anxiety*

In L2 studies, anxiety has been considered a situation-specific construct, meaning that its potential impact could be limited to certain situations (Horwitz, Horwitz, & Cope, 1986; Horwitz, 2010). There seem to be at least three types of anxiety that are discussed in the literature, which are negatively related to L2 performance. First, students' levels of anxiety in the classroom have been examined. Such anxiety – foreign language classroom anxiety – has been measured using the Foreign Language Classroom Anxiety Scale (FLCAS), and its negative, moderate relationship with L2 listening has been reported (e.g.,  $r = -.53$ , Elkhafaifi, 2005). Second, test anxiety is a perceived fear of failure when taking a test. Winke and Lim (2017) examined the relationship between test anxiety and L2 listening scores and reported a negative, weak correlation between them ( $r = -.267$ ). Similar but slightly lower correlations were reported in In'nami (2006) as well. Third, anxiety in listening context – foreign language listening anxiety – has been recently examined using the Foreign Language Listening Anxiety Scale (FLLAS). Brunfaut and Révész (2015) reported a negative, moderate relationship between foreign language listening anxiety and L2 listening performance ( $r = -.544$ ). Similar results were reported in Elkhafaifi (2005). Note that negative, moderate relationships were observed between three types of anxiety and L2 performance, which supports two previous meta-analyses ( $r = -.59$  in Karalık & Merç, 2019;  $-.46$  in Teimouri et al., 2019). Also note that, among the three types of anxiety, foreign language listening anxiety has been conceptualized and most closely studied in relation to L2 listening.

## 2.4 Core and peripheral components of proficiency

The review so far is related to Hulstijn's (2015, 2019) model of language proficiency. In this model, core components (i.e., linguistic knowledge and speed) are more strongly correlated with performance than are peripheral components (e.g., knowledge of discourse, strategic competences). In our meta-analysis, core components

correspond to linguistic knowledge while peripheral components correspond to cognitive ability. Linguistic knowledge is expected to be more strongly correlated with L2 listening than cognitive ability.

### 3. Research questions

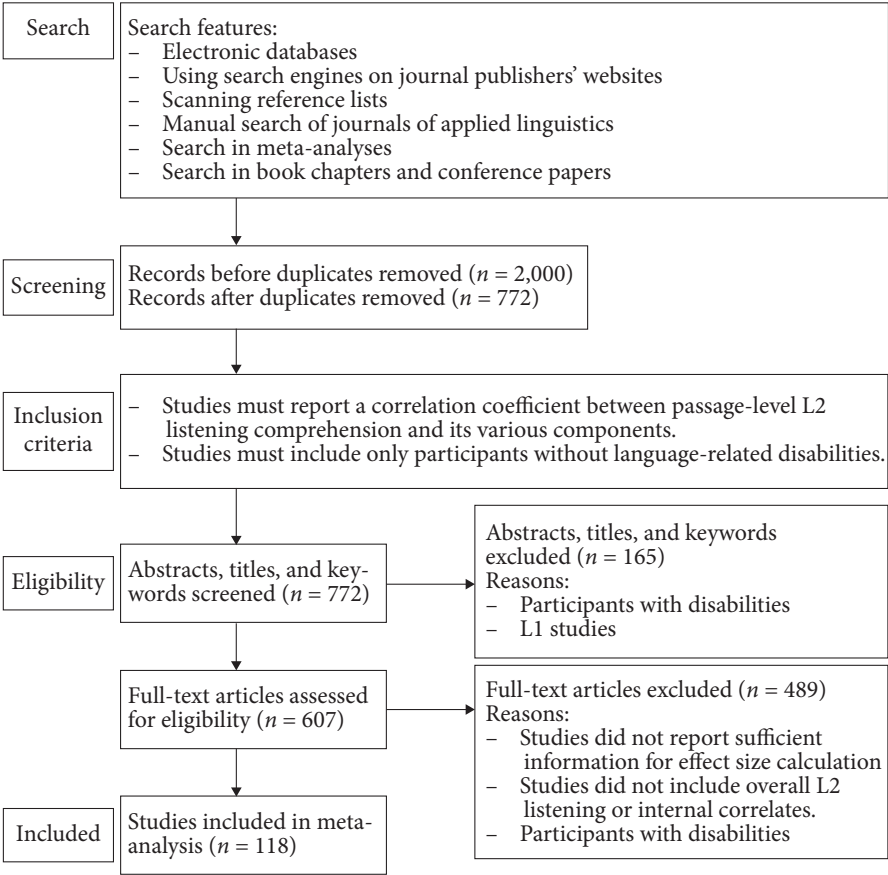
The current meta-analysis built on and expanded previous meta-analyses (see Table 1) but differed in two main points. First, we added four potential correlates to our meta-analysis: phonological awareness, morphological knowledge, attitude, and motivation. Second, we aimed to clarify the relationship between vocabulary, aptitude, metacognitive awareness, working memory, and anxiety, and L2 listening. These components are important to L2 listening and we conducted a detailed analysis on them. Two research questions were addressed. First, what is the overall relationship between L2 listening and all its components collectively? Second, what is the relationship between L2 listening and each of its components?

## 4. Method

### 4.1 Literature search

We searched previous studies extensively between January 2015 and October 2018. It was not possible to search studies at one sitting because searching studies for a study of this magnitude, followed by screening, locating, retrieving, and coding them takes place over an extended period of time. By the time the coding was complete, the previous batch of search results no longer represented the latest developments in the field. To include the most recent studies, we considered it essential to update our search results and conducted searches on a few, separate occasions. Across occasions, the search was conducted in three ways. They are summarized in Figure 1.

First, the search was conducted using databases (Educational Resources Information Center [ERIC], Linguistics and Language Behavior Abstracts [LLBA], MLA International Bibliography, PsycINFO, ScienceDirect, Scopus, and Web of Science) and Google Scholar. Second, the search was conducted using the following journals with search functionality available on their websites: *Annual Review of Applied Linguistics*, *Applied Language Learning*, *Applied Linguistics*, *Applied Psycholinguistics*, *Assessing Writing*, *Canadian Modern Language Review*, *ELT Journal*, *Foreign Language Annals*, *International Journal of Applied Linguistics*, *International Review of Applied Linguistics in Language Teaching*, *JALT Journal*, *Language Assessment*



**Figure 1.** Flow diagram for the literature search and inclusion of studies

*Quarterly, Language Learning, Language Learning & Technology, Language Teaching, Language Teaching Research, Language Testing, Modern Language Journal, RELC Journal, Second Language Research, Studies in Second Language Acquisition, System, and TESOL Quarterly.* The keywords used in the first and second search methods were L2, second language, foreign language, listening, AND XX, where XX was replaced in turn with correlate\*, component\*, subcomponent\*, construct\*, sub-construct\*, correlat\*, oral communica\*, oral skill\*, speak\*, spee\*, spoken, phon\*, pronuncia\*, vocab\*, lexic\*, word, working memory, grammar\*, syntac\*, discourse, speech produc.\*, anxiety, and motivation. These keywords were derived from the keywords and synonyms retrieved from the thesauruses supplied in databases, books, and articles reviewed, as well as authors' experiences. Abstract, title, and keyword searches were conducted; no restrictions on the publication date and language were imposed. When journals were available in print format in addition to

an online format, they were manually searched. Third, recent relevant resources were inspected, including books (e.g., Ockey & Wagner, 2018; see Appendix A), meta-analyses on L2 listening (Karalík & Merç, 2019; Li, 2016; Teimouri et al., 2019; Zhang & Zhang, 2022) and comprehension (Linck, Osthus, Koeth, & Bunting, 2014), Wen's (2016, Table 5.2) list of studies on working memory in relation to skills and its updated version (Wen, 2018), and review papers (Horwitz, 2010; Masgoret & Gardner, 2003). Across these three methods of searching the literature via databases, journal websites, and books, the reference list of each paper and chapter, both published and unpublished, was scrutinized for additional relevant materials.

## 4.2 Study inclusion criteria

To be included in the meta-analysis, a study had to (1) examine the relationship between passage-level L2 listening and its various components using correlations (Pearson or Spearman) and (2) target nonclinical L2 learners. Passage-level L2 listening was typically operationalized as the total score of the listening section of a test. Studies examining sentence-level comprehension only, for example, as seen in Part 1 of the listening section of the TOEFL PBT test (e.g., Man: *Do you mind if I smoke here?* Woman: *Not at all.* Question: *What does the woman mean?* Answer: *She refused.*) were excluded. The Spearman correlations included were four correlations from Brunfaut and Révész (2015), three correlations from Winke and Lim (2017), two correlations from Milton, Wade, and Hopkins (2010), and one correlation each from Harrington and Carey (2009) and Stæhr (2008). Two correlations each in Brunfaut and Révész (2015) and Andringa et al. (2012) were reported as nonsignificant, with no corresponding values; they were replaced with zero. Studies conducted on bilinguals were included if their L2 could be specified.

Studies were excluded if they (1) concerned only reading, speaking, or writing in L1 or L2, vocabulary or grammar in L1 because their role does not seem to be widely discussed in L2 listening models (e.g., Vandergrift & Goh, 2012); (2) concerned integrated-skills tasks because combined use of skills on these tasks could blur the contribution of each skill to the performance; (3) reported on using self-assessment as an indicator of listening proficiency because self-assessment does not necessarily indicate actual proficiency (Ross, 1998); (4) reported on using dictation as a measure of listening proficiency because dictation is considered to measure a broader skill than listening (Buck, 2001); or (5) concerned the relationship between task/item characteristics and task/item difficulty (e.g., Bachman, Davidson, Ryan, & Choi, 1995; Carr, 2006; Freedle & Kostin, 1999; Nissan, DeVincenzi, & Tang, 1996) or the effectiveness of a particular teaching method (e.g., as meta-analyzed in Norris & Ortega, 2000). These exclusion criteria reflected our interest in examining learners' individual difference variables so that the findings can be compared with

Jeon and Yamashita's (2014) reading meta-analysis, the updated version of which is reported in Chapter 3 of this book.

As shown in Figure 1, 772 studies were identified initially. They were examined for all the above criteria by all four authors (with the first author taking the lead), narrowing down the number of studies to 118 (see Appendix B).

### 4.3 Coding

The 118 selected studies (533 correlations) were coded as follows.

#### *Linguistic knowledge*

(a) grammar; (b) vocabulary (size, depth, recognition, production, aural, and/or written); (c) phonological awareness; and (d) morphological knowledge. Concerning (b), vocabulary was coded in terms of size and depth, recognition and production, and aural and written – a distinction that has been widely used in vocabulary research (see Literature Review above).

Concerning (a) and (b), grammar and vocabulary were also coded as a single variable, following the close relationship between them in natural, spoken language (Römer, 2017).

#### *Cognitive ability*

(e) aptitude (phonetic coding, language analytic ability, or rote memory); (f) metacognitive awareness; and (g) working memory (phonological or executive [updating or shifting]).

Regarding (e), aptitude was coded following Li (2016, Table 1) and Stansfield and Reed (2019). Concerning (f), metacognitive awareness in listening research has been often measured using the MALQ (Vandergrift et al., 2006). Studies using MALQ tended to report total scores and section scores sorted by the subcomponents of metacognitive awareness. Both total and section scores were coded. For (g), working memory measures were coded for the measurement of the phonological or executive memory component of working memory (e.g., Wen, 2016). Given the importance of executive memory in cognitive activity, including L2 listening (Wallace & Lee, 2020) as reviewed above, we coded studies measuring executive memory for focus on the updating or shifting function.

#### *Affective features*

(h) attitude; (i) motivation; and (j) anxiety (foreign language listening anxiety, test anxiety, or foreign language classroom anxiety). We explain some of these variables below (see also the Table 1 coding sheet in Appendix C).

As for (j) anxiety and its subcomponents that have been known to have negative relationships with L2 listening (i.e., amotivation, mental translation, and response time), the interpretation of their synthesized correlations across the data, was simplified by converting negative signs that had not been reverse coded in original studies into positive (i.e., the number of correlations was 23, 2, 1, and 5 for anxiety, amotivation, mental translation, and response time, respectively). In our data, all correlations with positive signs indicate a favorable relationship with L2 listening.

### *Other issues*

For longitudinal studies reporting multiple-timepoint data, we coded Time 1 data, as Time 2 or later data could have been influenced by time-varying, extraneous variables that were not the focus of our study. We had three longitudinal studies in our meta-analysis, Droop and Verhoeven (2003), Speciale, Ellis, and Bywater (2004), and Sáfár and Kormos (2008), which provided one, twelve, and five correlations, respectively.

When data were separately reported for higher, lower, and whole groups, the data for the whole groups were coded as they were considered to best represent the study sample. For example, Vogely (1995) reported on the correlation between listening recall scores and metacognitive strategy questionnaire responses from students who enrolled in first-semester ( $r = 0.52$ ,  $n = 25$ ), second-semester ( $r = 0.28$ ,  $n = 40$ ), and third-and-fourth-semester ( $r = 0.42$ ,  $n = 18$ ) Spanish courses. Vogely also reported on the correlation for the whole sample ( $r = 0.44$ ,  $N = 83$  [ $25 + 40 + 18$ ]). We coded the correlation for the whole sample ( $r = 0.44$ ,  $N = 83$ ).

When information was inconsistently reported within a study, we decided to act cautiously and coded less favorable information. This applied to sample size: When different sample sizes were reported in a single study, the smaller sample size was coded.

Finally, instrumental reliability (i.e., Cronbach's alpha and Kuder–Richardson Formula 20) was coded for each measure. If multiple reliability estimates were reported for a single measure (e.g., reliability for working memory span task 1 and another reliability for working memory span task 2), they were averaged. For 14 correlations, multiple reliability estimates were reported.

The inter-coder reliability was examined by having the first and fourth authors code 10 of the 118 studies. The agreement percentage ranged from 95% to 100% across all coded variables. Disagreement was resolved through discussion. The remaining studies were coded by the first author.

#### 4.4 Analyses

We followed the common methodology for meta-analysis of correlations (e.g., Plonsky & Oswald, 2015). Correlation coefficients were Fisher's  $z$ -transformed, aggregated to estimate an average weighted effect size, and the result was converted back to correlations for interpretation. This procedure was repeated for correlations between listening and its components. Please note that the reliability of listening and component measures was reported only in one-third of available correlations – 31% (167/533) and 35% (186/533) – respectively. These percentages were reduced to 25% (134/533) when correlations accompanied by the reporting of both listening and component measures were considered. As only a quarter of correlations was reported with instrument reliability, correlations were not corrected for attenuation.

Effect size aggregation was conducted using a random-effect model. To address dependent effect sizes (e.g., two or more effect sizes from a single study), a robust variance estimator was used with the *robumeta* package in R (Fisher & Tipton, 2015). This method allows researchers to model within-study dependencies and is preferred to averaging dependent effect sizes per study (Tanner-Smith, Tipton, & Polanin, 2016). As a sensitivity analysis, the data were also analyzed using median correlations from each study with the *metafor* package in R (Viechtbauer, 2010).

Throughout the analysis, significance was specified as  $p < .05$ . When confidence intervals associated with respective overall correlation did not overlap with each other, this was interpreted as an indication of a statistically significant difference between the correlations under comparison.  $I^2$  statistics were calculated to examine “what proportion of the variance in observed effects reflects variation in true effects, rather than sampling error” (Borenstein, 2019, p. 115); small  $I^2$  suggests that most of the variance in observed effects (derived from studies included) is due to sampling error; large values of  $I^2$  (e.g.,  $I^2 = 98.765$ ) suggest that a large percentage (e.g., 98.765%) of the variance in observed effects reflects variation in true effects. The minimum number of correlations for component/subcomponent analyses was three, following Li (2016). Thus, some potential components or subcomponents were not synthesized (e.g., integrativeness) because the number of correlations for each component was fewer than three.

Publication bias or a file drawer problem was examined using Orwin's (1983) fail-safe  $N$ , with a criterion effect size of .01 ( $r = .01$ ) and an effect size of 0 for missing values ( $r = .00$ ). The obtained fail-safe  $N$  suggests the number of studies needed to reduce the summary effect to the criterion effect. Fail-safe  $N$ s that reached or exceeded  $5k + 10$ , where  $k$  was the number of studies combined, were considered to indicate the trivial effect of publication bias and support the interpretation of the summary effect (Rosenthal, 1979). Fail-safe  $N$ s were calculated in Microsoft Excel

since the current version of the *robumeta* package does not conduct publication bias analyses. The results were interpreted considering correlations of .25, .40, and .60 as small, medium/moderate, and large effects (Plonsky & Oswald, 2014).

## 5. Results

### 5.1 L2 listening and its overall components

As Table 2 and Figure 2a show, the synthesized correlation between L2 listening and all its components collectively was moderate in size ( $r = .446$  [95% confidence interval: .402, .487],  $p < .001$ ). The  $I^2$  value indicates that 92% of the variance in observed effects reflects variances in the true overall correlations between L2 listening and its components rather than sampling error. It should be noted that the analyses of the variables listed below had a small  $I^2$  value each, indicating that the distribution of effects in studies included is far from the distribution of the true correlations and that it is necessary to be cautious about interpreting these results: motivation; aptitude subcomponents of phonetic coding and rote memory; and metacognitive awareness subcomponents of MALQ directed attention, mental translation, person knowledge, planning and evaluation, and problem-solving.

The overall results showed no evidence of a file drawer problem. For the overall data, the fail-safe  $N$  was 5,762, which indicates that 5,762 studies were needed to reduce the summary effect ( $r = .446$ ) to the criterion effect ( $r = .01$ ). This number exceeded Rosenthal's (1979) criterion ( $5 \times 118 + 10 = 600$ ) and indicates a probably trivial impact of publication bias in the current data. Overall, results from other analyses in terms of the file drawer problem were similar, suggesting the probably trivial impact of publication bias. Note that analyses of aptitude language analytic ability, working memory executive updating, and working memory executive shifting indicated an impact of publication bias. The results from these analyses should, therefore, be interpreted with caution. As a sensitivity analysis, the data were also analyzed using median correlations from each study (Linck et al., 2014). The synthesized correlation between L2 listening and its components was moderate in size ( $r = .441$  [.428, .454],  $p < .001$ ). These values were essentially the same as those above and suggest few impacts of dependent effect sizes.



Table 2. L2 listening and its correlates: Overall, component, and subcomponent analyses

Component	Subcomponent	No. of studies	No. of independent participants	No. of dependent participants	No. of rs	<i>r</i>	95%	CI	Min <i>r</i>	Max <i>r</i>	Fail-safe <i>N</i>	<i>I</i> <sup>2</sup>
Overall	–	118	15268	64630	533	.446**	.402	.487	–.600	.948	5762	91.694
<i>Linguistic knowledge</i>												
Grammar	–	25	3963	4848	35	.517**	.424	.599	.030	.770	1485	91.340
Vocabulary	–	66	9875	18424	182	.562**	.511	.609	–.195	.948	4418	91.020
Grammar-vocabulary	–	69	10320	23831	218	.554**	.504	.600	–.195	.948	4522	91.535
Phonological awareness	–	11	1609	3319	24	.359**	.281	.433	–.020	.640	412	77.280
Morphological knowledge	–	5	514	702	6	.525**	.304	.693	.045	.660	303	73.183
<i>Cognitive ability</i>												
Aptitude	–	10	2868	6601	36	.105*	.023	.185	–.340	.490	96	51.412
Metacognitive awareness	–	24	3457	10122	64	.275**	.194	.353	–.190	.860	662	82.355
Working memory	–	33	3308	14353	123	.297**	.225	.366	–.600	.720	993	79.228
<i>Affective feature</i>												
Attitude	–	3	358	2732	20	.098	–.242	.417	–.320	.704	27	83.254
Motivation	–	4	282	752	17	.106	–.076	.282	–.329	.634	39	32.663
Anxiety	–	14	1668	2777	26	.439*** <sup>d</sup>	.340	.530	–.038	.700	670	80.283
<i>Subset</i>												
Vocabulary	Size	64	9797	17333	170	.550**	.504	.592	–.195	.940	4151	89.848
	Depth	9	827	827	9	.789**	.628	.885	.234	.948	1147	90.493
	Recognition	60	8783	15799	141	.543**	.489	.593	–.195	.948	3820	90.504
	Production	15	2088	3184	41	.595**	.498	.677	.152	.850	1095	86.621
	Aural	25	3949	7272	77	.626**	.564	.680	–.195	.940	1982	90.242
Aptitude	Written	47	6253	11400	118	.526**	.460	.587	–.020	.948	2860	90.041
	Phonetic coding	6	692	734	9	.135*	.031	.236	.040	.447	76	0.000
	Language analytic ability	8	938	2858	16	.055	–.064	.173	–.340	.490	36 <sup>a</sup>	55.925
	Rote memory	7	1222	3009	11	.079	–.044	.199	–.151	.289	48	38.747

Table 2. (continued)

Component	Subcomponent	No. of studies	No. of independent participants	No. of dependent participants	No. of rs	<i>r</i>	95% CI	Min <i>r</i>	Max <i>r</i>	Fail-safe <i>N</i>	<i>I</i> <sup>2</sup>
Metacognitive awareness	MALQ <sup>b</sup> total score	6	1108	1108	6	.226**	.068 .373	.023	.370	133	73.382
	MALQ <sup>b</sup> directed attention	2	339	791	4	.272**	.006 .502	.226	.300	55	0.000
	MALQ <sup>b</sup> mental translation	2	339	791	4	.267 <sup>d</sup>	-.173 .619	.220	.318	53	0.000
	MALQ <sup>b</sup> person knowledge	2	339	791	4	.339*	.255 .419	.269	.397	70	0.000
	MALQ <sup>b</sup> planning and evaluation	3	454	1021	6	.102*	.042 .161	.060	.200	28	0.000
	MALQ <sup>b</sup> problem solving	3	454	906	5	.274**	.249 .298	.260	.289	82	0.000
Working memory	Phonological	24	2630	7953	78	.265**	.179 .347	-.600	.690	636	80.972
	Executive	15	1478	6400	45	.313**	.208 .411	-.170	.720	479	70.366
	Executive updating <sup>c</sup>	1	226	2034	9	.062**	.062 .062	.000	.129	5 <sup>a</sup>	55.537
	Executive shifting <sup>c</sup>	1	226	2034	9	.142**	.142 .142	.077	.202	13 <sup>a</sup>	43.606
Anxiety	Foreign language listening anxiety	9	1433	1526	10	.520** <sup>d</sup>	.414 .612	.340	.700	539	80.415
	Test anxiety	2	142	221	3	.130 <sup>d</sup>	-.923 .954	-.038	.267	24	59.801
	Foreign language classroom anxiety	5	614	614	5	.396** <sup>d</sup>	.152 .594	.143	.530	211	80.367

Note

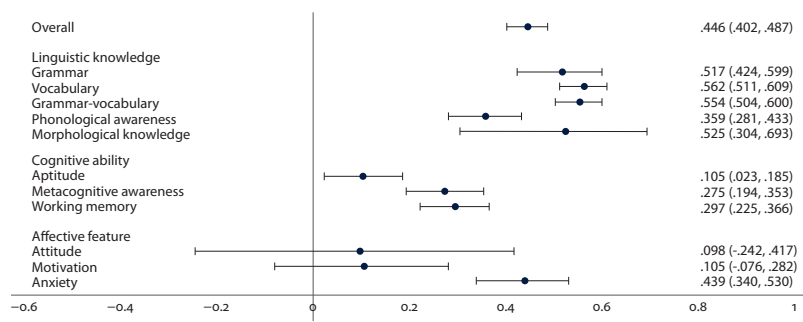
\*  $p < .05$ .\*\*  $p < .01$ .

a. This indicates an impact of publication bias; the results should be interpreted with caution.

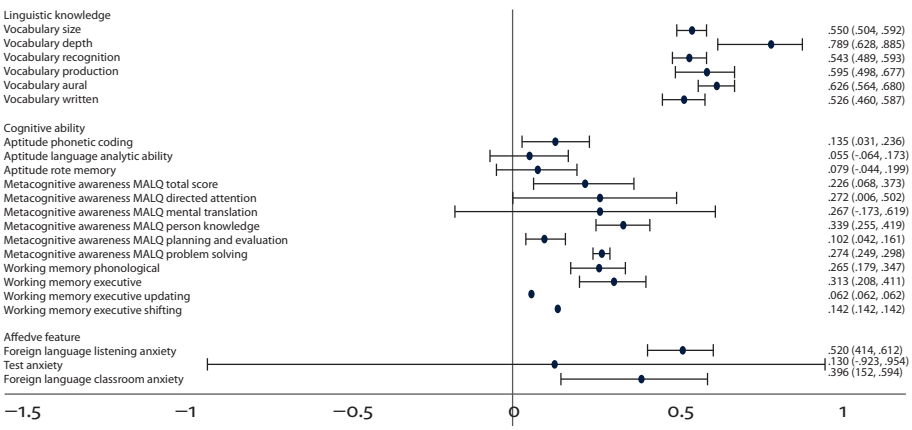
b. Metacognitive Awareness Listening Questionnaire (Vandergrift et al., 2006).

c. The point estimates and confidence intervals of updating and shifting had the same value (e.g., .062 [.062, .062]) because the correlations were similar in size and based on the same group of participants (.000 to .129 and .077 to .202, respectively).

d. The sign of values was reversed from negative to positive for consistency with other measures; thus, a positive correlation indicates that learners with less anxiety, for instance, are more likely to perform well in listening tests (see Coding).



**Figure 2a.** Forest plot of L2 listening and its correlates: Overall, component, and subcomponent analyses



**Figure 2b.** Forest plot of L2 listening and its correlates: Overall, component, and subcomponent analyses

## 5.2 L2 listening and each of its components

A more detailed relationship between L2 listening and each of its components is also shown in Table 2 and Figure 2a. Of particular interest was the finding that linguistic knowledge was overall moderately related to L2 listening ( $r = .562$  [.511, .609] for vocabulary through .359 [.281, .433] for phonological awareness). Although cognitive abilities were also found to be associated with L2 listening, they were less strongly related ( $r = .297$  [.225, .366] for working memory through .105 [.023, .185] for aptitude). As for affective features, anxiety was moderately related to L2 listening ( $r = .439$  [.340, .530]), unlike attitude and motivation.

Further, Table 2 and Figure 2b show results from the subset level analysis. As reported above, vocabulary was most strongly related to L2 listening ( $r = .562$  [.511,

.609]). Overall, this held true for the subset level where vocabulary size and depth were correlated with L2 listening to a similar degree ( $r = .550$  [.504, .592] and .789 [.628, .885], respectively). So were recognition and production vocabulary ( $r = .543$  [.489, .593] and .595 [.498, .677], respectively) and aural and written vocabulary ( $r = .626$  [.564, .680]) and .526 [.460, .587], respectively). However, the confidence intervals between size and depth did not overlap with each other, with size correlated less strongly with L2 listening than depth. The confidence intervals between recognition and production vocabulary and between aural and written vocabulary overlapped with each other.

Regarding aptitude, phonetic coding was significantly related to L2 listening ( $r = .135$  [.031, .236]) whereas language analytic ability and rote memory were not. Thus, among aptitude components, the phonetic coding component of aptitude was the most responsible for the relationship between aptitude and listening.

Metacognitive awareness, as often operationalized using MALQ scores, was associated with listening ( $r = .226$  [.068, .373]). The components of metacognitive awareness in MALQ – directed attention, person knowledge, planning and evaluation, and problem-solving – were significantly related to listening (e.g.,  $r = .272$  [.006, .502] for directed attention) whereas mental translation did not show its significant relationship with listening.

As for working memory, its phonological and executive components were significantly related to listening ( $r = .265$  [.179, .347] and .313 [.208, .411], respectively). Both updating and shifting functions of the executive component also significantly contributed to listening to a marginal degree ( $r = .062$  [.062, .062] and .142 [.142, .142], respectively).

Finally, foreign language listening anxiety and foreign language classroom anxiety were significantly related to listening ( $r = .520$  [.414, .612] and .396 [.152, .594], respectively), which contrasted with test anxiety, showing no significant relationship with listening. This suggests that different roles are played by different types of anxiety.

## 6. Discussion

This meta-analysis examined two research questions. Concerning the relationship between L2 listening and its components overall (Research Question 1), the synthesized correlation was moderate ( $r = .446$  [.402, .487]). This suggests a moderate relationship of L2 listening with its components in general. Nevertheless, note that, although the confidence interval of the synthesized correlation was narrow, the correlations ranged from  $-.600$  to  $.948$ . This suggests varying degrees of the relationships between L2 listening and its components, revealing the multidimensional nature of the L2 listening construct. In this respect, the results supported

the L2 listening literature: Successful listening comprehension requires adequate processing of a flow of incoming messages online by using linguistic knowledge, and cognitive and affective skills (e.g., Graham, 2006; Vandergrift & Goh, 2012). As an overall correlation like this was not reported in Karalík and Merç (2019) – the closest listening meta-analysis to our study – it was not possible to compare our findings with theirs. Previous meta-analyses were summarized at the component level in Table 1 and were compared with ours as discussed in Research Question 2.

Concerning the relationship between L2 listening and each of its components (Research Question 2), the results – both statistically significant and nonsignificant – are discussed below.

## 6.1 Linguistic knowledge

### 6.1.1 Grammar

Grammatical knowledge was moderately related to L2 listening ( $r = .517$ ), as expected from previous studies (e.g.,  $r = .69$  in Karalík & Merç, 2019). The strength of the correlations was similar to the correlation of L2 listening with vocabulary, phonological awareness, and morphological awareness, as the confidence intervals of the correlation with grammar and the ones with the remaining three components overlapped. The results suggest that grammatical knowledge itself is a strong correlate of L2 listening and that this is also true relative to other linguistic knowledge variables.

### 6.1.2 Vocabulary

Vocabulary overall was most strongly related to L2 listening ( $r = .558$ ). This was expected from and consistent with previous studies (e.g.,  $r = .56$  in Zhang & Zhang, 2022), suggesting that vocabulary plays an important role in comprehending aural information. It should be noted that the synthesized correlation with vocabulary was the highest but only moderate, which indicates that the use of a single component – even the strongest correlate, namely vocabulary – can only predict a limited portion of the variance of L2 listening. This suggests the need to measure multiple components of L2 listening to accurately predict and explain it.

When vocabulary was classified into different subdimensions, confidence intervals between size and depth did not overlap with each other; those between recognition and production vocabulary and between aural and written vocabulary overlapped, suggesting a similar range of relationships. The following section discusses each relationship.

First, vocabulary size was correlated with L2 listening less strongly than depth ( $r = .550$  and  $.789$ , respectively), and the difference of strengths of correlations was significant. The result was consistent with Karalík and Merç (2019;  $r = .62$  and  $.80$ ,

respectively) although their confidence intervals overlapped. This difference seems to be explained by the different correlations included in the meta-analyses: 170 and 9 in the current study, and 16 and 3 in Karalík and Merç (2019). The current results – vocabulary depth correlated more strongly with size – may suggest that having a deeper vocabulary knowledge (i.e., having a wider range of knowledge other than primary meaning) can better predict L2 listening. This may be because listening requires a smaller vocabulary size than reading, particularly in everyday conversation where people tend to speak in a simpler language. However, the lack of opportunity to listen back and forth (in contrast with reading) requires learners to use a variety of elements of vocabulary knowledge such as associations and collocations to comprehend listening texts and make effective inferences from context regarding word meaning using information of surrounding words (Milton & Masrai, 2021).

However, one caveat is that in the current meta-analysis, size and depth are based on a different number of correlations and different test characteristics: Size has 170 correlations, and depth has only 9 studies. The large difference in the number of studies investigating size and depth is a reflection on vocabulary studies in general (Koizumi & In'nami, 2020; Qian & Lin, 2020; Yanagisawa & Webb, 2020) – size has been extensively targeted because size is easier to define and assess given the availability of multiple vocabulary size tests (Qian & Lin, 2020) and probably because the form–meaning link, the core construct of size, is considered central in L2 learning and use (Schmitt, 2010; Webb, 2005). Further, 78% (7/9) of studies on depth used the Word Associates Test (Read, 1993) or its slightly modified version. In contrast, studies that examined size used a variety of tests, including widely used tests (e.g., Vocabulary Size Test; Nation & Beglar, 2007) and author-made tests (see Appendix C for details). As a result, the test quality and resultant constructs of size tests seem to vary substantially. Further meta-analyses are needed to better understand the relationship between size, depth, and L2 listening.

Second, vocabulary recognition was less strongly related to L2 listening than vocabulary production ( $r = .543$  and  $.595$ , respectively). This trend aligned well with Zhang and Zhang (2022;  $r = .49$  and  $.61$ ), although the difference between recognition and production in the current meta-analysis was much smaller. However, in both meta-analyses, confidence intervals of the two variables overlapped. These results indicate that the differences in vocabulary elicited using recognition and production formats are trivial, which is also indicated by previous studies, given that the distinction of recognition and production in terms of relationships with L2 listening has not received much focus (see Milton & Masrai, 2021; Qian & Lin, 2020).

Third, aural vocabulary was more strongly related to L2 listening than written vocabulary in our meta-analysis ( $r = .626$  and  $.526$ , respectively) as it was in Zhang and Zhang's meta-analysis (2022;  $r = .60$  and  $.52$ ). As listening requires aural comprehension, it is reasonable to observe a stronger relationship between aural

vocabulary and L2 listening. However, the confidence intervals of aural vocabulary and written vocabulary overlapped with each other in both meta-analyses. This can be explained by numerous other factors that are considered to affect L2 listening, such as L1 and L2 proficiency levels, vocabulary frequency levels, the balance of aural and written vocabulary in the mental lexicon, and test formats (Milton & Masrai, 2021; Qian & Lin, 2020). For example, Milton and Masrai (2021) argue that both aural and written vocabulary is necessary when answering L2 listening questions, with stems and options often provided in written form. Influenced by such factors, strengths of correlations between L2 listening and aural and written vocabulary may vary.

When grammar and vocabulary were analyzed as one construct, their relationship with L2 listening was similar to when they were analyzed separately. This suggests that the role of grammar and vocabulary in L2 listening is similar overall and that they could be considered a single construct (Römer, 2017) although further discussion is needed to support such conceptualization.

### 6.1.3 *Phonological awareness*

Phonological awareness was weakly related to L2 listening ( $r = .359$ ). Phonological awareness was not examined in previous meta-analyses but its weak relationship with L2 listening in our meta-analysis was consistent with previous primary studies (e.g.,  $r = .36$  in Geva & Farnia, 2012). The consistent relationship between phonological awareness and L2 listening reinforces the view that one's sensitivity to analyzing audio information into units contributes to the understanding of such information (Cheung et al., 2001; Wagner, Torgesen, & Rashotte, 1994). However, the phonological awareness-L2 listening relationship was overall weaker than the relationship between grammar, vocabulary, morphological knowledge, and L2 listening. This may be because acquiring phonological awareness is a necessary but not a sufficient condition for L2 listening, as it needs to be accompanied by knowledge about the syntactic structure of language and lexis to process audio input. Thus, phonological awareness plays an important yet minor role in L2 listening comprehension, compared with grammar, vocabulary, and morphological knowledge.

### 6.1.4 *Morphological knowledge*

Morphological knowledge was moderately related to L2 listening ( $r = .525$ ), which was not examined in previous meta-analyses but expected from previous primary studies (e.g.,  $r = .49$  in Goodwin et al., 2015). Again, comparable to the results for phonological awareness in relation to L2 listening, morphological knowledge is useful for inferring the meaning of new words. For instance, understanding inflectional morphemes by adding a suffix to a root word while maintaining the meaning of the

word (e.g., book to books) and derivational morphemes by adding a suffix to a root word while modifying the meaning of the word (e.g., friend to friendship) is an important aspect of comprehension. Furthermore, listening passages for adolescents and adults probably contain morphologically complex words, and this likely results in a relatively strong relationship of L2 listening to morphological knowledge and also to vocabulary, especially vocabulary depth, because morphological knowledge is part of vocabulary knowledge (Nation, 2020). The moderate relationship between L2 listening and morphological knowledge may partially explain why vocabulary depth was strongly related to L2 listening.

## 6.2 Cognitive Ability

### 6.2.1 *Aptitude*

Aptitude was the weakest correlate of the three correlates of cognitive ability ( $r = .105$ ). This marginal relationship seems to arise from phonetic coding, which was the only subcomponent of aptitude that was significantly related to L2 listening ( $r = .135$ ): language analytic ability and rote memory were not significantly related ( $r = .055$  and  $.079$ , respectively). Thus, the phonetic coding component of aptitude was the most responsible for the relationship between aptitude and L2 listening, suggesting that the ability to understand the relationship between a speech sound and written symbols is essential but of minor importance for successful listening. In contrast, the lack of association of language analytic ability and rote memory with L2 listening in the current study may appear puzzling. These subcomponents are considered to represent the ability to learn grammar and vocabulary, respectively. As grammar and vocabulary were found to relate to L2 listening in the current and previous studies (e.g., Karalik & Merç, 2019 for grammar and vocabulary; Wallace, 2018; Zhang & Zhang, 2022, for vocabulary), it is possible that language analytic ability and rote memory are constructs different from grammar and vocabulary, respectively. In fact, Li (2016) reported that language analytic ability was only weakly related to L2 grammar ( $r = .39$ ,  $p < .05$ ) and that rote memory was correlated negligibly with L2 vocabulary ( $r = .20$ ,  $ns$ ). This seems to explain the seemingly contradictory results of language analytic ability and rote memory vs. grammar and vocabulary, in relation to relationships with L2 listening.

The correlations evident in the current study were weaker than those found by Li (2016) except for language analytic ability (.105 and .30 in overall aptitude; .135 and .12 in phonetic coding; .055 and .25 in language analytic ability; and .079 and .21 in rote memory, respectively). Although language analytic ability was subject to potential publication bias in the current meta-analysis, the same seems to be true in Li (2016) because he had fewer correlations (i.e., 16 and 10). Different results seem to have been derived from a difference in the search and a longer search



period: Li (2016) did not include unpublished reports and searched studies up to May 2014; the current study included unpublished reports (which provided 8 additional correlations) and searched studies up to October 2018 (which provided 1 additional correlation).

### 6.2.2 *Metacognitive awareness*

Metacognitive awareness was significantly related to L2 listening, which is consistent with Karalík and Merç (2019). One point upon which the current study and Karalík and Merç (2019) seem to differ is the degree of correlation between metacognitive awareness and L2 listening, which was weaker ( $r = .275$  [.194, .353]) in the former than in the latter ( $r = .54$  [.29, .72]). This seems to have been caused by the number of studies included (i.e., 24 and 7, respectively). However, confidence intervals in both results overlapped, and both studies confirm the finding that metacognitive strategies play an important role in addressing the cognitive processes involved in L2 listening. A more detailed picture emerged from the results of studies using the MALQ. MALQ total scores and its subsections were significantly related to L2 listening ( $r = .102$  to  $.339$ ) except for mental translation ( $r = .267$ ), which had a non-significant correlation. This suggests that directed attention, person knowledge, planning and evaluation, and problem-solving strategies contribute to L2 listening success: Good listeners tend to direct focused attention to the listening text, feel less difficulty and anxiety about listening, set a goal to comprehend the text better, monitor their understanding, and solve problems that they encounter while listening, for example, by guessing the meaning of a word from context.

### 6.2.3 *Working memory*

Working memory was weakly related to L2 listening ( $r = .297$ ). This closely aligned with Karalík and Merç (2019;  $r = .30$ ) and Linck et al. (2014;  $r = .242$ , between L2 comprehension [including listening] and working memory). These consistent results across the three meta-analyses, including the current study, suggest that the relationship between L2 listening and working memory is consistently weak.

Further, both phonological and executive working memories were significantly and weakly related to L2 listening to a similar degree ( $r = .265$  and  $.313$ , respectively). The results appear inconsistent with literature on working memory, where executive working memory has been shown to be more closely related to the processing and comprehension of written language than phonological working memory (e.g., Wen, 2016). Yet, the results could be explained by the characteristics of spoken texts. For example, Buck (2001) argues that, in listening, the text is provided aurally (not written), can become easier or more difficult to understand by phonological modification (e.g., assimilation), prosodic features (e.g., intonation), accent,

speech rate, and hesitation (e.g., pauses; for a similar discussion, see Wagner's chapter in this book). These characteristics are outside the learners' control and may heavily tax phonological working memory (particularly phonological short-term storage). Further, to match multiple-choice questions with aural input can become cognitively demanding while trying to understand aural text. This makes it important for learners to recognize and decode questions (and stems) automatically, so that comprehension is enhanced and more resources can be directed to executive working memory to update, switch, and inhibit the information (e.g., Wen, 2016). Regardless of passage/sentence length and difficulty and the cognitive demand of the task on learners, they need to subvocally rehearse and remember the content of a passage/sentence while engaging in the task. Given these characteristics of spoken texts and questions, articulatory rehearsal is essential in listening. As articulatory rehearsal is involved in both phonological and executive working memories, it could be possible that they were related to L2 listening to a similar degree.

Finally, updating and shifting were significantly related to L2 listening ( $r = .062$  and  $.142$ , respectively), but the size of the correlations was very small. This suggests that the functions of executive working memory – updating information and shifting/switching attention between tasks while inhibiting/rejecting irrelevant information (Miyake et al., 2000) – are not much associated with L2 listening. However, caution should be exercised when interpreting the finding: All nine correlations taken from Wallace (2018), were flagged as being subject to publication bias. Wallace (2018) allowed learners to take notes and look at questions and options through the test session, which could have made listening less cognitively demanding (Wallace & Lee, 2020). Further, citing a study reporting that executive working memory is not fully developed until adulthood, Wallace and Lee (2020) stated that this could have been the case for Wallace's (2018) participants – they were all teenagers. These factors could have lowered the relationship between L2 listening, updating, and shifting.

### 6.3 Affective features

#### 6.3.1 *Attitude and motivation*

Attitude and motivation were not related to L2 listening ( $r = .098$  and  $.106$ ). This was consistent with previous studies, where such correlations were weak at best (e.g.,  $r = -.34$  to  $.16$  in Vandergrift, 2008). According to Vandergrift (2008), depending on a component of motivation and listening task types, there would be negative or no correlation with L2 listening. However, note that due to the lack of studies, we were not able to classify correlations into those measuring the relationship of L2 listening with amotivation, extrinsic motivation, or intrinsic motivation, with

L2 learning stratified by various listening task types. Therefore, our analysis was limited in that it only allowed us to examine a combined relationship of multiple components of motivation with L2 listening. We believe, however, that the recent development of motivation theories (e.g., Papi & Hiver, 2020) will prompt studies investigating relationships among attitude, motivation, and L2 learning and will lead to more studies being added to meta-analyses.

### 6.3.2 *Anxiety*

Anxiety was overall significantly and moderately related to L2 listening ( $r = .439$ ), suggesting that less anxious students are likely to perform better in listening tasks than those who feel more anxious. This was consistent with previous meta-analyses that reported moderate relationships ( $r = -.59$  in Karalík & Merç, 2019;  $-.46$  Teimouri et al., 2019).

Among the three types of anxiety, test anxiety was not significantly related to L2 listening ( $r = .130$ , which was reverse coded). However, we would like to note that this result was based on three correlations from two studies (In'nami, 2006; Winke & Lim, 2017) and should only be considered provisional. Further, the nonsignificant results may be attributed to two reasons. First, because these two studies were conducted in relatively low-stakes settings, participants may not have been nervous during the test. Second, test anxiety is anxiety about test-taking in general, as demonstrated in questionnaire items such as “During tests, I find myself thinking of the consequences of failing” and “Were you nervous at the test?” (Winke & Lim, 2017, p. 397). These items are designed to elicit a wide range of experiences students may have about test-taking and may not be limited to listening. The relationship between test anxiety and L2 listening may be detected if anxiety is defined and measured specifically in relation to L2 listening.

In contrast with test anxiety, foreign language listening anxiety, and foreign language classroom anxiety were both significantly related to L2 listening ( $r = .520$  and  $.396$ , respectively). The correlation of L2 listening with foreign language listening anxiety was slightly higher than the one with foreign language classroom anxiety, which supports the findings of Elkhafaifi (2005;  $r = -.70$  and  $-.53$ , respectively). This could be explained by situation specificity. Foreign language classroom anxiety is the type of anxiety that students may experience in the classroom. While listening may often be related to tasks and activities, all tasks or activities are not specific to listening. Instead, foreign language listening anxiety was a more situation-specific construct, with particular focus on the anxiety students may experience in L2 listening. This could be observed in items in the FLLAS, such as “It bothers me to encounter words I can't pronounce while listening to Arabic” and “By the time you get past the strange sounds in Arabic, it's hard to remember what you're listening to” (Elkhafaifi, 2005, p. 209). As indicated by these examples, foreign language listening

anxiety seems to be construct- sensitive to context. Its association with L2 listening could be observed if it was defined and measured specifically in relation to L2 listening (see also Pae [2013] for similar discussion).

#### 6.4 Relationship between core and peripheral components of proficiency

We found that the core and peripheral components of proficiency, grammar, vocabulary, and grammar-vocabulary (i.e., core components; Hulstijn, 2015, 2019) had non-overlapping confidence intervals with those of aptitude, metacognitive awareness, and working memory (i.e., peripheral components; Hulstijn, 2015, 2019). As we discussed above, core components were expected to be more strongly correlated with L2 listening than peripheral components. This was overall supported by our findings. This suggests that grammar and vocabulary are the core components that constitute language proficiency than peripheral components. It should be noted that there were some exceptions in core components, with confidence intervals of phonological awareness and morphological knowledge overlapping with all peripheral components except for aptitude. Thus, it should be considered that relationships between core and peripheral components of proficiency are generally supported in L2 listening, with a few exceptions.

#### 6.5 Reliability issues

The reliability of listening and its component measures was reported in one-third of correlations – 31% (167/533) and 35% (186/533) – respectively. These percentages were on the lower end of what Plonsky and Derrick (2016) reported concerning the frequency with which reliability was reported in various L2 subdomains (6%–64%). This suggests that, compared with other L2 fields, reliability was less often reported in our data. More importantly, the percentage of correlations accompanied by the reporting of both listening and component measures was 25% (134/533), meaning that only a quarter of the correlations in our data could be corrected for the attenuation of measurement error. As correlations are lowered with more error, if such error had been considered, the synthesized correlations from the current study would have been higher than those presented.

Further, reliability estimates in our data resembled Plonsky and Derrick's (2016) findings. As shown in Table 3, median values were overall slightly higher in Plonsky and Derrick than those in our study. However, for each skill or awareness, interquartile ranges overlapped between Plonsky and Derrick and our study. In sum, in our data, reliability estimates were infrequently reported; when they were reported, the values were comparable to those in the field.

**Table 3.** Comparison of instrument reliability across studies and skills

	Plonsky and Derrick (2016) <sup>a</sup>			Current study		
	N of <i>r</i>	Median	Interquartile range	N of <i>r</i>	Median	Interquartile range
Listening	38	.77	.17	167	.74	.15
Grammar	124	.82	.16	15	.74	.081
Vocabulary	150	.83	.14	45	.81	.142
Phonological awareness	34	.83	.16	17	.78	.13

*Note*

a. Adopted from part of Table 5 in their study.

7. Conclusion

This study examined the relationship between L2 listening and its components using meta-analysis. For Research Question 1, concerning the overall relationship between L2 listening and its components, the synthesized correlation was moderate. For Research Question 2, regarding the relationship between L2 listening and each of its components, we found that most linguistic variables (i.e., grammar, vocabulary, and morphological knowledge) were moderately related to L2 listening, although phonological awareness correlated weakly. Most cognitive variables (i.e., metacognitive awareness and working memory) were weakly related to L2 listening, with aptitude correlating with L2 listening marginally. Most affective variables (i.e., attitude and motivation) were not significantly related to L2 listening, with anxiety correlating with L2 listening moderately. Further, we found that subset components of language aptitude correlated with L2 listening strongly, moderately, weakly, or very weakly. The exceptions were language analytic ability and rote memory in aptitude, the use of mental translation in metacognitive awareness, and test anxiety, which were not significantly related to L2 listening, unlike other components of these constructs. The results will help researchers to refine their understanding of the construct of L2 listening in relation to its various components.

Along with theoretical implications related to L2 listening and Hulstijn’s (2015, 2019) model, the current results would help teachers, learners, and test developers be aware of potentially affecting factors (e.g., vocabulary, metacognitive awareness, and anxiety) when learners do not do well on listening tests to enhance their practices. For example, teachers can provide assistance to overcome listening difficulty. Test developers can consider these factors to create better test tasks and provide an environment where test takers can show their ability to the fullest.

We discuss three limitations of this study. The results suggest that three sub-components (i.e., language analytic ability, and updating and shifting functions of the executive working memory) suffered from publication bias. While the need for an addition of primary studies in future meta-analysis is applicable to all components/subcomponents, this is especially needed in the case of these three subcomponents with publication bias (and also the subcomponents with small  $I^2$  values; see L2 Listening and its Components Overall). For example, the two components of executive working memory – updating and shifting – were based on data from the same participants in a single study (i.e., Wallace, 2018). This may have led to a potential impact of publication bias on the relationship between these constructs. Future studies are warranted to better understand how the components of executive working memory relate to L2 listening and how such a relationship compares with other linguistic, cognitive, and affective variables.

Second, the current study has focused on the relationship between L2 listening and its components. Although this focus has clarified how each component was independently related to L2 listening, it remains unclear how the components were related to each other (e.g., metacognitive awareness and vocabulary) and how components were jointly related to L2 listening. These questions can be addressed using meta-analytic structural equation modeling (Jak, 2015); correlations among all sets of variables are calculated and used to examine direct and indirect relationships between L2 listening and its components. This would allow for the examination of an indirect effect of metacognitive awareness and working memory (particularly updating and shifting) on L2 listening through vocabulary, for instance. Such an effect was reported in a primary study (Wallace, 2022) and can be further examined by modeling grammar and vocabulary as moderating the relationship between cognitive variables and L2 listening, using a meta-analytic approach.

Finally, this study would benefit from the consideration of measurement error when modeling the relationship between listening and its components. This requires instrument reliability to be routinely reported in primary studies to allow for the correction for attenuation of measurement error in meta-analysis. If the results from the current study were corrected for attenuation effects, they could yield stronger correlations between listening and its components. Muchinsky (1996) argues that the corrected correlations could represent the overestimates of effects whereas the uncorrected correlations could represent the underestimates of effects. Zhang and Zhang (2022) state that the true effect may be located between these two estimates. By comparing the corrected and uncorrected synthesized correlations between listening and its components, we may be better able to understand how listening is related to its various components and how such a relationship is moderated by variables, for example, vocabulary size and depth.

## References

- Andersson, U. (2010). The contribution of working memory capacity to foreign language comprehension in children. *Memory*, 18(4), 458–472. <https://doi.org/10.1080/09658211003762084>
- Andringa, S., Olsthoorn, N., van Beuningen, C., Schoonen, R., & Hulstijn, J. (2012). Determinants of success in native and nonnative listening comprehension: An individual differences approach. *Language Learning*, 62(Suppl. 2), 49–78. <https://doi.org/10.1111/j.1467-9922.2012.00706.x>
- Bachman, L. F., Davidson, F., Ryan, K., & Choi, I.-C. (1995). *An investigation into the comparability of two tests of English as a foreign language: The Cambridge-TOEFL comparability study*. University of Cambridge Local Examinations Syndicate.
- Baddeley, A. D., & Logie, R. H. (1999). Working memory: The multiple-component model. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 28–61). Cambridge University Press. <https://doi.org/10.1017/CBO9781139174909.005>
- Borenstein, M. (2019). *Common mistakes in meta-analysis and how to avoid them*. Biostat.
- Brown, S. (2011). *Listening myths: Applying second language research to classroom teaching*. University of Michigan Press. <https://doi.org/10.3998/mpub.2132445>
- Brunfaut, T., & Révész, A. (2015). The role of task and listener characteristics in second language listening. *TESOL Quarterly*, 49(1), 141–168. <https://doi.org/10.1002/tesq.168>
- Buck, G. (2001). *Assessing listening*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511732959>
- Carr, N. T. (2006). The factor structure of test task characteristics and examinee performance. *Language Testing*, 23(3), 269–289. <https://doi.org/10.1191/0265532206lt3280a>
- Cheung, H., Chen, H. C., Lai, C. Y., Wong, O. C., & Hills, M. (2001). The development of phonological awareness: effects of spoken language experience and orthography. *Cognition*, 81(3), 227–241. [https://doi.org/10.1016/S0010-0277\(01\)00136-6](https://doi.org/10.1016/S0010-0277(01)00136-6)
- Clenton, J., & Booth, P. (Eds.). (2021). *Vocabulary and the four skills: Pedagogy, practice, and implications for teaching vocabulary*. Routledge.
- Daller, H., Milton, J., & Treffers-Daller, J. (2007). Editors' introduction: Conventions, terminology and an overview of the book. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modeling and assessing vocabulary knowledge* (pp. 1–32). Cambridge University Press. <https://doi.org/10.1017/CBO9780511667268.003>
- Droop, M., & Verhoeven, L. (2003). Language proficiency and reading ability in first- and second-language learners. *Reading Research Quarterly*, 38(1), 78–103. <https://doi.org/10.1598/RRQ.38.1.4>
- Elkhafaifi, H. (2005). Listening comprehension and anxiety in the Arabic language classroom. *The Modern Language Journal*, 89(2), 206–220. <https://doi.org/10.1111/j.1540-4781.2005.00275.x>
- Fisher, Z., & Tipton, E. (2015). Robust variance meta-regression (Version 2.0) [Software]. Retrieved from <http://cran.r-project.org/web/packages/robumeta/robumeta.pdf>
- Freedle, R., & Kostin, I. (1999). Does the text matter in a multiple-choice test of comprehension? The case for the construct validity of TOEFL's minitalks. *Language Testing*, 16(1), 2–32. <https://doi.org/10.1177/026553229901600102>
- Gardner, R. C. (2000). Correlation, causation, motivation, and second language acquisition. *Canadian Psychology*, 41(1), 10–24. <https://doi.org/10.1037/h0086854>
- Geva, E., & Farnia, F. (2012). Developmental changes in the nature of language proficiency and reading fluency paint a more complex view of reading comprehension in ELL and EL1. *Reading & Writing*, 25(8), 1819–1845. <https://doi.org/10.1007/s11145-011-9333-8>



- Goh, C. (1997). Metacognitive awareness and second language listeners. *ELT Journal*, 51(4), 361–369. <https://doi.org/10.1093/elt/51.4.361>
- Goh, C. C. M., & Hu, G. W. (2014). Exploring the relationship between metacognitive awareness and listening performance with questionnaire data. *Language Awareness*, 23(3), 255–274. <https://doi.org/10.1080/09658416.2013.769558>
- Goodwin, A. P., August, D., & Calderon, M. (2015). Reading in multiple orthographies: Differences and similarities in reading in Spanish and English for English learners. *Language Learning*, 65(3), 596–630. <https://doi.org/10.1111/lang.12127>
- Graham, S. (2006). Listening comprehension: The learners' perspective. *System*, 34(2), 165–182. <https://doi.org/10.1016/j.system.2005.11.001>
- Harley, B., & Hart, D. (1997). Language aptitude and second language proficiency in classroom learners of different starting ages. *Studies in Second Language Acquisition*, 19(3), 379–400. <https://doi.org/10.1017/S0272263197003045>
- Harrington, M., & Carey, M. (2009). The on-line Yes/No test as a placement tool. *System*, 37(4), 614–626. <https://doi.org/10.1016/j.system.2009.09.006>
- Horwitz, E. K. (2010). Foreign and second language anxiety. *Language Teaching*, 43(2), 154–167. <https://doi.org/10.1017/S026144480999036X>
- Horwitz, E. K., Horwitz, M. B., & Cope, J. (1986). Foreign language classroom anxiety. *The Modern Language Journal*, 70(2), 125–132. <https://doi.org/10.1111/j.1540-4781.1986.tb05256.x>
- Hulstijn, J. H. (2015). *Language proficiency in native and non-native speakers: Theory and research*. John Benjamins.
- Hulstijn, J. H. (2019). An individual-differences framework for comparing nonnative with native speakers: Perspectives from BLC theory. *Language Learning*, 69(s1), 157–183. <https://doi.org/10.1111/lang.12317>
- In'nami, Y. (2006). The effects of test anxiety on listening test performance. *System*, 34(3), 317–340. <https://doi.org/10.1016/j.system.2006.04.005>
- In'nami, Y., & Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing*, 26(2), 219–244. <https://doi.org/10.1177/0265532208101006>
- Jak, S. (2015). *Meta-analytic structural equation modelling*. Springer. <https://doi.org/10.1007/978-3-319-27174-3>
- Jeon, E.-H., & Yamashita, Y. (2014). L2 reading comprehension and its correlates: A meta-analysis. *Language Learning*, 64(1), 160–212. <https://doi.org/10.1111/lang.12034>
- Karalik, T., & Merç, A. (2019). Correlates of listening comprehension in L1 and L2: A meta-analysis. *Eurasian Journal of Applied Linguistics*, 5(3), 353–383. <https://doi.org/10.32601/ejal.651387>
- Koizumi, R., & In'nami, Y. (2020). Structural equation modeling of vocabulary size and depth using conventional and Bayesian methods. *Frontiers in Psychology: Language Sciences*, 11, 618. <https://doi.org/10.3389/fpsyg.2020.00618>
- Li, S. (2016). The construct validity of language aptitude: A meta-analysis. *Studies in Second Language Acquisition*, 38(4), 801–842. <https://doi.org/10.1017/S027226311500042X>
- Linck, J. A., Osthus, P., Koeth, J. T., & Bunting, M. F. (2014). Working memory and second language comprehension and production: A meta-analysis. *Psychonomic Bulletin & Review*, 21(4), 861–883. <https://doi.org/10.3758/s13423-013-0565-2>
- Masgoret, A. M., & Gardner, R. C. (2003). Attitudes, motivation, and second language learning: A meta-analysis of studies conducted by Gardner and Associates. *Language Learning*, 53(1), 123–163. <https://doi.org/10.1111/1467-9922.00212>
- McLean, S., Stewart, J., & Batty, A. O. (2020). Predicting L2 reading proficiency with modalities of vocabulary knowledge: A bootstrapping approach. *Language Testing*, 37(3), 389–411. <https://doi.org/10.1177/0265532219898380>



- Milton, J., & Masrai, A. (2021). Vocabulary and listening. In J. Clenton & P. Booth (Eds.), *Vocabulary and the four skills: Pedagogy, practice, and implications for teaching vocabulary* (pp. 45–59). Routledge.
- Milton, J., Wade, J., & Hopkins, N. (2010). Aural word recognition and oral competence in a foreign language. In R. Chacón-Beltrán, C. Abello-Contesse, & M. Torreblanca-López (Eds.), *Further insights into non-native vocabulary teaching and learning* (pp. 83–98). Multilingual Matters. <https://doi.org/10.21832/9781847692900-007>
- Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex “frontal lobe” tasks: A latent variable analysis. *Cognitive Psychology*, 41(1), 49–100. <https://doi.org/10.1006/cogp.1999.0734>
- Montero Perez, M., van den Noortgate, W., & Desmet, P. (2013). Captioned video for L2 listening and vocabulary learning: A meta-analysis. *System*, 41(3), 720–739. <https://doi.org/10.1016/j.system.2013.07.013>
- Muchinsky, P. M. (1996). The correction for attenuation. *Educational and Psychological Measurement*, 56(1), 63–75. <https://doi.org/10.1177/0013164496056001004>
- Nation, P. (2020). The different aspects of vocabulary knowledge. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 15–29). Routledge.
- Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *Language Teacher*, 31, 9–13. Retrieved on 11 January 2022 from [https://jalt-publications.org/files/pdf/the\\_language\\_teacher/07\\_2007tlt.pdf](https://jalt-publications.org/files/pdf/the_language_teacher/07_2007tlt.pdf)
- Nissan, S., DeVincenzi, F., & Tang, K. L. (1996). *An analysis of factors affecting the difficulty of dialogue Items in TOEFL listening comprehension* (Report Number: RR-95-37, TOEFL-RR-51). <https://doi.org/10.1002/j.2333-8504.1995.tb01671.x>
- Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 Instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, 50(3), 417–528. <https://doi.org/10.1111/0023-8333.00136>
- Ockey, G. J., & Wagner, E. (2018). *Assessing L2 listening: Moving towards authenticity*. John Benjamins. <https://doi.org/10.1075/llt.50>
- Orwin, R. (1983). A fail-safe N for effect size in meta-analysis. *Journal of Educational Statistics*, 8(2), 157–159. <https://doi.org/10.2307/1164923>
- Pae, T.-I. (2013). Skill-based L2 anxieties revisited: Their intra-relations and the inter-relations with general foreign language anxiety. *Applied Linguistics*, 34(2), 232–252. <https://doi.org/10.1093/applin/ams041>
- Papi, M., & Hiver, P. (2020). Language learning motivation as a complex dynamic system: A global perspective of truth, control, and value. *The Modern Language Journal*, 104(1), 209–232. <https://doi.org/10.1111/modl.12624>
- Plonsky, L., & Derrick, D. J. (2016). A meta-analysis of reliability coefficients in second language research. *The Modern Language Journal*, 100(2), 538–553. <https://doi.org/10.1111/modl.12335>
- Plonsky, L., & Oswald, F. L. (2014). How big is “big”? Interpreting effect sizes in L2 research. *Language Learning*, 64(4), 878–912. <https://doi.org/10.1111/lang.12079>
- Plonsky, L., & Oswald, F. L. (2015). Meta-analyzing second language research. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 106–128). Routledge. <https://doi.org/10.4324/9781315870908-6>
- Purpura, J. E. (2004). *Assessing grammar*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511733086>

- Qian, D. D., & Lin, L. H. (2020). The relationship between vocabulary knowledge and language proficiency. In S. Webb (ed), *The Routledge handbook of vocabulary studies* (pp. 66–80). Routledge.
- Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing*, 10(3), 355–371. <https://doi.org/10.1177/026553229301000308>
- Read, J. (2004). Plumbing the depths: How should the construct of vocabulary be defined? In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing* (pp. 209–227). John Benjamins. <https://doi.org/10.1075/llt.10.15rea>
- Römer, U. (2017). Language assessment and the inseparability of lexis and grammar: Focus on the construct of speaking. *Language Testing*, 34(4), 477–492. <https://doi.org/10.1177/0265532217711431>
- Rosenthal, R. (1979). The “file drawer problem” and tolerance for null results. *Psychological Bulletin*, 86(3), 638–641. <https://doi.org/10.1037/0033-2909.86.3.638>
- Ross, S. (1998). Self-assessment in second language testing: A meta-analysis and analysis of experiential factors. *Language Testing*, 15(1), 1–20. <https://doi.org/10.1177/026553229801500101>
- Sáfar, A., & Kormos, J. (2008). Revising problems with foreign language aptitude. *International Review of Applied Linguistics*, 46(2), 113–136. <https://doi.org/10.1515/IRAL.2008.005>
- Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Palgrave MacMillan. <https://doi.org/10.1057/9780230293977>
- Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language Learning*, 64(4), 913–951. <https://doi.org/10.1111/lang.12077>
- Shintani, N., & Wallace, P. M. (2014). Effects of listening support in second language classroom: A meta-analysis. *English Teaching and Learning*, 38(3), 71–101.
- Speciale, G., Ellis, N. C., & Bywater, T. (2004). Phonological sequence learning and short-term store capacity determine second language vocabulary acquisition. *Applied Psycholinguistics*, 25(2), 293–321. <https://doi.org/10.1017/S0142716404001146>
- Stansfield, C. W., & Reed, D. J. (2019). The MLAT at 60 years. In Z. Wen, P. Skehan, A. Biedron, S. Li., & R. L. Sparks (Eds.), *Language aptitude: Advancing theory, testing, research, research and practice* (pp. 15–32). Routledge. <https://doi.org/10.4324/9781315122021-2>
- Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing, *The Language Learning Journal*, 36(2), 139–152. <https://doi.org/10.1080/09571730802389975>
- Stæhr, L. S. (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Studies in Second Language Acquisition*, 31(4), 577–607. <https://doi.org/10.1017/S0272263109990039>
- Tafaghodtari, M. H., & Vandergrift, L. (2008). Second and foreign language listening: Unraveling the construct. *Perceptual and Motor Skills*, 107(1), 99–113. <https://doi.org/10.2466/pms.107.1.99-113>
- Taguchi, N. (2008). The effect of working memory, semantic access, and listening abilities on the comprehension of conversational implicatures in L2 English. *Pragmatics & Cognition*, 16(3), 517–539. <https://doi.org/10.1075/pc.16.3.05tag>
- Tanner-Smith, E. E., Tipton, E., & Polanin, J. R. (2016). Handling complex meta-analytic data structures using robust variance estimates: A tutorial in R. *Journal of Developmental and Life-Course Criminology*, 2(1), 85–112. <https://doi.org/10.1007/s40865-016-0026-5>
- Teimouri, Y., Goetze, J., & Plonsky, L. (2019). Second language anxiety and achievement: A meta-analysis. *Studies in Second Language Acquisition*, 41(2), 363–387. <https://doi.org/10.1017/S0272263118000311>

- Vafae, P. (2016). The relative significance of syntactic knowledge and vocabulary knowledge in second language listening comprehension (Unpublished doctoral dissertation). University of Maryland.
- Vandergrift, L. (2005). Relationships among motivation orientations, metacognitive awareness and proficiency in L2 listening. *Applied Linguistics*, 26(1), 70–89. <https://doi.org/10.1093/applin/amh039>
- Vandergrift, L., & Goh, C. C. M. (2012). *Teaching and learning second language listening: Metacognition in action*. Routledge.
- Vandergrift, L., Goh, C. C. M., Mareschal, C., & Tafaghodtari, M. H. (2006). The metacognitive awareness listening questionnaire (MALQ): Development and validation. *Language Learning*, 56(3), 431–462. <https://doi.org/10.1111/j.1467-9922.2006.00373.x>
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3), 1–48. <https://doi.org/10.18637/jss.v036.i03>
- Vogely, A. (1995). Perceived strategy use during performance on three authentic listening comprehension tasks. *The Modern Language Journal*, 79(1), 41–56. <https://doi.org/10.1111/j.1540-4781.1995.tb05414.x>
- Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (1994). Development of reading-related phonological processing abilities: New evidence of bidirectional causality from a latent variable longitudinal study. *Developmental Psychology*, 30(1), 73–87. <https://doi.org/10.1037/0012-1649.30.1.73>
- Wallace, M. P. (2018). Second language listening comprehension: Relationships among vocabulary knowledge, topical knowledge, metacognition, and working memory (Unpublished doctoral dissertation). National Institute of Education, Nanyang Technological University, Singapore.
- Wallace, M. P. (2022). Individual differences in second language listening: Examining the role of knowledge, metacognitive awareness, memory, and attention. *Language Learning*, 72(1), 5–44. <https://doi.org/10.1111/lang.12424>
- Wallace, M. P., & Lee, K. (2020). Examining second language listening, vocabulary, and executive functioning. *Frontiers in Psychology*, 11(1122), 1–14. <https://doi.org/10.3389/fpsyg.2020.01122>
- Webb, S. (2005). Receptive and productive vocabulary learning: the effects of reading and writing on word knowledge. *Studies in Second Language Acquisition*, 27(1), 33–52. <https://doi.org/10.1017/S0272263105050023>
- Wen, Z. (2016). *Working memory and second language learning: Towards an integrated approach*. Multilingual Matters. <https://doi.org/10.21832/9781783095735>
- Wen, Z. (2018). Working memory in first and second language: A comprehensive bibliography. Expanded and updated (9 Sept 2018) from: Wen, Zhisheng (2016). *Working memory in second language learning: An integrated approach* (References). Multilingual Matters. Retrieved from [https://www.academia.edu/12198656/Working\\_memory\\_in\\_first\\_and\\_second\\_language\\_A\\_comprehensive\\_bibliography](https://www.academia.edu/12198656/Working_memory_in_first_and_second_language_A_comprehensive_bibliography).
- Winke, P., & Lim, H. (2017). The effects of test preparation on second-language listening test performance. *Language Assessment Quarterly*, 14(4), 380–397. <https://doi.org/10.1080/15434303.2017.1399396>
- Yanagisawa, A., & Webb, S. (2020). Measuring depth of vocabulary knowledge. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 371–386). Routledge.
- Zhang, S., & Zhang, X. (2022). The relationship between vocabulary knowledge and L2 reading/listening comprehension: A meta-analysis. *Language Teaching Research*. <https://doi.org/10.1177/1362168820913998>

## Appendix A. Books examined to collect studies for the meta-analysis

Studies in Language Testing series published by Cambridge University Press.

- Brown, S. (2011). *Listening myths: Applying second language research to classroom teaching*. University of Michigan Press. <https://doi.org/10.3998/mpub.2132445>
- Buck, G. (2001). *Assessing listening*. Cambridge University Press.  
<https://doi.org/10.1017/CBO9780511732959>
- Field, J. (2009). *Listening in the language classroom*. Cambridge University Press.  
<https://doi.org/10.1017/CBO9780511575945>
- Field, J. (2018). *Rethinking the second language listening test: From theory to practice*. Equinox.
- Flowerdew, J., & Miller, L. (2005). *Second language listening: Theory and practice*. Cambridge University Press.
- Lynch, T. (2009). *Teaching second language listening*. Oxford University Press.
- Mendelsohn, D. J., & Rubin, J. (Eds.). (1995). *A guide for the teaching of second language listening*. Dominie Press.
- Nation, I. S. P., & Newton, J. (2009). *Teaching ESL/EFL listening and speaking*. Cambridge University Press.
- Ockey, G. J., & Wagner, E. (Eds.). (2018). *Assessing L2 listening: Moving towards authenticity*. John Benjamins. <https://doi.org/10.1075/llt.50>
- Rost, M. (2016). *Teaching and researching listening* (2nd ed.). Routledge.
- Vandergrift, L., & Goh, C. C. M. (2012). *Teaching and learning second language listening: Metacognition in action*. Routledge.
- Worthington, D. L., & Bodie, G. D. (Eds.). (2018). *The sourcebook of listening research: Methodology and measures*. Wiley-Blackwell.

## Appendix B. 118 studies included in the meta-analysis

- Afshari, S., & Tavakoli, M. (2017). The relationship between depth and breadth of vocabulary knowledge and Iranian EFL learners' listening comprehension. *International Journal of Research Studies in Language Learning*, 6(3), 13–24. Retrieved on 11 January 2022 from [http://consortiacademia.org/wp-content/uploads/IJRSLI/IJRSLI\\_v6i3/1438-5152-1-PB.pdf](http://consortiacademia.org/wp-content/uploads/IJRSLI/IJRSLI_v6i3/1438-5152-1-PB.pdf)
- Amin, I. A. R., Aly, M. A. S., & Mohammed, A. M. (2011). A correlation study between EFL strategic listening and listening comprehension skills among secondary school students. *Benha Faculty of Education Journal*, 23, 1–26. Retrieved on 11 January 2022 from <https://files.eric.ed.gov/fulltext/ED527448.pdf>
- Andersson, U. (2010). The contribution of working memory capacity to foreign language comprehension in children. *Memory*, 18(4), 458–472. <https://doi.org/10.1080/09658211003762084>
- Andringa, S., Olsthoorn, N., van Beuningen, C., Schoonen, R., & Hulstijn, J. (2012). Determinants of success in native and nonnative listening comprehension: An individual differences approach. *Language Learning*, 62(Suppl. 2), 49–78.  
<https://doi.org/10.1111/j.1467-9922.2012.00706.x>
- Aotani, M. (2011). Factors affecting the holistic listening of Japanese learners of English. Dissertation Abstracts International Section A: Humanities and Social Sciences, Vol 72(8-A), 2012, 2692.

- Atasheneh, N., & Izadi, A. (2012). The role of teachers in reducing/increasing listening comprehension test anxiety: A case of Iranian EFL learners. *English Language Teaching*, 5(3), 178–187. <https://doi.org/10.5539/elt.v5n3p178>
- Aukrust, V. G. (2008). Turkish-speaking first graders in Norway acquiring second language vocabulary, listening comprehension and literacy skills. *Scandinavian Journal of Educational Research*, 52(3), 293–314. <https://doi.org/10.1080/00313830802025108>
- Babayigit, S. (2014). The role of oral language skills in reading and listening comprehension of text: A comparison of monolingual (L1) and bilingual (L2) speakers of English language. *Journal of Research in Reading*, 37(S1), S22–S47. <https://doi.org/10.1111/j.1467-9817.2012.01538.x>
- Bekleyen, N. (2009). Helping teachers become better English students: Causes, effects, and coping strategies for foreign language listening anxiety. *System*, 37(4), 664–675. <https://doi.org/10.1016/j.system.2009.09.010>
- Brunfaut, T., & Révész, A. (2015). The role of task and listener characteristics in second language listening. *TESOL Quarterly*, 49(1), 141–168. <https://doi.org/10.1002/tesq.168>
- Call, M. E. (1985). Auditory short-term memory, listening comprehension, and the Input Hypothesis. *TESOL Quarterly*, 19(4), 765–781. <https://doi.org/10.2307/3586675>
- Carroll, J. B. (1967). *The foreign language attainments of language majors in the senior year – A survey conducted in U.S. colleges and universities*. Harvard University.
- Cheng, J., & Matthews, J. (2018). The relationship between three measures of L2 vocabulary knowledge and L2 listening and reading. *Language Testing*, 35(1), 3–25. <https://doi.org/10.1177/0265532216676851>
- Chiang, H.-H. (2018). English vocabulary size as a predictor of TOEIC Listening and Reading achievement among EFL Students in Taiwan. *Theory and Practice in Language Studies*, 8(2), 203–212. <https://doi.org/10.17507/tpls.o802.o4>
- Choi, I. C., Kim, K. S., & Boo, J. (2003). Comparability of a paper-based language test and a computer-based language test. *Language Testing*, 20(3), 295–320. <https://doi.org/10.1191/0265532203lt2580a>
- Cox, T. L., & Davies, R. S. (2012). Using automatic speech recognition technology with elicited oral response testing. *CALICO Journal*, 29(4), 601–618. <https://doi.org/10.11139/cj.29.4.601-618>
- Crosson, A. C., & Lesaux, N. K. (2010). Revisiting assumptions about the relationship of fluent reading to comprehension: Spanish-speakers' text-reading fluency in English. *Reading & Writing*, 23(5), 475–494. <https://doi.org/10.1007/s11145-009-9168-8>
- Dabbagh, A. (2016). The predictive role of vocabulary knowledge in listening comprehension: Depth or breadth? *International Journal of English Language & Translation Studies*, 4(3), 1–13.
- Droop, M., & Verhoeven, L. (2003). Language proficiency and reading ability in first- and second-language learners. *Reading Research Quarterly*, 38(1), 78–103. <https://doi.org/10.1598/RRQ.38.1.4>
- Elkhafaifi, H. (2005). Listening comprehension and anxiety in the Arabic language classroom. *The Modern Language Journal*, 89(2), 206–220. <https://doi.org/10.1111/j.1540-4781.2005.00275.x>
- Fay, A. (2012). Listening comprehension and working memory capacity in beginning L2 learners: An exploratory study (Unpublished doctoral dissertation). Pontifical Catholic University of Rio Grande do Sul, Brazil.
- Foomani, E. M. (2015). Lexical inferencing in listening: Depth of vocabulary knowledge and listening proficiency. *International Journal of English Language Education*, 3, 105–117. <https://doi.org/10.5296/ijele.v3i2.8081>
- French, L. M. (2006). *Phonological working memory and second language acquisition: A developmental study of francophone children learning English in Quebec*. Edwin Mellen Press.

- Gardner, R. C., & Lambert, W. E. (1965). Language aptitude, intelligence, and second-language achievement. *Journal of Educational Psychology*, 56(4), 191–199.  
<https://doi.org/10.1037/h0022400>
- Genesee, F., & Hamayan, E. (1980). Individual differences in young second language learners. *Applied Psycholinguistics*, 1(1), 95–110. <https://doi.org/10.1017/S0142716400000758>
- Geva, E., & Farnia, F. (2012). Developmental changes in the nature of language proficiency and reading fluency paint a more complex view of reading comprehension in ELL and EL1. *Reading & Writing*, 25(8), 1819–1845. <https://doi.org/10.1007/s11145-011-9333-8>
- Ghapanchi, Z., & Taheryan, A. (2012). Roles of linguistic knowledge, metacognitive knowledge and metacognitive strategy use in speaking and listening proficiency of Iranian EFL learners. *World Journal of Education*, 2(4), 64–75. <https://doi.org/10.5430/wje.v2n4p64>
- Goh, C. C. M., & Hu, G. W. (2014). Exploring the relationship between metacognitive awareness and listening performance with questionnaire data. *Language Awareness*, 23(3), 255–274.  
<https://doi.org/10.1080/09658416.2013.769558>
- Golchi, M. M. (2012). Listening anxiety and its relationship with listening strategy use and listening comprehension among Iranian IELTS learners. *International Journal of English Linguistics*, 2(4), 115–128. Retrieved on 22 January 2020 from <http://www.ccsenet.org/journal/index.php/ijel/article/view/17093>
- Goodwin, A. P., August, D., & Calderon, M. (2015). Reading in multiple orthographies: Differences and similarities in reading in Spanish and English for English learners. *Language Learning*, 65(3), 596–630. <https://doi.org/10.1111/lang.12127>
- Griffin, K. L. (1993). The relative contribution of reader variables to the comprehension of English texts by university-level Spanish speakers (Unpublished doctoral dissertation). The Ohio State University.
- Gu, S. S., & Wang, T. S. (2007). Study on the relationship between working memory and EFL listening comprehension. *CELEA Journal*, 30, 46–56. Retrieved from <http://www.celea.org.cn/teic/76/08031206.pdf>
- Gutierrez, X. (2016). Analyzed knowledge, metalanguage, and second language proficiency. *System*, 60, 42–54. <https://doi.org/10.1016/j.system.2016.06.003>
- Guo, S. (2001). A multidimensional analysis of reading English as a second language by native speakers of Chinese (Unpublished doctoral dissertation). University of Iowa.
- Harley, B., & Hart, D. (1997). Language aptitude and second language proficiency in classroom learners of different starting ages. *Studies in Second Language Acquisition*, 19(3), 379–400.  
<https://doi.org/10.1017/S0272263197003045>
- Harley, B., & Hart, D. (2002). Age, aptitude, and second language learning on a bilingual exchange. In P. Robinson (Ed.), *Individual differences and instructed language learning* (pp. 301–330). John Benjamins. <https://doi.org/10.1075/llt.2.15har>
- Harrington, M., & Carey, M. (2009). The on-line Yes/No test as a placement tool. *System*, 37(4), 614–626. <https://doi.org/10.1016/j.system.2009.09.006>
- Harsch, C., & Hartig, J. (2016). Comparing C-tests and Yes/No vocabulary size tests as predictors of receptive language skills. *Language Testing*, 33(4), 555–575.  
<https://doi.org/10.1177/0265532215594642>
- Henning, G. H., Ghawaby, S. M., Saadalla, W. Z., El-Rifai, M. A., Hannallah, R. K., & Mattar, M. S. (1981). Comprehensive assessment of language proficiency and achievement among learners of English as a foreign language. *TESOL Quarterly*, 15(4), 457–466.  
<https://doi.org/10.2307/3586486>



- Hutchinson, J. M., Whiteley, H. E., Smith, C. D., & Connors, L. (2003). The developmental progression of comprehension-related skills in children learning EAL. *Journal of Research in Reading*, 26(1), 19–32. <https://doi.org/10.1111/1467-9817.261003>
- In'nami, Y. (2006). The effects of test anxiety on listening test performance. *System*, 34(3), 317–340. <https://doi.org/10.1016/j.system.2006.04.005>
- Irvine, P., Atai, P., & John W. Oller, Jr. (1974). Cloze, dictation, and the Test of English as a foreign language. *Language Learning*, 24(2), 245–252. <https://doi.org/10.1111/j.1467-1770.1974.tb00506.x>
- Jeon, E. H. (2011). Contribution of morphological awareness to second-language reading comprehension. *The Modern Language Journal*, 95(2), 217–235. <https://doi.org/10.1111/j.1540-4781.2011.01179.x>
- Joyce, P. D. (2003). The breadth of vocabulary learning at a Japanese university. In Korea Tesol (Ed.), *KOTESOL conference proceedings 2003* (pp. 171–182). Korea TESOL.
- Joyce, P. (2008). Linguistic knowledge and psycholinguistic processing skills as components of L2 listening comprehension (Unpublished doctoral dissertation). University of Roehampton, London.
- Kanzaki, M. (2010). Vocabulary size, TOEIC scores, and testwiseness. In A. M. Stoke (Ed.), *JALT2009 conference proceedings* (pp. 740–748). JALT.
- Kanzaki, M. (2015). Comparing TOEIC and vocabulary test scores. In G. Brooks, M. Grogan & M. Porter (Eds.), *The 2014 PanSIG conference proceedings* (pp. 52–58). JALT.
- Kassem, H. M. (2015). The relationship between listening strategies used by Egyptian EFL college sophomores and their listening comprehension and self-efficacy. *English Language Teaching*, 8, 153–169. <https://doi.org/10.5539/elt.v8n2p153>
- Keitges, D. J. (1986). Relationships between foreign-language aptitude, attitudes and motivation, and personality traits, and the attained English language proficiency of Japanese university students (Unpublished doctoral dissertation). University of Texas at Austin.
- Kim, J. H. (2000). Foreign language listening anxiety: A study of Korean students learning English (Unpublished doctoral dissertation). University of Texas.
- Kök, İ. (2017). Relationship between listening comprehension strategy use and listening comprehension proficiency. *International Journal of Listening*, 32(3), 163–179. <https://doi.org/10.1080/10904018.2016.1276457>
- Kormos, J., & Sáfár, A. (2008). Phonological short-term memory, working memory and foreign language performance in intensive language learning. *Bilingualism: Language and Cognition*, 11(2), 261–271. <https://doi.org/10.1017/S1366728908003416>
- Lee, E.-H., Chae, Y., Ahn, J., & Kim, Y. (2011). 한국어 학습자의 인지적·정의적 개인차 요인과 제2언어 성취 간의 관계 [Relationships of cognitive and affective individual difference factors to L2 achievement and proficiency of learners of Korean]. *Bilingual Research*, 46, 253–296.
- Lin, Y. (2004). Chinese vocabulary acquisition and learning Chinese as a second language (Unpublished doctoral dissertation). The University of Iowa.
- Liu, M. (2011). 英语词汇知识和听力理解的关系研究 [The relationship between English vocabulary knowledge and listening comprehension]. 中国电力教育 (China Electric Power Education), 17, 193–196.
- Liu, S. (1995). 英语词汇量与听力理解关系的研究 [The relationship between English vocabulary size and listening comprehension]. 海南大学学报 (Journal of Hainan University (Social Science)), 2, 84–89.

- Mahdavi Zafarghandi, A., & Bahrpeyma, M. (2017). The relationship between short-term memory and listening comprehension ability of IELTS test takers at different language proficiency levels. *International Journal of Research Studies in Language Learning*, 6(3), 33–45. Retrieved on 11 January 2022 from [http://consortiacademia.org/wp-content/uploads/IJRSL/IJRSL\\_v6i3/1441-5234-1-PB.pdf](http://consortiacademia.org/wp-content/uploads/IJRSL/IJRSL_v6i3/1441-5234-1-PB.pdf)
- Matthews, J. (2018). Vocabulary for listening: Emerging evidence for high and mid-frequency vocabulary knowledge. *System*, 72, 23–36. <https://doi.org/10.1016/j.system.2017.10.005>
- Matthews, J., & Cheng, J. (2015). Recognition of high frequency words from speech as a predictor of L2 listening comprehension. *System*, 52, 1–13. <https://doi.org/10.1016/j.system.2015.04.015>
- McLean, S., Kramer, B., & Beglar, D. (2015). The creation and validation of a listening vocabulary levels test. *Language Teaching Research*, 19(6), 741–760. <https://doi.org/10.1177/1362168814567889>
- Mecartty, F. H. (1994). Lexical and grammatical knowledge in second language reading and listening comprehension (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign.
- Mihaljević Djigunović, J., & Legac, V. (2008). Foreign language anxiety and listening comprehension of monolingual and bilingual EFL learners. *Studia Romanica et Anglici Zagrabienia*, 53, 327–347. Retrieved on 11 January 2020 from <https://hrcak.srce.hr/40456>
- Mills, N., Pajares, F., & Herron, C. (2006). A reevaluation of the role of anxiety: Self-efficacy, anxiety, and their relation to reading and listening proficiency. *Foreign Language Annals*, 39(2), 276–295. <https://doi.org/10.1111/j.1944-9720.2006.tb02266.x>
- Milton, J., Wade, J., & Hopkins, N. (2010). Aural word recognition and oral competence in a foreign language. In R. Chacón-Beltrán, C. Abello-Contesse, & M. Torreblanca-López (Eds.), *Further insights into non-native vocabulary teaching and learning* (pp. 83–98). Multilingual Matters. <https://doi.org/10.21832/9781847692900-007>
- Miralpeix, I., & Munoz, C. (2018). Receptive vocabulary size and its relationship to EFL language skills. *International Review of Applied Linguistics in Language Teaching*, 56(1), 1–24. <https://doi.org/10.1515/iral-2017-0016>
- Naiman, N., Frohlich, M., Stern, H. H., & Todesco, A. (1978). *The good language learner*. Research in Education Series No. 7. Ontario Institute for Studies in Education.
- Nakamoto, J., Lindsey, K. A., & Manis, F. R. (2008). A cross-linguistic investigation of English language learners' reading comprehension in English and Spanish. *Scientific Studies of Reading*, 12(4), 351–371. <https://doi.org/10.1080/10888430802378526>
- Oh, E. J., & Lee, J. M. (2014). The role of linguistic knowledge and listening strategies in bottom-up and top-down processing of L2 listening. *English Teaching*, 69(2), 149–173. <https://doi.org/10.15858/engtea.69.2.201406.149>
- Oh, E. (2016). Comparative studies on the roles of linguistic knowledge and sentence processing speed in L2 listening and reading comprehension in an EFL tertiary setting. *Reading Psychology*, 37(2), 257–285. <https://doi.org/10.1080/02702711.2015.1049389>
- Onaha, H. (2004). Effect of shadowing and dictation on listening comprehension ability of Japanese EFL learners based on the theory of working memory. *JACET Bulletin*, 39, 137–148. Retrieved on 11 January 2022 from <https://dl.ndl.go.jp/info:ndljp/pid/10501447/1>
- Onaha, H. (2006). How the phonological loop related to listening. *SELT-Okinawa Review*, 6, 95–110. Retrieved on 11 January 2022 from <http://ir.lib.u-ryukyu.ac.jp/bitstream/20.500.12000/18587/1/No19p025.pdf>
- Onaha, H. (2010). Phonological short-term memory and early foreign language reading. *Scripsimus*, 19, 25–36. Retrieved on 11 January 2022 from <http://ir.lib.u-ryukyu.ac.jp/bitstream/20.500.12000/18587/1/No19p025.pdf>



- Pae, H.-K., & O'Brien, B. (2018). Overlap and uniqueness: Linguistic componential traits contributing to expressive skills in English as a foreign language. *Reading Psychology*, 39(4), 384–412. <https://doi.org/10.1080/02702711.2018.1443298>
- Park, E. C. (2004). The relationship between morphological awareness and lexical inference ability for English language learning children with Korean first-language background (Unpublished doctoral dissertation). Carnegie Mellon University.
- Poelmans, P. (2003). Developing second-language listening comprehension: Effects of training lower-order skills versus higher-order strategy (Unpublished doctoral dissertation). Universiteit van Amsterdam, The Netherlands.
- Proctor, C. P., Carlo, M., August, D., & Snow, C. (2005). Native Spanish-speaking children reading in English: Toward a model of comprehension. *Journal of Educational Psychology*, 97(2), 246–256. <https://doi.org/10.1037/0022-0663.97.2.246>
- Sáfar, A., & Kormos, J. (2008). Revisiting problems with foreign language aptitude. *International Review of Applied Linguistics in Language Teaching*, 46(2), 113–136. <https://doi.org/10.1515/IRAL.2008.005>
- Sağlam, S. (2014). The role of vocabulary breadth, syntactic knowledge, and listening strategy use on listening comprehension. *Route Educational and Social Science Journal*, 1, 54–72. <https://doi.org/10.17121/ressjournal.49>
- Sakuma, Y. (2004). The characteristics of memory representations in listening span tests and the EFL abilities. *Annual Review of English Language Education in Japan*, 15, 91–100.
- Sato, Y. (2017). *The relationship among working memory, short-term memory, and L2 proficiency*. Paper presented at the 43rd annual conference of the Japan Society of English Language Education (*Conference proceedings*, pp. 514–515).
- Satori, M. (2011). The effect of working memory on component processes of L2 listening. *International Journal of Social and Cultural Studies*, 4, 43–59. Retrieved from [http://reposit.lib.kumamoto-u.ac.jp/bitstream/2298/28788/1/SCS0004\\_043-059.pdf](http://reposit.lib.kumamoto-u.ac.jp/bitstream/2298/28788/1/SCS0004_043-059.pdf)
- Satori, M. (2012a). The effect of L1 and L2 working memory on L2 listening comprehension. *International Journal of Social and Cultural Studies*, 5, 13–27. Retrieved from [http://reposit.lib.kumamoto-u.ac.jp/bitstream/2298/28790/2/SCS0005\\_013-027.pdf](http://reposit.lib.kumamoto-u.ac.jp/bitstream/2298/28790/2/SCS0005_013-027.pdf)
- Satori, M. (2012b). The role of working memory in L2 listening comprehension. *Proceedings of the 17th Conference of Pan-Pacific Association of Applied Linguistics* (pp. 8–9).
- Satori, M. (2013). Working memory and L2 linguistic knowledge as components of L2 listening comprehension. *Kumamoto University Studies in Social and Cultural Science*, 11, 109–127. Retrieved from [http://reposit.lib.kumamoto-u.ac.jp/bitstream/2298/28690/3/SB0011\\_109-127.pdf](http://reposit.lib.kumamoto-u.ac.jp/bitstream/2298/28690/3/SB0011_109-127.pdf)
- Serraj, S., & Noordin, N. B. (2013). Relationship among Iranian EFL students' foreign language anxiety, foreign language listening anxiety and their listening comprehension. *English Language Teaching*, 6(5), 1–12. <https://doi.org/10.5539/elt.v6n5p1>
- Service, E. (1992). Phonology, working memory, and foreign-language learning. *Quarterly Journal of Experimental Psychology Section A*, 45(1), 21–50. <https://doi.org/10.1080/14640749208401314>
- Service, E., & Kohonen, V. (1995). Is the relation between phonological memory and foreign language learning accounted for by vocabulary acquisition? *Applied Psycholinguistics*, 16(2), 155–172. <https://doi.org/10.1017/S0142716400007062>
- Speciale, G., Ellis, N. C., & Bywater, T. (2004). Phonological sequence learning and short-term store capacity determine second language vocabulary acquisition. *Applied Psycholinguistics*, 25(2), 293–321. <https://doi.org/10.1017/S0142716404001146>

- Spies, T. G. (2011). Academic language proficiency development and its impact on reading comprehension: Within and across languages (Unpublished doctoral dissertation). Texas A&M University.
- Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *The Language Learning Journal*, 36(2), 139–152. <https://doi.org/10.1080/09571730802389975>
- Stæhr, L. S. (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Studies in Second Language Acquisition*, 31(4), 577–607. <https://doi.org/10.1017/S0272263109990039>
- Tafaghodtari, M. H., & Vandergrift, L. (2008). Second and foreign language listening: Unraveling the construct. *Perceptual and Motor Skills*, 107(1), 99–113. <https://doi.org/10.2466/pms.107.1.99-113>
- Taguchi, K. (2015). An exploratory study on correlations among vocabulary size, vocabulary learning strategy use, TOEIC scores and self-efficacy. *Economic Review of Toyo University*, 41, 55–67. Retrieved from [https://toyo.repo.nii.ac.jp/?action=repository\\_action\\_common\\_download&item\\_id=8306&item\\_no=1&attribute\\_id=22&file\\_no=1](https://toyo.repo.nii.ac.jp/?action=repository_action_common_download&item_id=8306&item_no=1&attribute_id=22&file_no=1)
- Taguchi, N. (2008). The effect of working memory, semantic access, and listening abilities on the comprehension of conversational implicatures in L2 English. *Pragmatics & Cognition*, 16(3), 517–539. <https://doi.org/10.1075/pc.16.3.o5tag>
- Teng, F. (2014). Assessing the depth and breadth of vocabulary knowledge with listening comprehension. *PASAA: A Journal of Language Teaching and Learning*, 48(2), 29–56. Retrieved on 12 January 2022 from <https://files.eric.ed.gov/fulltext/EJ1077893.pdf>
- Teng, F. (2016). An in-depth investigation into the relationship between vocabulary knowledge and academic listening comprehension. *TESL-EJ*, 20, 1–17. Retrieved on 12 January 2022 from <http://www.tesl-ej.org/pdf/ej78/a5.pdf>
- Tsuhihira, T. (2007). L2 working memory capacity and L2 listening test scores of Japanese junior college students. *Bunkyo Gakuin Foreign Language Department of Bunkyo Gakuin Junior College*, 7, 159–175. Retrieved on 12 January 2022 from [https://www.u-bunkyo.ac.jp/center/library/image/fsell2007\\_159-175.PDF](https://www.u-bunkyo.ac.jp/center/library/image/fsell2007_159-175.PDF)
- Uchikoshi, Y. (2013). Predictors of English reading comprehension: Cantonese-speaking English language learners in the U.S. *Reading and Writing*, 26(6), 913–939. <https://doi.org/10.1007/s11145-012-9398-z>
- Vafaei, P. (2016). The relative significance of syntactic knowledge and vocabulary knowledge in second language listening comprehension (Unpublished doctoral dissertation). University of Maryland.
- Valizadeh, M. R., & Alavinia, P. (2013). Listening comprehension performance viewed in the light of emotional intelligence and foreign language listening anxiety. *English Language Teaching*, 6(12), 11–26. <https://doi.org/10.5539/elt.v6n12p11>
- Vandergrift, L. (2005). Relationships among motivation orientations, metacognitive awareness and proficiency in L2 listening. *Applied Linguistics*, 26(1), 70–89. <https://doi.org/10.1093/applin/amh039>
- Vandergrift, L., & Baker, S. C. (2015). Learner variables in second language listening comprehension: An exploratory path analysis. *Language Learning*, 65(2), 390–416. <https://doi.org/10.1111/lang.12105>
- Vandergrift, L., & Baker, S. C. (2018). Learner variables important for success in L2 listening comprehension in French immersion classrooms. *Canadian Modern Language Review*, 74(1), 79–100. <https://doi.org/10.3138/cmlr.3906>

- Vandergrift, L., Goh, C. C. M., Mareschal, C., & Tafaghodtari, M. H. (2006). The metacognitive awareness listening questionnaire (MALQ): Development and validation. *Language Learning*, 56(3), 431–462. <https://doi.org/10.1111/j.1467-9922.2006.00373.x>
- Vogely, A. (1995). Perceived strategy use during performance on three authentic listening comprehension tasks. *The Modern Language Journal*, 79(1), 41–56. <https://doi.org/10.1111/j.1540-4781.1995.tb05414.x>
- Wallace, M. P. (2018). Second language listening comprehension: Relationships among vocabulary knowledge, topical knowledge, metacognition, and working memory (Unpublished doctoral dissertation). National Institute of Education, Nanyang Technological University, Singapore.
- Wang, S. (2015). An empirical study on the role of vocabulary knowledge in EFL listening comprehension. *Theory and Practice in Language Studies*, 5(5), 989–995. <https://doi.org/10.17507/tpls.0505.14>
- Wang, Y., & Treffers-Daller, J. (2017). Explaining listening comprehension among L2 learners of English: the contribution of general language proficiency, vocabulary knowledge and metacognitive awareness. *System*, 65, 139–150. <https://doi.org/10.1016/j.system.2016.12.013>
- Wilson, I., Kaneko, E., Lyddon, P., Okamoto, K., & Ginsburg, J. (2011). Nonsensesyllable sound discrimination ability correlates with second language (L2) proficiency. In W.-S. Lee & E. Zee (Eds.), *Proceedings of the 17th International Congress of Phonetic Sciences* (pp. 2133–2136). City University of Hong Kong.
- Winke, P. (2005). Individual differences in adult Chinese second language acquisition: The relationships among aptitude, memory, and strategies for learning (Unpublished doctoral dissertation). Georgetown University, Washington, DC.
- Winke, P. (2013). An investigation into second language aptitude for advanced Chinese language learning. *The Modern Language Journal*, 97(1), 109–130. <https://doi.org/10.1111/j.1540-4781.2013.01428.x>
- Winke, P., & Lim, H. (2017). The effects of test preparation on second-language listening test performance. *Language Assessment Quarterly*, 14(4), 380–397. <https://doi.org/10.1080/15434303.2017.1399396>
- Wong, S. W. L., Mok, P. P. K., Chung, K. K.-H., Leung, V. W. H., Bishop, D. V. M., & Chow, B. W.-Y. (2017). Perception of native English reduced forms in Chinese learners: Its role in listening comprehension and its phonological correlates. *TESOL Quarterly*, 51(1), 7–31. <https://doi.org/10.1002/tesq.273>
- Yamashita, J., & Shiotsu, T. (2017). Comprehension and knowledge components that predict L2 reading: A latent-trait approach. *Applied Linguistics*, 38(1), 43–67. <https://doi.org/10.1093/applin/amu079>
- Yang, R.-L. (1994). *A study of the communicative anxiety and self-esteem of Chinese students in relation to their oral and listening proficiency in English*. 語文學報 (Language Journal), 1, 124–208. National Hsinchu University of Education. Retrieved from [http://readopac3.ncl.edu.tw/nclJournal/search/detail.jsp?sysId=0005415286&dtId=000040&search\\_type=detail&la=ch%20?iframe=true](http://readopac3.ncl.edu.tw/nclJournal/search/detail.jsp?sysId=0005415286&dtId=000040&search_type=detail&la=ch%20?iframe=true)
- Yoshizawa, K. (2002). Relationships among strategy use, foreign language aptitude, and second language proficiency: A structural equation modelling approach (Unpublished PhD dissertation). Temple University.
- Zhang, X. D. (2011). 词汇知识与二语听力理解关系研究 [On the relationship of lexical knowledge and listening comprehension]. *Foreign Language World*, 143, 36–42.
- Zuo, X. (2013). 英语听力理解策略与听力水平的相关性研究 [The relevant research on English listening comprehension strategies and listening levels]. *Disciplines Exploration*, 50–51.

## Appendix C. Coding Sheet

Variable	Subset	Example instrument
Listening	–	Large-scale proficiency tests (e.g., Cambridge First Certificate Exam listening section; CET4 listening section; IELTS listening section; TOEFL listening section; TOEIC listening section), textbook-based and author-made listening tests
<i>Linguistic knowledge</i>		
Grammar	–	Grammaticality judgement test, Iowa Test of Basic Skills (ITBS) grammar section, and Michigan Test of English Language Proficiency grammar section
Vocabulary	Size	Aural vocabulary knowledge (AVK) test (Matthew, 2018), Listening Vocabulary Levels Test (LVL; McLean, Kramer, & Beglar, 2015), Productive Levels Test (Laufer & Nation, 1999), Vocabulary Levels Test (Nation, 1983; Schmitt, Schmitt, & Clapham, 2001), Vocabulary Size Test (Nation & Beglar, 2007), and WLPB–Picture Vocabulary subtest (Woodcock, 1991)
	Depth	Depth of vocabulary knowledge test (DVKT; developed by Qian & Schedl, 2004), vocabulary depth test (Wesche & Paribakht, 1996), Word Associates Test (Read, 1993, 1998), and modified version of lexical inference tasks (Mori & Nagy, 1999)
	Recognition	Depth of vocabulary knowledge test (DVKT; developed by Qian & Schedl, 2004), Listening Vocabulary Levels Test (LVL; McLean, Kramer, & Beglar, 2015), vocabulary depth test (Wesche & Paribakht, 1996), Vocabulary Levels Test (Nation, 1983; Schmitt et al., 2001), Vocabulary Size Test (Nation & Beglar, 2007), and Word Associates Test (Read, 1993, 1998)
	Production	Aural vocabulary knowledge (AVK) test (Matthew, 2018) and WLPB–Picture Vocabulary subtest (Woodcock, 1991)
	Aural	Aural vocabulary knowledge (AVK) test (Matthew, 2018), Listening Vocabulary Levels Test (LVL; McLean, Kramer, & Beglar, 2015), and WLPB–Picture Vocabulary subtest (Woodcock, 1991)
	Written	Depth of vocabulary knowledge test (DVKT; developed by Qian & Schedl, 2004), vocabulary depth test (Wesche & Paribakht, 1996), Vocabulary Levels Test (Nation, 1983; Schmitt et al., 2001), Vocabulary Size Test (Nation & Beglar, 2007), and Word Associates Test (Read, 1993, 1998)
Phonological awareness	–	Phonemic discrimination, segmentation accuracy; The blending, elision, and segmenting subtests of the Comprehensive Test of Phonological Processing (CTOPP; Wagner, Torgesen, & Rashotte, 1999)
Morphological knowledge	–	Test adapted from the Derivation task of Carlisle's (2000) Test of Morphological Structure; test adapted from Schmitt and Meara's (1997) test of verbal suffix knowledge; and the Recalling Sentences subtest from the CELF-4UK (Semel et al., 2006)

Variable	Subset	Example instrument
<i>Cognitive ability</i>		
Aptitude	Phonetic coding	MLAT Part 2 (phonetic script), MLAT Parts 2 and 3 (spelling cues) combined <sup>a</sup> , LLAMA D (speech recognition), LLAMA E (sound-symbol connect)
	Language analytic ability (grammatical sensitivity)	MLAT Part 4 (words in sentences), LLAMA F (grammatical reasoning), PLAB 4
	Rote memory	MLAT Part 5 (paired associates), MLAT Parts 5 and 1 (number learning) combined <sup>a</sup> , LLAMA B (vocabulary learning)
Metacognitive awareness	–	Metacognitive Awareness Listening Questionnaire (Vandergrift et al., 2006), Metacognitive Awareness Strategy Questionnaire (MASQ; Carrell, 1989), Strategy Inventory for Language Learning (SILL; Oxford, 1990), and vocabulary learning questionnaire (Mizumoto, Someya, & Yamanishi, 2014)
	MALQ <sup>b</sup> total score	Metacognitive Awareness Listening Questionnaire (Vandergrift et al., 2006) total score
	MALQ <sup>b</sup> directed attention	Metacognitive Awareness Listening Questionnaire (Vandergrift et al., 2006) directed attention total score
	MALQ <sup>b</sup> mental translation	Metacognitive Awareness Listening Questionnaire (Vandergrift et al., 2006) mental translation total score
	MALQ <sup>b</sup> person knowledge	Metacognitive Awareness Listening Questionnaire (Vandergrift et al., 2006) person knowledge total score
	MALQ <sup>b</sup> planning and evaluation	Metacognitive Awareness Listening Questionnaire (Vandergrift et al., 2006) planning and evaluation total score
	MALQ <sup>b</sup> problem solving	Metacognitive Awareness Listening Questionnaire (Vandergrift et al., 2006) problem solving total score
Working memory	Phonological	Backward digit span, counting span, digit span, nonword span, pseudoword repetition, sentence repetition, and word span <sup>c</sup>
	Executive	Listening span, reading span <sup>c</sup>
	Executive updating	Japanese letter-memory test (adapted from Miyake et al., 2000 and Friedman et al., 2008), keep-track test (adapted from Miyake et al., 2000 and Friedman et al., 2008), and visual-spatial 3-back test (adapted from Schmiedek, Hildebrandt, Lovden, Lindenberger, & Wilhelm, 2009)
	Executive shifting	Global-local test (Miyake et al., 2000), number-letter test (Miyake et al., 2000), and plus-minus test (Miyake et al., 2000)

Variable	Subset	Example instrument
<i>Affective feature</i>		
Attitude	–	Test adapted from Attitude/Motivation Test Battery (AMTB; Gardner & Smythe, 1975)
Anxiety	Foreign language listening anxiety	Foreign Language Listening Anxiety Scale (FLLAS; Elkhafaifi, 2005)
	Test anxiety	Test Anxiety Scale (TAS; Sarason, 1975), Test Influence Inventory (TII; Fujii, 1993), and test-taking anxiety questionnaire (adapted from Cassady & Johnson, 2002)
	Foreign language classroom anxiety	Foreign Language Classroom Anxiety Scale (FLCAS; Horwitz et al., 1986)
Motivation	–	Test adapted from Attitude/Motivation Test Battery (AMTB; Gardner & Smythe, 1975); Language Learning Motivation Orientation Scale (Noels, Pelletier, Clement, & Vallerand, 2000)

*Note.*

a. According to Stansfield and Reed (2019), the MLAT consists of five sections. Part 1 (number learning) measures phonetic coding ability (weakly), rote learning ability (weakly), and inductive learning (weakly). Part 2 (phonetic script) measures phonetic coding ability. Part 3 (spelling cues) measures phonetic coding ability (weakly). Part 4 (words in sentences) measures grammatical sensitivity. Part 5 (paired associates) measures rote learning ability. Parts 2, 4, and 5 measure one component of aptitude each, whereas Part 1 measures three components of aptitude weakly, and Part 3 measures one component weakly. Compared to Parts 2, 4, and 5, where there was a one-to-one correspondence between aptitude components and sections, it was not clear what “weakly” referred to in Parts 1 and 3. Thus, Part 3 was coded as a measure of phonetic coding only when it was reported as a combined score with Part 2. Part 1 was coded as a measure of rote memory only when it was reported as a combined score with Part 5.

b. Metacognitive Awareness Listening Questionnaire (Vandergrift et al., 2006).

c. Working memory tests were classified into measuring phonological or executive working memory following Linck et al. (2014).

Instruction for coding: Read each study carefully. When information is partially reported or not reported at all, search online for further details (e.g., test manuals; studies that used the same test) and check other studies conducted by the same author(s).

## References for Appendix C

- Carlisle, J. F. (2000). Awareness of the structure and meaning of morphologically complex words: Impact on reading. *Reading and Writing*, 12(3/4), 169–190. <https://doi.org/10.1023/A:1008131926604>
- Carrell, P. L. (1989). Metacognitive awareness and second language reading. *The Modern Language Journal*, 73(2), 121–134. <https://doi.org/10.1111/j.1540-4781.1989.tb02534.x>

- Cassady, J. C., & Johnson, R. E. (2002). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology*, 27(2), 270–295. <https://doi.org/10.1006/ceps.2001.1094>
- Elkhafafi, H. (2005). Listening comprehension and anxiety in the Arabic language classroom. *The Modern Language Journal*, 89(2), 206–219. <https://doi.org/10.1111/j.1540-4781.2005.00275.x>
- Friedman, N. P., Miyake, A., Young, S. E., DeFries, J. C., Corley, R. P., & Hewitt, J. K. (2008). Individual differences in executive functions are almost entirely genetic in origin. *Journal of Experimental Psychology: General*, 137(2), 201–225. <https://doi.org/10.1037/0096-3445.137.2.201>
- Fujii, Y. (1993). Tesuto eikyo inventori no sakusei [Construction of a Test Influence Inventory (TII)]. *Japanese Journal of Psychology*, 64(2), 135–139. <https://doi.org/10.4992/jjpsy.64.135>
- Gardner, R. C., & Smythe, P. C. (1975). Motivation and second-language acquisition. *Canadian Modern Language Review*, 31(3), 218–230. <https://doi.org/10.3138/cmrlr.31.3.218>
- Horwitz, E. K., Horwitz, M., & Cope, J. (1986). Foreign language classroom anxiety. *The Modern Language Journal*, 70(2), 125–132. <https://doi.org/10.1111/j.1540-4781.1986.tb05256.x>
- Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing*, 16(1), 33–51. <https://doi.org/10.1177/026553229901600103>
- Linck, J. A., Osthus, P., Koeth, J. T., & Bunting, M. F. (2014). Working memory and second language comprehension and production: A meta-analysis. *Psychonomic Bulletin & Review*, 21(4), 861–883. <https://doi.org/10.3758/s13423-013-0565-2>
- Matthews, J. (2018). Vocabulary for listening: Emerging evidence for high and mid-frequency vocabulary knowledge. *System*, 72, 23–36. <https://doi.org/10.1016/j.system.2017.10.005>
- McLean, S., Kramer, B., & Begler, D. (2015). The creation and validation of a listening vocabulary levels test. *Language Teaching Research*, 19(1), 1–20. <https://doi.org/10.1177/1362168814567889>
- Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex “Frontal Lobe” tasks: A latent variable analysis. *Cognitive Psychology*, 41(1), 49–100. <https://doi.org/10.1006/cogp.1999.0734>
- Mizumoto, A., Someya, Y., & Yamanishi, H. (2014). Goigakushu wo sokushinsuru burendido ranningu no kokoromi: “Grammar & Vocabulary Development” no rinen to sono koka nikan-suru chukanhokoku [Using blended-learning for fostering vocabulary learning: An interim report of the newly implemented “Grammar & Vocabulary Development” Course]. *Journal of Faculty of Foreign Language Studies, Kansai University*, 11, 71–92. Retrieved on 12 January 2022 from [https://kansai-u.repo.nii.ac.jp/?action=pages\\_view\\_main&active\\_action=repository\\_view\\_main\\_item\\_detail&item\\_id=10221&item\\_no=1&page\\_id=13&block\\_id=21](https://kansai-u.repo.nii.ac.jp/?action=pages_view_main&active_action=repository_view_main_item_detail&item_id=10221&item_no=1&page_id=13&block_id=21)
- Mori, Y. & Nagy, W. E. (1999). Integration of information from context and word elements in interpreting novel kanji compounds. *Reading Research Quarterly*, 34(1), 80–101. <https://doi.org/10.1598/RRQ.34.1.5>
- Nation, P. (1983). Testing and teaching vocabulary. *Guidelines*, 5, 12–25. Retrieved on 12 January 2022 from <https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/paul-nations-publications/publications/documents/1983-Testing-and-teaching.pdf>
- Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *Language Teacher*, 31, 9–13. Retrieved on 12 January 2022 from [https://jalt-publications.org/files/pdf/the\\_language\\_teacher/07\\_2007tlt.pdf](https://jalt-publications.org/files/pdf/the_language_teacher/07_2007tlt.pdf)
- Noels, K. A., Pelletier, L. G., Clement, R., & Vallerand, R. (2000). Why are you learning a second language? Motivational orientations and self-determination theory. *Language Learning*, 50(1), 57–85. <https://doi.org/10.1111/0023-8333.00111>
- Oxford, R. L. (1990). *Language learning strategies: What every teacher should know*. Heinle & Heinle.



- Qian, D. D., & Schedl, M. (2004). Evaluation of an in-depth vocabulary knowledge measure for assessing reading performance. *Language Testing*, 21(1), 28–52.  
<https://doi.org/10.1191/0265532204lt2730a>
- Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing*, 10(3), 355–371. <https://doi.org/10.1177/026553229301000308>
- Read, J. (1998). Validating a test to measure depth of vocabulary knowledge. In A. J. Kunnan (ed.), *Validation in language assessment* (pp. 41–60). Lawrence Erlbaum Associates.  
<https://doi.org/10.4324/9780203053768>
- Sarason, I. G. (1975). The Test Anxiety Scale: concept and research. In I. G. Sarason, C. D. Spielberger (Eds.), *Stress and anxiety* (Vol. 2, pp. 193–217). Hemisphere.
- Schmiedek, F., Hildebrandt, A., Lovden, M., Lindenberger, U., & Wilhelm, O. (2009). Complex span versus updating tasks of working memory: The gap is not that deep. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(4), 1089–1096.  
<https://doi.org/10.1037/a0015730>
- Schmitt, N., & Meara, P. (1997). Researching vocabulary through a word knowledge framework: Word associations and verbal suffixes. *Studies in Second Language Acquisition*, 19(1), 17–36. <https://doi.org/10.1017/S0272263197001022>
- Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the vocabulary levels test. *Language Testing*, 18(1), 55–88.  
<https://doi.org/10.1177/026553220101800103>
- Semel, E., Wiig, E. H., & Secord, W. A. (2006). *Clinical evaluation of language fundamentals* (4th ed.). Pearson Assessment.
- Stansfield, C. W., & Reed, D. J. (2019). The MLAT at 60 years. In Z. Wen, P. Skehan, A. Biedron, S. Li., & R. L. Sparks (Eds.), *Language aptitude: Advancing theory, testing, research, research and practice* (pp. 15–32). Routledge. <https://doi.org/10.4324/9781315122021-2>
- Vandergrift, L., Goh, C. C. M., Mareschal, C., & Tafaghodtari, M. H. (2006). The metacognitive awareness listening questionnaire (MALQ): Development and validation. *Language Learning*, 56(3), 431–462. <https://doi.org/10.1111/j.1467-9922.2006.00373.x>
- Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (1999). *Comprehensive test of phonological processing*. Pro-Ed.
- Wesche, M. B., & Paribakht, T. S. (1996). Assessing second language vocabulary knowledge: Depth versus breadth. *Canadian Modern Language Review*, 53(1), 13–40.  
<https://doi.org/10.3138/cmlr.53.1.13>
- Woodcock, R. W. (1991). *Woodcock language proficiency battery-revised*. Riverside Publishing.



## CHAPTER 9

# L2 speaking

## Theory and research

Jie Gao and April Ginther

Fudan University / Purdue University

This chapter unpacks L2 speaking theories through a cognitive approach, where models such as the Blueprint (Levelt, 1989) and the bilingual speech model of Kormos (2006) are used to explain speech production. From the perspective of language testing and assessment, we further discuss fluency and vocabulary as measurable constructs, which reflects their contribution to the utterance-based strand of L2 speaking assessment research. In addition, we review the categorization framework of cognitive fluency, utterance fluency, and perceived fluency proposed by Segalowitz (2010). This framework emphasizes the operational mechanisms underlying speaking performance, meanwhile highlighting participation from both speakers and listeners. From the perspective of language teaching, we summarize different instructional approaches to L2 speaking, as well as the impact of segmental and suprasegmental features on pronunciation proficiency. We also synthesize the operationalization methods for intelligibility, comprehensibility, and accentedness, which are closely connected with both L2 speaking assessment and pronunciation instruction.

### 1. Introduction

In applied linguistics and language testing, four research themes permeate discussions of the characteristics, instruction, development, and assessment of speaking proficiency: (1) models of first- (L1) and second-language (L2) speech production; (2) discussion and operationalizations of fluency; (3) the place and importance of pronunciation and accent; and (4) the development of the now familiar trinity of accentedness, comprehensibility, and intelligibility. To put these four themes in perspective, we start this chapter with a brief introduction to a cognitive approach to speaking proficiency. As for the first theme, we discuss how speech production has been modeled by reviewing Levelt's (1989) Blueprint model. As for the second theme, we review how fluency has been defined and operationalized in

the literature; we also review, although to a lesser extent, fluency in relation to vocabulary. Our selection of fluency and vocabulary reflects the predominance of utterance-based, fluency-focused research in applied linguistics and testing over the last twenty years. The contribution of vocabulary to the development of fluency represents an important subfield within the broader domain of fluency research. As for the third issue, we review pronunciation and accent while considering their role in pedagogy. Finally, as for the fourth theme, we shift our focus to issues of accentedness, intelligibility, and comprehensibility. Primarily developed in parallel with the research on utterance-based fluency, this line of research largely focuses on perceived fluency and has contributed to substantial changes in our approaches to the instruction and assessment of speaking proficiency.

Why fluency? A standard move in reviews of this kind is to highlight the familiar phrase *a fluent speaker*, which captures the strength of the association between fluency and proficiency for laypersons, but interest in fluency and its relationship to proficiency is shared by second language learners, teachers, researchers, and language testers alike. Fluency, however defined, is accepted as the default for the native speaker (or speakers' L1s), but the exception for non-native speakers (L2s). Acquisition of first language(s) by children through interaction with primary caregivers not only establishes notions of identity and community but also knowledge and facility, prompting Preston (1989) to remark early on that "From a psycholinguistic point of view, nativeness is almost the entire question of SLA" (p. 78). While the notion of "nativeness" has been appropriately challenged on many fronts (see Davies, 2011 for a comprehensive review), and first language speakers' performances are not homogeneous and vary in response to context, topic, and function, understanding the differences across L1s and L2s remains a central pursuit in applied linguistic research.

Addressing disparities in acquisition and achievement suggests, as Segalowitz (2010) remarks, "... acknowledging that people are never inherently incapable of achieving high levels of L2 fluency. One simply must find the right circumstances and conditions to promote learning [...]" (p. 166). Researchers' representations of both L1 and L2 fluency can be understood as attempts to identify the most felicitous conditions for L2 acquisition, teaching, and learning.

Gardner's (1985) *The Mind's New Science: A History of the Cognitive Revolution* traces the origins of cognitive approaches to the 1950s and identifies Miller's (1956) article "The Magical Number Seven" as an important starting point. Gardner (1985) focuses on intersections among psychology, linguistics, and artificial intelligence/computer science. Segalowitz (2010) further articulates the cognitive approach in *Cognitive Bases of Second Language Acquisition*, a volume that has already exerted considerable influence on research of speaking. In it, he argues that adopting a fully-fledged, cognitive perspective is required to address the complexity involved

and “[b]ecause fluency is a multidimensional construct, the multidisciplinary afforded by a cognitive science perspective is indispensable” (p. 25).

Although the value of the cognitive perspective has been recognized, it is not the only approach to examining language proficiency. For example, Luoma (2004) described L2 speaking as a “social” and “situation-based” activity that, as such, requires the inclusion of broad array of perspectives in order to represent it fully. Characteristics of communicative competence (Canale & Swain, 1980; Hymes, 1972), interactional competence (He & Young, 1998; Kramsch, 1986; Young, 1999, 2011), and concretized components as language knowledge and strategic competence (Bachman & Palmer, 1996) have broadened the scope of L2 speaking inquiries and investigations. We recognize the essential nature of speaking as a social activity; however, in this review, we focus more closely on what Segalowitz (2010) has categorized as utterance and perceived fluencies. His discussion of fluency as three complementary components (i.e., cognitive fluency, utterance fluency, and perceived fluency) provides a helpful system for the discussion of the research that has been conducted on the relationship between speaking proficiency and fluency.

## **2. Theme 1: Models explaining L2 speaking and differences between L1 and L2 speech production**

A good place to start is with models of speech production. Among the explanatory models representing fluent speech production, the widely cited and adapted modular (i.e., Blueprint) model proposed by Levelt (1989) has played a central role. This model consists of three phases of speech production: conceptualization, formulation, and articulation. Aspects of these phases function simultaneously, collaboratively, and independently. Speaking begins with a conceptual preparation stage, passes on to the phase of grammatical, morpho-phonological and phonetic encoding, and is finally realized by articulating overt speech. Lemmas, or base morphological forms in the mental lexicon, are activated at the earliest phase, and thus the availability of base forms, mental lexicon, or vocabulary is essential. Levelt argues that lemma activation leads to the surface structure of language output embedded in syntactic structures and further triggers morpho-phonological encoding. In the last step immediately preceding overt speech production, phonetic encoding results in the ultimate articulation of syllabary outcomes.

While developed as an explication of L1 processing and production, the Blueprint model is often applied in L2 domains. De Bot (1992) constructed a bilingual speech production framework based on Levelt (1989), which elaborated classifications for code-switching and cross-linguistic inferences discussed by Nortier (1989). In this framework, intended or situationally motivated code-switching

helped to justify the co-existence of different subsystems. With the conceptualizer phase being non-language specific, first-language speakers possess a single lexicon for lexical element storage. However, the connection between lemma and form characteristics is often not one-to-one for bilingual speakers, echoing arguments that different languages may have their own formulators. Projecting the theories of L1 speech processing and production to L2 speech, Kormos (2006) presented an integrated bilingual speech model that further modified Levelt's Blueprint model. Such models, which emphasize modular processes of speech production and the function of memory, help categorize the main components in representing and then assessing speaking from the speakers' perspective. The final step, articulation, is typically represented and measured in terms of fluency and pronunciation.

### **3. Theme 2: Representation and operationalization of fluency and vocabulary**

When defining a construct to develop a measurement instrument, a typical approach researchers take is to collect information, for example, by reading relevant literature and consulting experts in the field. Then, researchers define the construct and operationalize it so that it can be measured well with their instrument. Although constructs can be defined and operationalized in this manner, this may not seem to be the case for fluency. Segalowitz (2010) argued:

On this point, the literature is at times confusing and disappointing. There are a multitude of meanings for fluency as the term is used in English. Anyone reading the scientific literature will quickly find that researchers provide many different ways of operationalizing what they mean by the term fluency, as will be seen shortly. Moreover, there is no generally accepted model or framework to allow one to think about fluency in a systematic way, although there are some proposals that provide promising elements for such a framework. (p. 22)

Segalowitz' approach derives in part from his observation that "Despite several decades of work, researchers have not discovered universally applicable, objective measures of oral fluency" (p. 39). Indeed, fluency research has been characterized by considerable variation in operationalization and measurement: speed (speech rate, articulation rate, phonation-time ratio, and mean length of run); pausing (number, duration, and location of silent and filled pauses); and repair (false starts, repetitions, and reformulations). A promising element of Segalowitz' approach, as mentioned above, lies in his distinction between "(a) cognitive fluency – the efficiency of operation of the underlying processes responsible for the production of utterances; (b) utterance fluency – the features of utterances that reflect the

speaker's cognitive fluency; and (c) perceived fluency – the inferences listeners make about speakers' cognitive fluency based on their perceptions of their utterance fluency" (p. 240). Examinations of variables associated with utterance fluency as the most valid indicators of cognitive fluency has captured the lion's share of fluency research thus far.

More specifically, Segalowitz (2010) identifies cognitive fluency as the overarching concept and ultimate domain, or construct, of concern. He further categorized speaking performance as utterance and perceived fluencies, highlighting the roles of both speaker and listener. Referring to the operational mechanisms underlying speaking performance, cognitive fluency shares much in common with Lennon's (2000) broad sense of fluency or the ability to marshal resources needed for communicative purposes. Cognitive fluency is described as the ability to process and make adjustments during the speech planning phase, assemble utterances and express individual interpretation with linguistic resources, along with the capability to realize sociolinguistic functions of a language. Both Segalowitz' cognitive fluency and Lennon's broad sense of fluency are associated with "overall proficiency" and correspond to the commonsense notion of being *a fluent speaker*. Based on Fillmore (1979), who describes fluency as a phenomenon that covers pausing, coherence, appropriateness, and creativity, Ginther, Dimova, and Yang (2010) summarized the connections between L1 and L2 fluency in the following:

This broad sense of fluency extends into the domain of second language acquisition where the term is used to refer to mastery and ease of acquired second language performance (Faerch et al., 1984). First and second language domains are thought to converge when second language performance becomes 'nativelike' at high levels of proficiency. (p. 381)

Constructs, however, remain abstractions, and their characteristics must be inferred through observation. Utterance fluency, interpreted as an observable phenomenon, reflects Lennon's (2000) idea of narrow representations of fluency focused on temporal, measurable variables, primarily those involving the quantity, speed, and pausing. Investigations of utterance fluency operationalized in the narrow sense correspond to "the speed and smoothness of oral delivery" (p. 25) and include speech rate, hesitation and pausing phenomena, categorized as speed and repair fluency in Tavakoli and Skehan (2005). The use of temporal representations alone to fully represent fluency has never been argued adequate; nevertheless, the characteristics of temporal aspects of fluency have produced a rich and varied catalogue of research (Kormos & Dénes, 2004; Segalowitz & Freed, 2004; Segalowitz, French, & Guay, 2017; Towell, Hawkins, & Bazergui, 1996).

The narrow sense of fluency has been analyzed as a measurable construct manifested in language task performances. In a series of studies examining the



characteristics of speaker fluency in relation to different tasks, Skehan (1996, 1998) investigated complexity, accuracy, fluency, and lexical measures as overall representations of language performance. From a cognitive perspective, fluency indicates development of performance control with routinized and lexicalized language elements. As summarized by Skehan (2003):

Regarding fluency, it is now increasingly accepted that finer grained analyses of fluency require separate measures of (a) silent (breakdown fluency), (b) reformulation, replacement, false starts, and repetition (repair fluency), (c) speech rate (e.g., words/syllables per minute), and (d) automatization, through measures of length of run. (p. 8)

Tavakoli and Skehan (2005) later combined sub-dimensions of fluency-related measures with speech rate and length of run. Speech rate “refers to how fast and dense the produced language is in terms of the time units” (p. 255). Length of run, as defined in Freed (2000), is continuous speech produced between pauses or hesitations (usually pauses of 250 milliseconds or greater). Skehan (2003) identified the combined sub-dimension as speed fluency.

Indices related to the measurement of utterance fluency, when reorganized with Tavakoli and Skehan’s (2005) framework of speed fluency, breakdown fluency and repair fluency, include: examinations of filled pauses and unfilled pauses (Kormos, 2006; Kowal, O’Connell & Sabin, 1975; Riggenbach, 1991); length of pauses (Goldman-Eisler, 1968; Kormos & Dénes, 2004; Towell et al., 1996); repeat, reformulation, false starts, and other disfluencies in speech (Freed, 1995; Hieke, 1985; Tonkyn, 2012); rate of speech (Freed, 2000; Blake, 2006); mean syllables per run (Ginther, Dimova, & Yang, 2010); and mean length of syllables (Bosker et al., 2013). Utterance fluency measures such as the length of pauses have also been applied in research of conversational cues. They serve as detectable speech signals of speakers’ emotion and attitude. For example, long pauses usually resulted in a perception of refusal from listeners and would also generate a discrepancy of expectations for negative responses (Foti & Roberts, 2016; Roberts & Francis, 2013). Roberts and Norris (2016) also noticed that longer response delay is tolerated more from male speakers than from female speakers.

Based on the cognitive and componential view of L2 speaking proficiency, research has also attempted to disentangle the relationship between utterance fluency and perceived fluency. Segalowitz’s (2010) distinctions between cognitive, utterance, and perceived fluencies expands on Lennon’s broad notion of fluency, and the features of utterance fluency concretize Lennon’s (2000) narrow sense of fluency. Perceived fluency, however, stems from the listeners’ perceptions. Segalowitz’s (2010) introduction of perceived fluency involves “... the inferences listeners make about a speaker’s cognitive fluency based on their perception of utterance fluency”

(p. 48). The separation of perceived fluency from cognitive fluency and utterance fluency fully demonstrates the role of listeners in communication. Utterance fluency manifests through objective measurement but is also inevitably connected with the listener's perceptions. Bosker et al. (2013) investigated the contribution of pauses, speed, and repairs to perceived fluency of L2 Dutch speech. Untrained raters were asked to evaluate overall fluency of L2 speech, which was also measured by sets of acoustic fluency features. Fluency ratings were mostly explained by breakdown fluency and speed fluency measures, followed by repair fluency. Different groups of raters were asked for subjective evaluation of pauses, speed of delivery, and the use of hesitations and corrections. When subjective ratings on specific fluency aspects were used to predict overall fluency, ratings for pausing explained most of the variance, closely followed by those for speed fluency. The major contributing role of speed fluency and breakdown fluency is also attributed to the strong correlation between the two, which aligns with arguments presented in Derwing, Rossiter, Munro and Thomson (2004) and Rossiter (2009).

Further investigation of fluency, or the assessment of speaking in general, is closely connected to vocabulary knowledge as temporal measurement indices (e.g., rate of speech, length of pauses) for gauging utterance fluency only constitute partial operationalization or the concept of fluency. From a cognitive stance, automatic and controlled processing in cognitive fluency have been explored in tasks associated with word recognition (Segalowitz, 2001; Segalowitz, Watson, & Segalowitz, 1995). Word recognition progresses in a stable fashion. Once the task is initiated, it draws little resource from other ongoing activities and cannot be stopped before completion. Faster performance thus involves a mix of automatic and controlled components simultaneously. Reduction in time consumed may be due to a practice effect, which does not necessarily lead to speed increase in a blend of automatic and controlled components. In other words, a characteristic of an automatic process is not only that it is fast, but that it is reliably so across both familiar and unfamiliar tasks. This argument can be applied in evaluating linguistic features representing utterance fluency. Speed alone does not possess enough explanatory power to reliably indicate automatic or controlled processing. Because the measurement of fluency, here a proxy for overall oral proficiency, cannot be confined to examining variables directly related to speakers' delivery speed, investigations from cognitive perspectives have focused on disintegrating the composite structure of speaking proficiency, where linguistic knowledge and processing skills are both emphasized (De Jong et al., 2012; Koizumi & In'nami, 2013). Knowledge of vocabulary has been consistently demonstrated to be a valid predictor of speaking proficiency. Broader vocabulary knowledge contributes to speed fluency, as it empowers speakers at the formulator stage with faster access to lexical resources.

The relationship between vocabulary and fluency has strong precedents, as the contribution of vocabulary knowledge to cognitive fluency has ensured its critical place. Pawley and Syder's (1983) article discussed the puzzle of "nativelike selection" and "nativelike fluency" and argued that lexis, especially lexicalized sentence stems or collocations, play a critical role in accounting for nativelike fluency. They questioned the assumption that generative rules alone could account for L1 linguistic competence, especially with regards to fluency, and argued for the critical role of nativelike selection (fixed, highly frequent phrases or collocations) in support of nativelike fluency. The production of long stretches of language output require not only high levels of automaticity but also selection of expressions that sound natural and idiomatic. Nativelike selection was argued to facilitate both speakers' and listeners' real time production and processing and to assist in explaining the puzzle of nativelike fluency.

The use of vocabulary measures as proxies for speaking proficiency has also been addressed in a number of studies. Lu (2012) examined the effects of lexical diversity on ratings of overall quality of Chinese students' oral narratives and found lexical diversity strongly correlated with the overall quality. Koizumi and In'nami (2013) considered breadth, depth, and processing speed as predictors for the second language speaking proficiency among Japanese students studying English as a foreign language across novice and intermediate levels; they found both depth and breadth to be effective predictors. Similarly, De Jong et al. (2012) found that linguistic measures (primarily vocabulary) and processing speed accounted for 76% of the variance in a measure of communicative success. In addition, limitations in vocabulary have been posited as a source of disfluency, with lexical and inflectional diversity being foundational to fluent speech production and grammatical development (Hilton, 2008; Marsden & David, 2008).

Another pertinent line of research addresses the number of words needed for a language learner to accomplish an everyday speaking task. Estimates of the number of words needed range from 2,000 (Schonell, Meddleton, & Shaw, 1965) to 5,000 (Adolphs & Schmitt, 2003), which would allow 90–98% of words to be known to the speaker. Adolphs and Schmitt (2003) cautioned that these estimates may vary widely across domains of use, from transactional uses requiring the fewest words (2,000) and sociocultural and pedagogical uses requiring many more (5,000). Lower estimates for the successful completion of everyday speaking tasks is associated, in part, with the greater affordances (e.g., gesture, visual aids) that accompany and scaffold speaking performance.

#### 4. Theme 3: The place and importance of pronunciation and accent

Pronunciation stood at the forefront of L2 speaking pedagogy for many years as a critical aspect associated primarily with the work of Robert Lado and Peter Fries. A proponent of structural linguistics, Fries adhered to the belief that language could be appropriately represented by decomposing its structure into sets of discrete phonological, morphological, and grammatical units, which would then lend themselves to instruction (e.g., Fries, 1945). He argued that foreign language learning required, first, the mastery of the sound system, where segmental fidelity to the native speaker norm was considered fundamental, followed by the mastery of morphology, syntax and lexis. Fries' stance was complemented by the work of Lado, a founder of the linguistic sub-discipline of contrastive analysis, who argued that systematic comparison between languages would yield a system of comparative classification and allow identification of areas of ease and difficulty in foreign language learning. Lado's (1957) volume, *Linguistics across cultures: Applied linguistics for language teachers*, outlines a system for comparing a learner's first language to their second with respect to pronunciation/phonology, grammar, vocabulary, and writing. Stockwell, Bowen, and Martin (1965) applied the model in their comparison of Spanish to English and hypothesized eight levels of difficulty given particular phonological, morphological, or syntactic points of comparison; they were subsequently criticized for the selection of 'critical' phonological, morphological, and syntactic contrasts. While these efforts were abandoned, cross-linguistic studies of L1 and L2 fluency features remains as a gap to be filled in order to fully understand the development of L2 proficiency (Segalowitz, 2010). Studies comparing speaking performance across L1s and L2s, which were uncommon with occasional exceptions (e.g., Riazantseva, 2001), have now (re)emerged as a topic of considerable interest and import (e.g., Ciaccio & Clahsen, 2020; Duran-Karaoz & Tavakoli, 2020).

Structuralism and contrastive analysis fit well with behaviorism, the dominant theory of psychology at the time, and the melding of structuralism and behaviorism provided the theoretical underpinnings of the Audiolingual method. Audiolingualism was characterized in the classroom by systematic attention to pronunciation accompanied by intensive oral drilling and memorization of the basic sentence patterns of the target language (Chastain & Woerdehoff, 1968; Larsen-Freeman, 2000; Prator & Celce-Murcia, 1979). Lado's (1961) volume, *Language testing: The construction and use of foreign language tests*, extended structuralism and contrastive analysis to language testing. Again, predicated on the idea that it is possible to identify and then select a representative sample of phonological and structural items based on the similarities and differences between the test taker's L1 and L2, Lado argued that, through intentional selection (sampling) of sets of items targeting particular structural characteristics, difficulty could be controlled,

and tests could be constructed whose scores represented the test taker's mastery of the L2. Providing the basis for the first version of the Test of English as a Foreign Language (TOEFL), Lado's assumptions have remained an influential underpinning to the selection of grammar items in L2 proficiency tests (for a summary, see Ginther & McIntosh, 2018; Spolsky, 1995).

As an instructional and testing approach, audiolingualism had at least two weaknesses: the emphasis on drill and rote memorization, and the insistence on fidelity to native speaker norms. In the 1960s, the structuralist approach and its instructional offshoot, audiolingualism, was under attack from all sides: from theoretical linguists who championed cognitive approaches over structuralist, data-driven approaches to linguistic description (Chomsky, 1965; Gardener, 1985), and from applied linguists who championed broader approaches to language learning and testing (Bachman, 1990; Bachman & Palmer, 1996; Canale, 1983; Canale & Swain, 1980). Communicative language teaching and its focus on ability for use (Gumperz, 1972) began its ascent and would become the dominant instructional paradigm.

Audiolingualism was replaced largely by the belief that language learners could and would acquire their L2s through active use and social exposure; however, this belief has been called into question as researchers stressed the importance of explicit, consistent form-focused practice (De Jong, 2014; Schmitt, 2008; Segalowitz, 2010). Calls to reconsider the value of practice suggest that old-school perspectives are beginning to resonate with contemporary approaches to instruction.

Note that structuralists assumed that learner difficulty with L2 acquisition could be captured through comparison across linguistic systems. Nevertheless, this assumption did not seem to be reasonable given that there is a countless number of combinations of all possible L1 and L2 contrasts (e.g., English-Chinese; English-Spanish; English-Vietnamese; Chinese-Vietnamese...). However, just as the prospects for practice has re-emerged, so too have calls for comparisons of phonological and fluency characteristics across speakers' L1s and L2s. The fidelity to a native norm, especially with reference to segmental phonology, was challenged for being too narrow, unrealistic, and inappropriate (Bosker et al., 2014; Jenkins, 1998). Nevertheless, such comparisons may not only be necessary to provide appropriate learner baselines for subsequent L2 learning outcomes but also for complete explications of proficiency and fluency (Bent & Bradlow, 2003; Bosker et al., 2014; Davies, 2011; Segalowitz, 2010).

Following the discussion so far, one may wonder to what extent pronunciation can be improved through instruction. Saito and Plonsky's (2019) meta-analysis of the effects of pronunciation instruction answers this question by presenting a three-tiered model for conceptualizing instructed L2 pronunciation proficiency. Their parameters contrasted "(a) the constructs being focused on (global vs.

specific), (b) the scoring method (human raters vs. acoustic analyses), and (c) the type of knowledge being elicited (controlled vs. spontaneous)” (p. 669). They concluded:

...pronunciation teaching can be beneficial at a controlled level, as providing explicit phonetic information enables learners to notice and practice the accurate production of L2 segmental, syllabic, prosodic and temporal features in a careful fashion. On the other hand, our results cast doubt on (a) whether and to what degree pronunciation teaching can subsequently lead to perceptible changes in learners’ relatively spontaneous and automatized pronunciation performance; and (b) whether pronunciation teaching can ultimately impact global pronunciation proficiency (e.g., comprehensibility). In general, when it comes to the analysis of specific segmental and suprasegmental L2 pronunciation accuracy, subjective measures (expert judgements) seem to involve more variability than objective measures (acoustic analyses), especially when they are applied to spontaneous speech tasks. (p. 696)

In terms of both research and practice, we seem to be coming full circle and arriving at a place where a synthesis of approaches (Kuhn, 1962) is possible. A closer look at the shift from form-focused pronunciation to accentedness, comprehensibility, and intelligibility is promising in this regard, particularly with respect to what was intended, what was actually gained, and where we have arrived.

## 5. Theme 4: Accentedness, comprehensibility, and intelligibility

Given both the historical precedents and the strength of the relationship between accent and identity, we will begin with a discussion of accentedness. Being a fluent speaker of at least one language is so familiar a part of our identities that listeners reliably recognize an accent different from their own after listening to as little as 30 milliseconds of recorded speech (Flege, 1984). Listeners are also able to reliably identify L1 and L2 speakers of languages that they do not speak (Major, 2007), and the presence of an accent has been found to affect language processing strategies of children as young as 16 months of age (Weatherhead & White, 2018). As Scovel (1988) remarked: “accent features are exceptionally salient, and as a result we’re very good at detecting perceived outsiders on the basis of their speech patterns” (p. 477).

Who we are, where we are from, and where we feel we belong are marked by how we speak and sound to others. Research in social psychology has found that listeners attribute a variety of characteristics to speakers based on accent, including nationality, regional membership, ethnicity, and social class (Labov, 2006), intelligence (Lambert, Hodgson, Gardner, & Fillenbaum, 1960), social desirability

(Kinzler & DeJesus, 2013), and suitability for employment (Kalin & Rayko, 1978). Being an L2 speaker increases the potential effect of negative attributions as perceptions of a foreign accent have been found to influence listeners' beliefs about L2 speakers' general communication skills (Hosoda, Stone-Romero, & Walter, 2007) and overall competence (Nelson, Signorella, & Botti, 2016). Furthermore, speakers identified as having foreign accents have been assumed less credible (Bourdieu & Thompson, 1991; Lev-Ari & Keysar, 2010; Livingston, Schilpzand, & Erez, 2017), less educated (Fraser & Kelly, 2012), and less intelligent (Anderson et al., 2007; Fuertes, Potere, & Ramirez, 2002). Given the association of identity and accent, along with the potential bias of our accent-based attributions, it is only natural that applied linguists, language testers, and language teachers have paid a great deal of attention to accentedness.

Few researchers have contributed more to the development of the concepts of accentedness, comprehensibility, and intelligibility than Tracey M. Derwing and Murray J. Munro (Derwing & Munro, 1997, 2005, 2009; Munro & Derwing, 1995a, 1995b, 2011). While accentedness, comprehensibility, and intelligibility were long present in the literature (Abercrombie, 1949; Morley, 1994; Pennington & Richards, 1986), Derwing and Munro developed these concepts in a series of related studies that left no stone unturned in their attempts to clarify the relationships involved. Discussions of accentedness now go hand-in-hand with the concepts of comprehensibility and intelligibility. In fact, the association between accentedness, comprehensibility, and intelligibility has so permeated the discussion of oral proficiency that it is now difficult to find a currently-used oral proficiency scale that explicitly refers to pronunciation. For all intents and purposes, the term has been replaced by comprehensibility and/or intelligibility.

Derwing and Munro (2005) operationalize accentedness as strength of the perception of difference from a local norm (from no accent to a very strong accent), comprehensibility as listener processing ease (from extremely easy to impossible to understand), and intelligibility as "the extent to which a listener actually understands an utterance" (p. 385), thus clearly emphasizing Segalowitz' (2010) perceived fluency.

Accentedness and comprehensibility are typically estimated through the use of 9-point Likert scales, while intelligibility is usually associated with more explicit means of estimation, such as percent/number of correctly identified words and phrases or performance on true/false and listening comprehension questions. While highly influential, Derwing and Munro's operationalizations of comprehensibility and intelligibility have not led to consensus in the use of the terms nor on how they interact. Part of the problem lies in the overlap between the operationalizations of comprehensibility, as the listener's ease of processing, and intelligibility, as the degree of actual comprehension. In fact, the terms – comprehensibility and intelligibility – have appeared to be used interchangeably in the literature. Despite



the extended efforts of Derwing and Munro, comprehensibility and intelligibility remain difficult to distinguish.

Yet, this difficulty does not seem to undermine the importance of their research. Perhaps the most influential contribution of their research has been to shift the instructional focus from the goal of achieving some facsimile of a native-like speech to a more realistic goal of accented but comprehensible and/or intelligible speech. Derwing and Munro referenced studies in which pronunciation instruction has been found to have a positive effect on both intelligibility and comprehensibility, and they argue that prioritized pronunciation instruction, or “a conceptualization of intelligibility that assists teachers in setting priorities” and “empirical evidence that identifies effective practices” (Munro & Derwing, 2011, p. 317), should focus on helping learners produce intelligible speech. Derwing and Munro (2009) also explained:

If time is spent on something that doesn't affect intelligibility or comprehensibility (such as the infamous interdental fricatives in English), something that really does matter will be neglected. Evidence is accumulating that what's important are the macroscopic things, including general speaking habits, volume, stress, rhythm, syllable structure and segmentals with a high functional load. (pp. 482–483)

As Harmer (1991) stated, “our aim should be to make sure that students can always be understood to say what they want to say. They will need good pronunciation for this, though they may not need to have perfect accents” (p. 22). Derwing and Munro's (2009) suggestion that instruction shift to intelligibility and comprehensibility to include features beyond segmental fidelity to native-like speech has been complemented by other studies that have examined the contributions of both segmental and suprasegmental aspects of speech. Investigations of intelligibility and comprehensibility in relation to speakers' overall oral language proficiency, support inclusion of an expanded set of variables – features of speech production that extend beyond segmentals.

The effects of segmental features on pronunciation perception are examined in combination with suprasegmental elements and fluency variables. Anderson-Hsieh, Johnson, and Koehler (1992) explored the relationship between native speakers' judgements of pronunciation in three areas: prosody, segmentals, and syllable structure. The strongest relationship was found between prosodic features and pronunciation ratings. Trofimovich and Baker (2006) investigated L2 acquisition of suprasegmentals by analyzing stress timing and tonal peak alignment in adult L2 speech, together with fluency measures of speech rate, pause frequency, and pause duration. Correlational results found speakers' approximation of English stress timing and fluency measures both impacted listeners' evaluation of speakers' accentedness. The predictive power of suprasegmental factors, however, was not as

strong as that for standard fluency measures (e.g., speech rate and mean length of run). What remains problematic about these arguments is that prosodic characteristics of speech (rhythm, timing, intonation) overlap with many fluency variables (speed, pausing, and pause placement or parsing). Nevertheless, the shift from segmental to suprasegmentals is important, despite the fact that it complicates the representation and operationalization of all of the constructs involved.

The contributions of segmental and suprasegmental features to accentedness and intelligibility were also examined in Winters and O'Brien (2012). Intonation contours and syllable duration were interchangeably mapped onto L2 speech and native speech of English and German. Listeners then completed cloze tests and comprehension tests to assess intelligibility. L2 segmental production was shown to have stronger effect on accentedness perception. In addition, results for intelligibility tasks demonstrated an interaction between shared speaker and listener L1 background (i.e., an interlanguage speech intelligibility effect) (Bent & Bradlow, 2003; Hahn, 2004). Isaacs and Trofimovich (2012) examined listeners' comprehensibility ratings in relation to 19 quantitative speech measures represented by segmental, suprasegmental, fluency, lexical, grammatical, and discourse-level variables, which were then correlated with three L1 English listeners' scalar judgements of L2 speech comprehensibility. Correlational results found L2 comprehensibility ratings related to a wide range of variables not restricted to the domain of phonology and fluency. Reports from three experienced raters were also collected for more detailed description of the features they found most noticeable. Their comments highlighted speakers' word stress, grammar, vocabulary, and fluency along with discourse structure and context representation. Raters' familiarity with the speakers' L1s was also mentioned as an influence on the comprehensibility ratings.

A developing line of speaking research examines the effects of different combinations of speakers' and listeners' L1s and L2s on perceptions of speaking performance. Reminiscent of Lado's (1964) call for comparisons between L1s and L2s to predict a learners' areas of difficulty, a growing number of studies have examined how listeners with different language backgrounds process speech. Despite the well-received argument that strong accents do not necessarily lead to reported listener difficulty (e.g., comprehensibility), some studies are finding that L1 listeners appear to process L1 and L2 speech differently (Gibson et al., 2017), and that native speaker "disfluencies" (e.g., pausing) are strategic and may ease listener processing. As Bosker et al. (2014) commented, "It has been previously found that native disfluencies may help the listener in word recognition (Corley & Hartsuiker, 2011), in sentence integration (Corley, MacGregor, & Donaldson, 2007), and in reference resolution (Arnold et al., 2007)" (p. 609).

However, there appears to be an initial processing cost of accentedness for listeners. Ockey, Papageorgiou and French (2016) and Ockey and French (2016)

discussed performance effects on subjects' performance on listening comprehension items on a monologic task and then on an interactive lecture and found consistent debilitation of performance when even slightly accented speech was included in the input. Of their second set of findings, they explained that even accents judged to be light and completely comprehensible can influence test takers' understanding of interactive lectures. Strength of accent is a variable that needs to be carefully considered in listening comprehension assessment.

Studies examining the extent to which listeners adjust to speakers' accentedness have reported mixed but encouraging results for both listeners and test takers. That is, the cost of accentedness may be overcome through exposure. Listeners experience a lower degree of cognitive burden after spending certain amount of time with a particular accent. Gass and Veronis (1984) reported that listener processing costs diminish, often rapidly, with increased familiarity. Floccia, Goslin, Girard, and Konopczynski (2006), along with Clarke and Garrett (2004), also found evidence of initial processing difficulty followed by processing adjustment/normalization after as few as 2 to 4 utterances. In a review of validity concerns for speaking assessments, Harding (2017) argued that the time has come for listener variables to be explicitly considered in construct definitions. Raters are invited to take a more active role in the drafting and application of speaking assessment rubrics, especially when a more clearly defined pronunciation scale is needed. Raters' familiarity and attitude towards speakers' pronunciation are also concerns for the generalizability inference in speaking assessment. By considering these listener variables in relation to speaking, researchers can better understand how they function with actual oral performance. Accumulating the findings would help researchers better define speaking and operationalize it.

## 6. Might practice make perfect?

It is a challenge to implement a multidisciplinary and cognitive approach to understanding L2 speaking. However, the development of fluency, vocabulary, and pronunciation research provides a strong foundation for integrating cognitive models. The categorization of fluency into complementary components of cognitive fluency, utterance fluency, and perceived fluency allows contrasts and comparisons across different approaches. Speakers' use of vocabulary is also an essential element of speech formulation. Studies related to accentedness, intelligibility, and comprehensibility push pronunciation perception beyond the investigation of segmental features only, enriching the understanding of articulation and listener comprehension.

The time has come to re-evaluate the status and importance of practice. While hardly a return to the heyday of audiolingualism, with its emphasis on skill and

drill, many researchers have suggested that the importance of practice has been underestimated. De Jong (2014) explained that speaking, when examined through cognitive approaches, requires practice before it becomes automatic and unconscious. Increased proceduralization of linguistic knowledge through explicit practice has also been advocated by Schmitt (2008) and Snellings, van Gelderen, and Gloppe (2002), who acknowledged improvements in lexical retrieval speed after explicit attention was given to vocabulary exercises over a specific period of time. Segalowitz (2010) agreed that, from the perspective of cognitive science, learning activities in the classroom need to consider “the whole set of mental processes involved in the planning, assembling, and execution of the speech act, and this must occur within genuinely communicative contexts and naturally repetitive at the same time” (p. 176).

These studies, reviews, and comments all suggest that the next challenge for researchers and instructors concerned with the development of speaking proficiency will be to synthesize the perspectives of the past with our current understandings: How can the development of speaking proficiency be facilitated through an instructional focus on vocabulary and fluency? How can we address similarities and differences with respect to intelligibility and comprehensibility as we hone rubrics that include pronunciation? How can we assess L2 speaking in conversational settings? Such findings will prove useful to the continued development of theories and practices regarding L2 speaking proficiency.

## References

- Abercrombie, D. (1949). Teaching pronunciation. *English Language Teaching*, 3, 113–122.  
<https://doi.org/10.1093/elt/III.5.113>
- Adolphs, S., & Schmitt, N. (2003). Lexical coverage of spoken discourse. *Applied Linguistics*, 24(4), 425–438. <https://doi.org/10.1093/applin/24.4.425>
- Anderson-Hsieh, J., Johnson, R., & Koehler, K. (1992). The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure. *Language Learning*, 42(4), 529–555.  
<https://doi.org/10.1111/j.1467-1770.1992.tb01043.x>
- Anderson, A., Downs, S. D., Faucette, K., Griffin, J., King, T., & Woolstenhulme, S. (2007). How accents affect perception of intelligence, physical attractiveness, and trustworthiness of Middle-Eastern-, Latin-American-, British, and Standard American-English-accented speakers. *BYU Undergraduate Journal of Psychology*, 3, 5–11.
- Arnold, J. E., Hudson Kam, C. L., & Tanenhaus, M. K. (2007). If you say thee uh you're describing something hard: The on-line attribution of disfluency during reference comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(5), 914–930. <https://doi.org/10.1037/0278-7393.33.5.914>
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford University Press.
- Bent, T., & Bradlow, A. R. (2003). The interlanguage speech intelligibility benefit. *Journal of the Acoustical Society of America*, 114(3), 1600–1610. <https://doi.org/10.1121/1.1603234>
- Blake, C. (2006). The potential of text-based Internet chats for improving ESL oral fluency. (Unpublished doctoral dissertation). Purdue University, West Lafayette, IN.
- Bosker, H. R., Pinget, A.-F., Quené, H., Sanders, T., & De Jong, N. H. (2013). What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing*, 30(2), 159–175. <https://doi.org/10.1177/0265532212455394>
- Bosker, H. R., Quené, H., Sanders, T., & De Jong, N. (2014). The perception of fluency in native and nonnative speech, *Language Learning*, 64(3), 579–614. <https://doi.org/10.1111/lang.12067>
- Bourdieu, P., & Thompson, J. B. (1991). *Language and symbolic power*. Harvard University Press.
- Canale, M. (1983). From communicative competence to communicative language pedagogy. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 2–27). Longman.
- Canale, S., & Swain, M. (1980). Theoretical bases of communicative language approaches to second language teaching and testing. *Applied Linguistics*, 1(1), 1–47. <https://doi.org/10.1093/applin/1.1.1>
- Chastain, K. D., & Woerdehoff, F. J. (1968). A methodological study comparing the audio-lingual habit theory and the cognitive code-learning theory. *The Modern Language Journal*, 52(5), 268–279. <https://doi.org/10.1111/j.1540-4781.1968.tb01905.x>
- Chomsky, N. (1965). *Aspects of the theory of syntax*. The MIT Press.
- Ciaccio, L. A. & Clahsen, H. (2020). Variability and consistency in first and second language processing: A masked morphological priming study on prefixation and suffixation. *Language Learning*, 70(1), 103–136. <https://doi.org/10.1111/lang.12370>
- Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America*, 116(6), 3647–3658. <https://doi.org/10.1121/1.1815131>
- Corley, M., MacGregor, L. J., & Donaldson, D. I. (2007). It's the way that you, er, say it: Hesitations in speech affect language comprehension. *Cognition*, 105(3), 658–668. <https://doi.org/10.1016/j.cognition.2006.10.010>
- Corley, M., & Hartsuiker, R. J. (2011). Why um helps auditory word recognition: the temporal delay hypothesis. *Plos One*, 6: e19792. <https://doi.org/10.1371/journal.pone.0019792>
- Davies, A. (2011). Does language testing need the native speaker? *Language Assessment Quarterly*, 8(3), 291–308. <https://doi.org/10.1080/15434303.2011.570827>
- De Bot, K. (1992). A bilingual production model: Levelt's "speaking" adapted. *Applied Linguistics*, 13(1), 1–24. <https://doi.org/10.1093/applin/13.1.1>
- De Jong, N. H. (2014). Teaching speaking. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Blackwell. <https://doi.org/10.1002/9781405198431.wbeal1437>
- De Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition*, 34(1), 5–34. <https://doi.org/10.1017/S0272263111000489>
- Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition*, 20(1), 1–16. <https://doi.org/10.1017/S0272263197001010>
- Derwing, T. M., & Munro, M. J. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly*, 39(3), 379–397. <https://doi.org/10.2307/3588486>
- Derwing, T. M., & Munro, M. J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching*, 42(4), 476–490. <https://doi.org/10.1017/S026144480800551X>

- Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgements on different tasks. *Language Learning*, 54(4), 655–679. <https://doi.org/10.1111/j.1467-9922.2004.00282.x>
- Duran-Karaoz, Z., & Tavakoli, P. (2020). Predicting L2 fluency from L1 fluency behavior: The case of L1 Turkish and L2 English speakers. *Studies in Second Language Acquisition*, 42(4), 671–695. <https://doi.org/10.1017/S0272263119000755>
- Faerch, K., Haastруп, K., & Phillipson, R. (1984). *Learner language and language learning*. Multilingual Matters.
- Fillmore, C. (1979). On fluency. In C. Fillmore, D. Kempler, & W. Wang (Eds.), *Individual differences in language ability and language behavior* (pp. 85–101). Academic Press. <https://doi.org/10.1016/B978-0-12-255950-1.50012-3>
- Flege, J. E. (1984). The detection of French accent by American listeners. *Journal of the Acoustical Society of America*, 76(3), 692–707. <https://doi.org/10.1121/1.391256>
- Floccia, C., Goslin, J., Girard, F., & Konopczynski, G. (2006). Does a regional accent perturb speech processing? *Journal of Experimental Psychology Human Perception and Performance*, 32(5), 1276–1293. <https://doi.org/10.1037/0096-1523.32.5.1276>
- Foti, D. & Roberts, F. (2016). The neural dynamics of speech perception: Dissociable networks for processing linguistic content and monitoring speaker turn-taking. *Brain & Language*, 157–158, 63–71. <https://doi.org/10.1016/j.bandl.2016.05.001>
- Fraser, C., & Kelly, B. F. (2012). Listening between the lines: Social assumptions around foreign accents. *Australian Review of Applied Linguistics*, 35(1), 74–93. <https://doi.org/10.1075/ara1.35.1.04fra>
- Freed, B. (1995). What makes us think that students who study abroad become fluent? In B. Freed (Ed.) *Second language acquisition in a study abroad context* (pp. 123–148). John Benjamins. <https://doi.org/10.1075/sib1.9.09fre>
- Freed, B. (2000). Is fluency, like beauty, in the eyes (and ears) of the beholder? In H. Riggensbach (Ed.), *Perspectives on fluency* (pp. 243–265). University of Michigan Press.
- Fries, C. (1945). *Teaching and learning English as a foreign language*. University of Michigan Press.
- Fuertes, J. N., Potere, J. C., & Ramirez, K. Y. (2002). Effects of speech accents on interpersonal evaluations: Implications for counseling practice and research. *Cultural Diversity & Ethnic Minority Psychology*, 8(4), 346–356. <https://doi.org/10.1037/1099-9809.8.4.347>
- Gardner, H. (1985). *The mind's new science: A history of the cognitive revolution*. Basic Books.
- Gass, S., & Varonis, E. M. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language Learning*, 34(1), 65–89. <https://doi.org/10.1111/j.1467-1770.1984.tb00996.x>
- Gibson, E., Tan, C., Futrell, R., Mahowald, K., Konieczny, L., Hemforth, B., & Fedorenko, E. (2017). Don't underestimate the benefits of being misunderstood. *Psychological Science*, 28(6), 703–712. <https://doi.org/10.1177/0956797617690277>
- Ginther, A., & McIntosh, K. (2018). Language testing and assessment. In A. Phakiti, P. De Costa, L. Plonsky, & S. Starfield (Eds.), *The Palgrave handbook of applied linguistics research methodology* (pp. 845–867). Palgrave. [https://doi.org/10.1057/978-1-137-59900-1\\_39](https://doi.org/10.1057/978-1-137-59900-1_39)
- Ginther, A., Dimova, S., & Yang, R. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implication for automated scoring. *Language Testing*, 27(3), 379–399. <https://doi.org/10.1177/0265532210364407>
- Goldman-Eisler, F. (1968). *Psycholinguistics: Experiments in spontaneous speech*. Academic Press.
- Gumperz, J. J. (1972). The communicative competence of bilinguals: Some hypotheses and suggestions for research. *Language in Society*, 1(1), 143–154. <https://doi.org/10.1017/S0047404500006606>



- Hahn, L. D. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly*, 38(2), 201–223. <https://doi.org/10.2307/3588378>
- Harding, L. (2017). Validity in pronunciation assessment. In O. Kang & A. Ginther (Eds.), *Assessment in second language pronunciation* (pp. 30–48). Routledge. <https://doi.org/10.4324/9781315170756-3>
- Harmer, J. (1991). *The practice of English language teaching*. Longman.
- He, A. W., & Young, R. (1998). Language proficiency interviews: A discourse approach. In R. Young & A. W. He (Eds.), *Talking and testing: Discourse approaches to the assessment of oral proficiency* (pp. 1–24). John Benjamins. <https://doi.org/10.1075/sibill.14.02he>
- Hieke, A. E. (1985). A componential approach to oral fluency evaluation. *The Modern Language Journal*, 69(2), 135–142. <https://doi.org/10.1111/j.1540-4781.1985.tb01930.x>
- Hilton, H. (2008). The link between vocabulary knowledge and spoken L2 fluency. *Language Learning Journal*, 36(2), 153–166. <https://doi.org/10.1080/09571730802389983>
- Hosoda, M., Stone-Romero, E. F., & Walter, J. N. (2007). Listeners' cognitive and affective reactions to English speakers with standard American English and Asian accents. *Perceptual and Motor Skills*, 104(1), 307–326. <https://doi.org/10.2466/pms.104.1.307-326>
- Hymes, D. (1972). On communicative competence. In J. B. Pride & A. Holmes (Eds.), *Sociolinguistics: Selected readings* (pp. 269–293). Penguin.
- Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition*, 34(3), 475–505. <https://doi.org/10.1017/S0272263112000150>
- Jenkins, J. (1998). Which pronunciation norms and models for English as an International Language? *ELT Journal*, 52(2), 119–126. <https://doi.org/10.1093/elt/52.2.119>
- Kalin, R., & Rayko, K. (1978). Discrimination in evaluative judgements against foreign-accented job candidates. *Psychological Reports*, 43(3), 1203–1209. <https://doi.org/10.2466/pro.1978.43.3f.1203>
- Kinzler, K. D., & DeJesus, J. M. (2013). Northern = smart and Southern = nice: The development of accent attitudes in the United States. *The Quarterly Journal of Experimental Psychology*, 66(6), 1146–1158. <https://doi.org/10.1080/17470218.2012.713695>
- Koizumi, R., & In'nami, Y. (2013). Vocabulary knowledge and speaking proficiency among second language learners from novice to intermediate levels. *Journal of Language Teaching and Research*, 4(5), 900–913. <https://doi.org/10.4304/jltr.4.5.900-913>
- Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners, *System*, 32(2), 146–164. <https://doi.org/10.1016/j.system.2004.01.001>
- Kormos, J. (2006). *Speech production and second language acquisition*. Lawrence Erlbaum Associates.
- Kowal, S., O'Connell, D., & Sabin, E. (1975). Development of temporal patterning and vocal hesitations in spontaneous narratives. *Journal of Psycholinguistic Research*, 4, 195–207. <https://doi.org/10.1007/BF01066926>
- Kramsch, C. (1986). From language proficiency to interactional competence. *The Modern Language Journal*, 70(4), 366–372. <https://doi.org/10.1111/j.1540-4781.1986.tb05291.x>
- Kuhn, T. (1962). *The structure of scientific revolution*. University of Chicago Press.
- Labov, W. (2006). *The social stratification of English in New York City* (2nd ed.). Cambridge University Press. <https://doi.org/10.1017/CBO9780511618208>
- Lado, R. (1957). *Linguistics across cultures: Applied linguistics for language teachers*. University of Michigan Press.



- Lado, R. (1961). *Language testing: The construction and use of foreign language tests*. McGraw-Hill.
- Lado, R. (1964). *Language teaching: A scientific approach*. McGraw-Hill.
- Lambert, W. E., Hodgson, R., Gardner, R., & Fillenbaum, S. (1960). Evaluational reactions to spoken languages. *The Journal of Abnormal and Social Psychology*, 60(1), 44–51.  
<https://doi.org/10.1037/h0044430>
- Larsen-Freeman, D. (2000). *Techniques and principles in language teaching*. Oxford University Press.
- Lennon, P. (2000). The lexical element in spoken second language fluency. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 25–42). University of Michigan Press.
- Lev-Ari, S., & Keysar, B. (2010). Why don't we believe nonnative speakers? The influence of accent on credibility. *Journal of Experimental Social Psychology*, 46(6), 1093–1096.  
<https://doi.org/10.1016/j.jesp.2010.05.025>
- Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. The MIT Press.
- Livingston, B. A., Schilpzand, P., & Erez, A. (2017). Not what you expected to hear: Accented messages and their effect on choice. *Journal of Management*, 43(3), 804–833.  
<https://doi.org/10.1177/0149206314541151>
- Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, 96(2), 190–208. [https://doi.org/10.1111/j.1540-4781.2011.01232\\_1.x](https://doi.org/10.1111/j.1540-4781.2011.01232_1.x)
- Luoma, S. (2004). *Assessing speaking*. Cambridge University Press.  
<https://doi.org/10.1017/CBO9780511733017>
- Major, R. C. (2007). Identifying a foreign accent in an unfamiliar language. *Studies in Second Language Acquisition*, 29(4), 539–556. <https://doi.org/10.1017/S0272263107070428>
- Marsden, E., & David, A. (2008). Vocabulary use during conversation: A cross-sectional study of development from year 9 to year 13 among learners of Spanish and French. *Language Learning Journal*, 36(2), 181–198. <https://doi.org/10.1080/09571730802390031>
- Miller, G. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, 63(2), 81–97.  
<https://doi.org/10.1037/h0043158>
- Morley, J. (1994). *Pronunciation pedagogy and theory: New views, new directions*. TESOL.
- Munro, M. J., & Derwing, T. M. (1995a). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, 45(1), 73–97.  
<https://doi.org/10.1111/j.1467-1770.1995.tb00963.x>
- Munro, M. J., & Derwing, T. M. (1995b). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech*, 38(3), 289–306.  
<https://doi.org/10.1177/002383099503800305>
- Munro, M. J., & Derwing, T. M. (2011). The foundations of accent and intelligibility in pronunciation research. *Language Teaching*, 44(3), 316–327. <https://doi.org/10.1017/S0261444811000103>
- Nelson, L. R., Signorella, M. L., & Botti, K. G. (2016). Accent, gender, and perceived competence. *Hispanic Journal of Behavioral Sciences*, 38(2), 166–185.  
<https://doi.org/10.1177/0739986316632319>
- Nortier, J. (1989). Dutch and Moroccan-Arabic in contact: Code-switching among Moroccans in the Netherlands (Unpublished doctoral dissertation). University of Amsterdam.
- Ockey, G., & French, R. (2016). From one to multiple accents on a test of L2 listening comprehension. *Applied Linguistics*, 37(5), 693–715. <https://doi.org/10.1093/applin/amu060>
- Ockey, G., Papageorgiou, S., & French, R. (2015). Effects of strength of accent on an L2 interactive lecture listening comprehension test. *The International Journal of Listening*, 30(1/2), 84–98. <https://doi.org/10.1080/10904018.2015.1056877>

- Pawley, A., & Syder, F. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. Richards & R. Schmidt (Eds.), *Language and communication* (pp. 191–226). Longman.
- Pennington, M., & Richards, J. (1986). Pronunciation revisited. *TESOL Quarterly*, 20(2), 207–225. <https://doi.org/10.2307/3586541>
- Prator, C. H., & Celce-Murcia, M. (1979). An outline of language teaching approaches. In M. Celce-Murcia & L. McIntosh (Eds.), *Teaching English as a second or foreign language*. Newbury House.
- Preston, D. (1989). *Sociolinguistics and second language acquisition*. Blackwell.
- Riazantseva, A. (2001). Second language proficiency and pausing: A study of Russian speakers of English. *Studies in Second Language Acquisition*, 23(4), 497–526. <https://doi.org/10.1017/S027226310100403X>
- Riggenbach, H. (1991). Towards an understanding of fluency: A microanalysis of nonnative speaker conversations. *Discourse Processes*, 14(4), 423–441. <https://doi.org/10.1080/01638539109544795>
- Roberts, F. & Francis, A. L. (2013). Identifying a temporal threshold of tolerance for silent gaps after requests. *The Journal of the Acoustic Society of America*, 133(6), EL471–477. <https://doi.org/10.1121/1.4802900>
- Roberts, F., & Norris, A. (2016). Gendered expectations for “agreeableness” in responses to requests and opinions. *Communication Research Reports*, 33(1), 16–23. <https://doi.org/10.1080/08824096.2015.1117437>
- Rossiter, M. J. (2009). Perception of L2 fluency by native and non-native speakers of English. *The Canadian Language Review/La Revue Canadienne des Langues Vivantes*, 65(3), 395–412. <https://doi.org/10.3138/cmlr.65.3.395>
- Saito, K., & Plonsky, L. (2019). Effects of second language pronunciation teaching revisited: A proposed measurement framework and meta-analysis. *Language Learning*, 69(3), 652–708. <https://doi.org/10.1111/lang.12345>
- Schmitt, N. (2008). Review article: Instructed second language vocabulary learning. *Language Teaching Research*, 12(3), 329–363. <https://doi.org/10.1177/1362168808089921>
- Schonell, F. J., Meddleton, I. G., & Shaw, B. A. (1965). *A study of the oral vocabulary of adults*. University of Queensland Press.
- Scovel, T. (1988). *A time to speak: A psycholinguistic investigation into the critical period for human speech*. Harper and Row.
- Segalowitz, N. (2001). Automaticity and attentional skill in fluent performance. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 200–219). University of Michigan Press.
- Segalowitz, N. (2010). *Cognitive bases of second language fluency*. Routledge. <https://doi.org/10.4324/9780203851357>
- Segalowitz, N., & Freed, B. F. (2004). Context, contact, and cognition in oral fluency acquisition: Learning Spanish in at home and study abroad context. *Studies in Second Language Acquisition*, 26(2), 173–199. <https://doi.org/10.1017/S0272263104262027>
- Segalowitz, N., French, L., & Guay, J. (2017). What features characterize adult second language utterance fluency and what do they reveal about fluency gains in short-term immersion? *Canadian Journal of Applied Linguistics/Revue Canadienne de Linguistique Appliquée*, 20, 90–116.
- Segalowitz, N., Watson, V., & Segalowitz, S. (1995). Vocabulary skill: Single-case assessment of automaticity of word recognition in a time lexical decision task. *Second Language Research*, 11(2), 121–136. <https://doi.org/10.1177/026765839501100204>

- Skehan, P. (1996). Second language acquisition and task-based instruction. In J. Willis & D. Willis (Eds.), *Challenge and change in language teaching* (pp. 17–30). Heinemann.
- Skehan, P. (1998). *A cognitive approach to language learning*. Oxford University Press.
- Skehan, P. (2003). Task-based instruction. *Language Teaching*, 36(1), 1–14.  
<https://doi.org/10.1017/S026144480200188X>
- Snellings, P., van Gelderen, A., & de Glopper, K. (2002). Lexical retrieval: An aspect of fluent second language production that can be enhanced. *Language Learning*, 52(4), 723–754.  
<https://doi.org/10.1111/1467-9922.00202>
- Spolsky, B. (1995). *Measured words: The development of objective language testing*. Oxford University Press.
- Stockwell, R., Bowen, J., & Martin, J. (1965). *The grammatical structure of English and Spanish*. Chicago University Press.
- Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure, and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239–276). John Benjamins. <https://doi.org/10.1075/llt.11.15tav>
- Tonkyn, A. (2012). Measuring and perceiving changes in oral complexity, accuracy and fluency. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 proficiency and performance: Complexity, accuracy and fluency in SLA* (pp. 221–245). John Benjamins.  
<https://doi.org/10.1075/llt.32.10ton>
- Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics*, 17(1), 84–119. <https://doi.org/10.1093/applin/17.1.84>
- Trofimovich, P., & Baker, W. (2006). Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies of Second Language Acquisition*, 28(1), 1–30. <https://doi.org/10.1017/S0272263106060013>
- Weatherhead, D., & White, K. S. (2018). And then I saw her race: Race-based expectations affect infants' word processing. *Cognition*, 177, 87–97.  
<https://doi.org/10.1016/j.cognition.2018.04.004>
- Winters, S., & O'Brien, M. G. (2012). Perceived accentedness and intelligibility: The relative contributions of F0 and duration. *Speech Communication*, 55(3), 486–507.  
<https://doi.org/10.1016/j.specom.2012.12.006>
- Young, R. F. (1999). Sociolinguistic approaches to SLA. *Annual Review of Applied Linguistics*, 19, 105–132. <https://doi.org/10.1017/S0267190599190068>
- Young, R. F. (2011). Interactional competence in language learning, teaching and testing. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning volume II* (pp. 426–443). Routledge.

# L2 speaking and its internal correlates

## A meta-analysis

Rie Koizumi, Yo In'nami and Eun Hee Jeon

Seisen University / Chuo University / University of North Carolina  
at Pembroke

The current meta-analysis examines relationships between second-language (L2) speaking (assessed using global ratings) and its internal features, such as fluency and accuracy (assessed using analytic ratings or measures) that were derived by analyzing speaking performance during the same speaking tasks. A synthesis of 39 studies (284 correlations) suggests that internal features are strongly correlated to L2 speaking in general ( $r = .649$ ) and that the strength of the correlations varies according to oral features (e.g.,  $r = .713$  to  $.888$  for fluency, delivery, grammar, vocabulary, pronunciation, and content). The results highlight the relative importance of various internal features in L2 speaking and suggest areas in need of further research.

### 1. Introduction

Second-language (L2) assessment researchers have been interested in how L2 speech is produced and what underlies this process (Kormos, 2006). It is well known that L2 speaking performance or proficiency is influenced by numerous variables (Fulcher, 2003). Sometimes, L2 speaking researchers measure those variables externally, that is, independently of L2 speaking. For the purpose of this study, we refer to them as external variables. Examples of such external variables are language aptitude, anxiety, L2 vocabulary knowledge, and L2 grammatical knowledge. These variables are present in the L2 speaker in the form of cognitive or affective traits (in the case of language aptitude and anxiety) and L2 proficiency (in the case of L2 vocabulary and grammar knowledge), all of which influence the quality of the L2 speaker's oral performance. At other times, L2 speaking researchers assess the quality of L2 speaking by measuring its quantity and quality of oral features such as fluency, accuracy, lexical or grammatical complexity of the spoken discourse itself. For example, speaking performance is likely to be superior when it has features

such as better fluency, accuracy, or higher grammatical or lexical complexity. For the purpose of this study, we shall refer to these features as internal variables. The relationship between external variables and L2 speaking is examined in Chapter 11. In the present chapter, we investigate the relationships between L2 speaking and its internal features (i.e., internal correlates) using a meta-analysis.

## 2. Literature review

Complex processes are involved in effective L2 communication. Based on Levelt (1989), Kormos (2006) developed a model of L2 speaking involving conceptualization (creation of preverbal messages), formulation (encoding of lexis, grammar, and phonology), articulation (vocalization of encoded messages), and monitoring (modifying messages and errors at each stage). These processes occur automatically and in a parallel manner among native speakers or highly proficient L2 learners, but novice or intermediate L2 learners tend to have difficulty executing automatic and parallel processing. While struggling with this difficulty, L2 learners attempt to speak fluently, accurately, comprehensibly, and coherently, to make an effective delivery. Oral features such as vocabulary, grammar, pronunciation, fluency, accuracy, comprehensibility, delivery, coherence, and content are involved in these processes, and they are considered essential internal features. Additionally, Housen, Kuiken, and Vedder (2012) have proposed that complexity, accuracy, and fluency (CAF) are the principal components of L2. Skehan (2009) has proposed considering lexis together with CAF. The models of Kormos (2006) and Housen et al. (2012) have led to numerous empirical studies and advanced L2 speaking research.

Previous studies suggest that while all internal features contribute to L2 speaking, some features contribute more than others. Adams (1980) examined how five internal features (i.e., accent, comprehension or comprehensibility, fluency, grammar, and vocabulary) affected the overall assessment of the quality of L2 speaking (assessed using global holistic ratings) in an oral interview. It was found that vocabulary, fluency, comprehension, and grammar were the most important features at some proficiency levels. Iwashita, Brown, McNamara, and O'Hagan (2008) examined relationships between overall holistic scores of L2 speaking and measures of five internal features in monologues: grammatical accuracy, grammatical complexity, vocabulary, pronunciation, and fluency. The data showed that while all the features were related to L2 speaking, vocabulary (the number of type and token) and fluency (speech rate, or the number of syllables divided by response time) were more closely related to L2 speaking. While previous studies demonstrated that some features are better predictors of the overall judgment of L2 speaking, each study included a limited range of internal features. Therefore, a meta-analysis

that synthesizes previous primary studies can deepen our understanding of these relationships. Internal features that have been examined in previous studies are reviewed below, beginning with the CAF features and then other features, such as pronunciation and content.

## 2.1 Fluency

Fluency refers to “surface smoothness in speaking, born of an ease in mobilising linguistic knowledge under the pressure” (Foster, 2020, p. 446). It can be defined broadly or narrowly, and a narrow definition has been used in L2 research (e.g., Housen et al., 2012; Tavakoli & Wright, 2020). According to Segalowitz (2010), fluency can be classified into three types: cognitive fluency or “the speaker’s ability to efficiently mobilize and integrate the underlying cognitive processes responsible for producing utterances with the characteristics that they have”; utterance fluency or “the features of an utterance”; and perceived fluency or “a judgment made about speakers based on impressions drawn from their speech samples” (p. 48). This chapter focuses on utterance and perceived fluency as internal features of L2 speaking, rather than cognitive fluency as it is typically measured externally (i.e., assessed independently from the speech sample) using cognitive tests, such as lexical retrieval measures (to assess picture naming) and sentence completion tasks (to measure sentence building speed), as used in De Jong, Steinel, Florijn, Schoonen, and Hulstijn (2013; for more detail, see Tavakoli & Wright, 2020).

Perceived fluency is typically evaluated by raters using rating scales or rubrics and usually called “fluency” in this context. It has been reported as one of the most influential features of L2 speaking ability (Adams, 1980). Sato (2012) also observed a strong correlation between the speaking ratings and fluency ratings of monologues ( $r = .74$ ).

Utterance fluency is typically divided into three dimensions (Tavakoli & Skehan, 2005): speed, breakdown, and repair fluency. Speed fluency is related to how fast a learner produces language and is typically separated into the following three features (De Jong, 2018; Tavakoli & Wright, 2020):

- Speech rate (assessed using quantitative measures such as the number of syllables or words or tokens, divided by total response time, which is the total time including silent pause time)
- Articulation rate (e.g., the number of syllables divided by total phonation time, or speaking time that excludes silent pause time)
- Mean length of run (e.g., the number of syllables divided by the number of the runs), where runs refer to utterances between silent pauses (e.g., pauses of 0.25 seconds or more)

Breakdown fluency concerns the length, number, and location of pauses and has three features:

- Phonation time ratio (e.g., total phonation time divided by total response time)
- Pause ratio (e.g., total silent pause time divided by total response time): The pause ratio is calculated as 1 minus the phonation time ratio.
- Pause length or mean duration of silent pause (e.g., total silent or unfilled pausing time divided by the number of silent pauses)

Repair fluency involves “reformulation, replacement, false starts, and repetition of words and phrases” (Tavakoli & Skehan, 2005, p. 255) and has one feature:

- Disfluency rate (e.g., the number of disfluencies [e.g., repetitions, self-corrections, and/or filled pauses] divided by total response time or by the number of units or words)

Among the three dimensions of utterance fluency, speed fluency has been reported as the most important in predicting L2 speaking. For example, Iwashita et al. (2008) showed that speed fluency, as measured by the speech rate, has a stronger relationship with the overall assessment of L2 speaking ( $\eta^2 = .60$ ) than breakdown fluency ( $\eta^2 = .20$  to  $.30$ ) or repair fluency ( $\eta^2 = .02$ ).<sup>1</sup> Further, in relation to fluency ratings, Saito, Ilkan, Magne, Tran, and Suzuki (2018) reported that repair fluency (measured by repetition and self-correction ratios) was not correlated with overall L2 fluency ratings of speech ( $\eta^2 = .01$  to  $.05$ ) in contrast to speed fluency (measured by the articulation rate:  $\eta^2 = .48$ ) and breakdown fluency (measured by the mid- and end-clause pause ratio:  $\eta^2 = .13$  to  $.40$ ). Based on the results, Saito et al. (2018) argued that speed fluency and breakdown fluency were generally strong predictors of L2 fluency and that repair fluency was much less related to L2 fluency judgments (see also Suzuki, Kormos, & Uchihara, 2021).

## 2.2 Accuracy, grammar, and vocabulary

Accuracy is defined as “the extent to which an L2 learner’s performance (and the L2 system that underlies this performance) deviates from a norm” (Housen et al., 2012, p. 4), and it usually refers to the degree to which grammatical, lexical, and/or phonological forms are used correctly. There are two dimensions: global and specific accuracy (Iwashita et al., 2008). Global accuracy assesses all types of errors, whereas specific accuracy focuses on certain types of error, such as verb tense and

---

1. Although the effect size, eta, was reported by the authors, our recalculation suggests that it was an eta-squared ( $\eta^2$ ). We report the eta-squared value here.



plural markers. While specific accuracy is used in studies investigating how accuracy develops, its relationship with L2 speaking is not often tested. Global accuracy is separated into the following three features:

- Error-free clause ratio (e.g., the number of error-free clauses divided by the number of clauses): Error-free clauses refer to those without any errors, although some types of errors may be exempted.
- Error-free unit ratio (e.g., the number of error-free units divided by the number of units): A unit could be a T-unit, c-unit, or Analysis of Speech Unit (AS-unit), the last of which is defined as “a single speaker’s utterance consisting of an independent clause, or sub-clausal unit, together with any subordinate clause(s) associated with either” (Foster, Tonkyn, & Wigglesworth, 2000, p. 365).
- Error ratio (e.g., the number of errors divided by the number of words [tokens], clauses, or units).

Previous studies have reported that global accuracy is related to overall ratings of L2 speaking to a limited degree (e.g.,  $\eta^2 = .22$  in Iwashita et al., 2008). Gan (2008) also reported no correlation between speaking scores and grammatical accuracy in group discussions ( $r = -.003$ ).

Grammar typically includes grammatical (morphological and syntactic) accuracy and other aspects, such as the range of grammar used (i.e., grammatical complexity; e.g., Clark & Swinton, 1980). Vocabulary covers lexical accuracy and appropriateness and the range of words and phrases used (i.e., lexical complexity; e.g., Clark & Swinton, 1980). Grammar and vocabulary have been considered to be the central aspects that learners should acquire and use in speaking an L2. Sato (2012) reported strong correlations between overall L2 speaking scores and grammar and vocabulary ratings of monologues ( $r = .59$  and  $.69$ ). As seen in Sato’s (2012) study, grammar and vocabulary, when conceptualized comprehensively, are typically measured using rating scales (see Koizumi, 2013, for more details about vocabulary and speaking).

## 2.3 Grammatical complexity and lexical complexity

According to Bulté and Housen (2012), complexity is a concept that has been defined and operationalized in various manners across studies. However, in general, linguistic complexity can be defined as the degree to which a speaker’s performance includes the use of “a wide and varied range of sophisticated structures and vocabulary” (Housen et al., 2012, p. 2). Although linguistic complexity includes lexical, morphological, syntactic, and phonological complexity, two types of complexity have been typically studied: grammatical and lexical complexity (Bulté & Housen, 2012).

Grammatical complexity includes syntactic and morphological complexity (Bulté & Housen, 2012, p. 27). However, two syntactic features are studied most frequently:

- Unit length or mean length of unit (e.g., the number of words [tokens] divided by the number of units or utterances)
- Sentential subordination (e.g., the number of clauses or subordinate clauses divided by the number of units).

Lexical complexity is also called lexical richness and typically comprises three features:

- Lexical diversity or variation, which refers to the extent to which diverse words are used (e.g., the number of word types divided by the number of words or tokens [Type token ratio or TTR], and variants of TTR, such as the Guiraud's index [the number of word types divided by the square root of the number of tokens] and D [also called voc-D or the hypergeometric distribution diversity index: HDD; see Zenker & Kyle, 2021, for how to compute D])
- Lexical density (e.g., the number of lexical words or content words, divided by the number of words, both lexical and function words)
- Lexical sophistication (e.g., the number of sophisticated, or less frequently used, word types divided by the number of types)

Grammatical complexity and lexical complexity have shown significant correlations with L2 speaking, ranging from marginal to moderate in size. In an analysis of presentation and discussion performances, Gan (2012) reported marginal correlations between global ratings of L2 speaking and measures of grammatical complexity, in terms of both unit length ( $r = .086$  to  $.154$ ) and sentential subordination ( $r = -.062$  to  $.118$ ). Lu (2012) conducted a meta-analysis that synthesized 12 spoken corpora of Chinese university students' speeches. He found marginal correlations between overall L2 speaking scores and measures of lexical density ( $r = .011$ ) and lexical sophistication ( $r = .048$  to  $.166$ ) and a weak to moderate correlation between L2 speaking scores and lexical diversity measures (e.g.,  $r = .352$  to  $.430$ ).

## 2.4 Pronunciation, comprehensibility, delivery, content, and coherence

Pronunciation, comprehensibility, delivery, content, and coherence are also internal features affecting L2 speaking. Pronunciation includes "many features of the speech stream, such as individual sounds, pitch, volume, speed, pausing, stress and intonation" (Luoma, 2004, p. 11). Comprehensibility is typically defined as "listeners' perception of how easy it is for them to understand L2 speech," which has been known to be affected by lexical accuracy, lexical complexity, and fluency (Saito, Webb,

Trofimovich, & Isaacs, 2016, p. 597). Delivery is a broad concept and defined rather differently across studies and tasks. For example, Xi and Mollaun (2006) defined it as “the pace and clarity of the speech,” including “the speakers’ pronunciation, intonation, rate of speech, and degree of hesitancy” (p. 10) in assessing computer-based monologue tasks, whereas Pietilä (1999) included “clear and audible voice, varied and well-modulated voice, and non-verbal-behavior (e.g., gestures, mannerisms)” (p. 80) in assessing oral presentation. Content typically refers to “content elaboration/development, indicating the degree to which the test-taker is conveying relevant and well-elaborated/developed ideas on given topics” (Sato, 2012, p. 226). Coherence refers to the “organization of ideas,” which includes dimensions of logic, topic sentences, transitions, and relevance (Cao, 2014, p. x). All of these concepts are typically measured using rating scales. A limited number of studies have shown that these features are substantial predictors of global ratings of L2 speaking (e.g.,  $r = .55$  with pronunciation and  $r = .77$  with content, both in Sato, 2012).

## 2.5 Measuring L2 speaking and its internal features

L2 speaking has been measured using rating scales, while L2 internal oral features have been assessed using either rating scales or measures. Rating scales are typically used by human raters who judge the quality of the speech. Measures are computed using a formula (e.g., the number of error-free clauses divided by the number of clauses); after a speech is transcribed, the text is segmented into certain units (e.g., clauses or AS-units), and the relevant parts are counted (e.g., the number of error-free clauses). Rating scales can be holistic or analytic and are often called subjective measures, whereas measures are called objective (quantitative) measures or discourse analytic measures. Both involve human judgment, although humans are less involved in the latter. Researchers have used ratio (or average) measures, in which the count of the relevant features is divided by a unit of measurement since measures using simple counts are not usually comparable across study participants and across studies (Tavakoli & Wright, 2020) due to different text (the number of tokens) and speech lengths (response time length).

As multiple measures can be computed to assess a single internal feature, the question of which measure better reflects its intended construct is important (e.g., Bulté & Housen, 2012). For example, researchers interested in lexical diversity have long examined this question because of serious validity concerns about the measures (Koizumi & In’nami, 2012; Kyle, Crossley, & Jarvis, 2021). The TTR, a measure that used to be employed most and that is easily computed, is known to be heavily affected by text length. Since eliminating the impact of text length was not simple, studies have devised and compared numerous measures, such as the Guiraud’s

index, D, and the measure of textual lexical diversity (MTLD; see Zenker, & Kyle, 2021, for details). Meta-analyses conducted for each measure in each feature would provide insights into its characteristics.

## 2.6 Relative strengths of relationships between L2 speaking and internal features

As reviewed above, previous studies suggest that some relationships between L2 speaking and its internal features seem to be stronger than others (e.g., fluency is more closely related to L2 speaking than accuracy). To the authors' knowledge, studies analyzing speaking ability or performance in terms of CAF or language knowledge (e.g., Bachman & Palmer, 2010; Housen et al., 2012) have not examined the relative importance of CAF or its components. Further, in Hulstijn's (2015) core-peripheral model, L2 proficiency is affected by core and peripheral components; core components are hypothesized to have stronger correlations with L2 speaking than peripheral components do. Using this model, it can be hypothesized that vocabulary, grammar, pronunciation, and fluency as core components (which reflect linguistic knowledge and processing speed) have stronger correlations with L2 speaking proficiency than do peripheral components such as interactional ability, strategic competence, and metalinguistic knowledge (see Hulstijn, 2015). The current meta-analysis focuses on core components and examines whether vocabulary, grammar, pronunciation, and fluency are strongly related to L2 speaking.

## 3. Current study and research questions

The purpose of our study was to examine the relationship between L2 speaking and its internal features overall and to examine how relationships vary across internal features, as the correlations are likely to change according to the specific internal features and measures used. Table 1 shows the internal features examined in our study in order to address these research questions:

- Research question 1: What is the relationship between L2 speaking and its internal features overall?
- Research question 2: How does this relationship vary across individual internal features?

Previous studies have focused on particular features in their own research context. The current chapter focuses on a wide range of oral features and presents a meta-analysis of the degree to which each feature is related to L2 speaking in order to develop a broad picture. To the authors' knowledge, this is the first meta-analysis conducted on this topic that includes various internal features (see Lu, 2012, for a

**Table 1.** Summary of internal features

Feature	Specific ratio measure (Formula)
Fluency (RS)	
Speed fluency (RM)	<p>Speech rate:</p> <ol style="list-style-type: none"> <li>1. Syllable/response time (the number of syllables divided by response time length)</li> <li>2. Word/response time (the number of words divided by response time length)</li> </ol> <p>Articulation rate:</p> <p>Syllable/phonation time (the number of syllables divided by phonation time length [i.e., response time minus time with filled pauses])</p> <p>Mean length of run:</p> <p>The number of words divided by the number of runs (i.e., utterances between silent pauses [e.g., pauses of 0.25 seconds or more])</p>
Breakdown fluency (RM)	<p>Phonation time ratio: Phonation time/response time (phonation time length divided by response time length)</p> <p>Pause ratio<sup>a</sup>: Silent pause time/response time (silent pause time length divided by response time length)</p> <p>Pause length<sup>a</sup>: Silent pause time/silent pauses (silent pause time length divided by the number of silent pauses)</p>
Repair fluency (RM)	Disfluency rate <sup>a</sup> (e.g., the number of disfluencies [e.g., repetitions, self-corrections] divided by response time length)
Accuracy (RM)	<p>Error-free clause ratio: Error-free clause/clause (the number of error-free clauses divided by the number of clauses)</p> <p>Error-free unit ratio: Error-free unit/unit (the number of error-free units divided by the number of units)</p> <p>Error ratio<sup>a</sup>: Error/word (the number of errors divided by the number of words)</p>
Grammar (RS)	
Vocabulary (RS)	
Grammatical complexity (RM)	<p>Unit length (the number of words divided by the number of units)</p> <p>Sentential subordination (the number of clauses divided by the number of units)</p>
Lexical complexity (RM)	1. Guiraud's index: the number of word types divided by the root square of the number of tokens
Lexical diversity	2. D
Lexical density	E.g., the number of lexical words divided by the number of words
Lexical sophistication	E.g., the number of sophisticated word types divided by the number of types
Pronunciation, Comprehensibility, Delivery, Content, Coherence (all RSs)	

Note. RS = Rating scale. RM = Ratio measure.

a. Lower values mean more fluent or accurate speeches.

meta-analysis of lexical complexity). Our meta-analysis will help researchers and practitioners better understand what underlies L2 speaking and what L2 speaking test scores represent.

4. Method

4.1 Literature search

We employed the same methods as those used in Chapter 8 (listening). We included published and unpublished studies that were completed by October 2018 that we obtained via literature searches conducted between January 2015 and October 2018. Figure 1 shows our search trajectory. Points that differ from the meta-analysis of listening and its components are explained below.

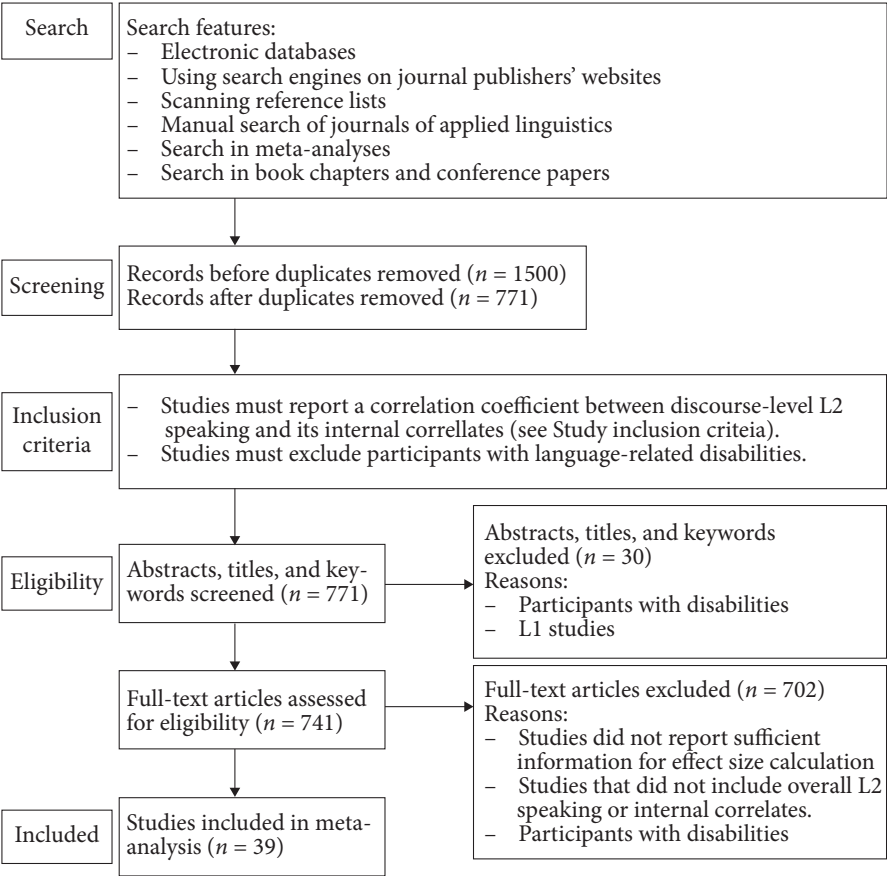


Figure 1. PRISMA chart: Flow diagram for the literature search and study inclusion criteria

We used the following keywords to locate studies with databases: “L2 AND speaking AND XX,” where XX was replaced in turn with *correlate\**, *component\**, *sub-component\**, *construct\**, *subconstruct\**, *correlat\**, *oral communica\**, *oral skill\**, *speak\**, *spee\**, *spoken*, *phon\**, *pronuncia\**, *vocab\**, *lexic\**, *word*, *working memory*, *gramma\**, *syntac\**, *discourse*, and *speech produc.\** We searched studies of both L2 assessment and acquisition.

## 4.2 Study inclusion criteria

To be included, a study had to do the following: (1) examine the relationships between L2 speaking and its various internal features using correlations (e.g., Pearson or Spearman); (2) use the global judgment of discourse-level L2 speaking, which was derived typically using a holistic rating scale (e.g., American Council on the Teaching of Foreign Languages Oral Proficiency Interview [ACTFL OPI] in Freed, 1995) or analytic rating scale with multiple criterion scores combined (e.g., totaling four dimension scores of presentation: Pronunciation & Delivery, Communication Strategies, Vocabulary & Language Patterns, and Ideas & Organization, in Gan, 2012); (3) report on one or more correlations between L2 speaking and internal features obtained from the same speech production; and (4) target L2 learners without language-related disabilities. Regarding (1), we excluded studies that reported only factor correlations (e.g., Farnsworth, 2013), partial correlations (e.g., Saito et al., 2018), regression coefficients (e.g., Révész, Ekiert, & Torgersen, 2016), and/or means and standard deviations (Iwashita et al., 2008). The Spearman correlations included consisted of 21 correlations from Lu (2012), eight correlations from Ushigusa (2008), six correlations from Koizumi and Yamanouchi (2003), five correlations from Koizumi and Kurizaki (2002), four correlations from Segalowitz and Freed (2004), two correlations from Milton, Wade, and Hopkins (2010), and one correlation each from Koizumi (2005b) and Malvern and Richards (2002). Regarding (2), we included studies using only global ratings that assessed various aspects of speaking rather than those including narrower ratings (e.g., Kormos & Dénes, 2004, focusing on fluency; Saito et al., 2016, focusing on comprehensibility and accentedness). We also excluded studies that only involved reading aloud, simple repetition, or imitation (e.g., Cucchiari, Strik, & Boves, 2000). Concerning (3), we excluded studies that correlated L2 speaking scores and oral features derived from different tasks (e.g., TOEFL iBT speaking scores and fluency scores for group oral discussions in Ockey, Koyama, Setoguchi, & Sun, 2015), because the inclusion of studies using different tasks could blur the relationships. Regarding (4), we included studies with bilingual participants if we could identify their L2 in the reports. A total of 741 studies were examined by the authors using these criteria, resulting in 39 studies being included in the meta-analysis (see Appendix A).



### 4.3 Coding

These 39 studies (284 effect sizes) were coded in terms of correlations and the internal features, which were categorized according to features and measures (see Table 1 and Appendix B for details).

Only ratio measures (e.g., the number of clauses per unit rather than just the number of clauses) were included. When a study focused on grammatical accuracy and used a ratio measure, it was coded as accuracy (e.g., Jin & Mak, 2013). We did not include the TTR because of its low validity.<sup>2</sup> We examined the rating scale or formula for calculating ratio measures and categorized them accordingly when studies did not label the features precisely. We assumed that measures would be the same if they were mathematically equivalent when they were standardized. For example, “words per second” and “words per minute” were coded as the same measure because the latter is calculated using a formula of the number of words divided by the time taken by a candidate to speak in seconds (i.e., words per second), multiplied by 60. Similarly, “errors per token” and “errors per 100 tokens” were coded as the same measure because the latter is calculated using a formula of the number of errors divided by the number of tokens (i.e., errors per token), multiplied by 100. When higher values meant less quality (e.g., higher values of pause ratio measures mean less fluency), the signs of correlations were changed (e.g., from negative to positive). This applied to ratio measures of breakdown fluency (i.e., pause ratio and pause length), repair fluency, and accuracy (i.e., error ratio). We reclassified and analyzed features or measures that had been used in primary studies three or more times. Features or measures that appeared less than that were not included (e.g., specific accuracy).

Regarding intercoder reliability, the first and second authors coded seven of the 39 studies and obtained agreement percentages of 90% to 100% across all coded variables. The second and third authors coded two other studies and obtained the agreement percentages of 95% to 100%. We resolved disagreement through discussion. Either the first or second author coded the remaining studies.

### 4.5 Analyses

Effect sizes were aggregated using a random-effect model with a robust variance estimator in the *robumeta* package in R (Fisher & Tipton, 2015) to consider the within-study dependency of multiple effect sizes. Thus, two or more effect sizes

---

2. The TTR (i.e., the number of types divided by the number of tokens) was not included in the meta-analysis because it is seriously affected by the number of tokens. However, for reference, the synthesized results of the TTR are presented here:  $N$  of studies = 4;  $N$  of independent participants = 637,  $N$  of dependent participants = 652,  $N$  of  $r$ s = 5;  $r = -.380$ ; 95% confidence interval:  $-.801, .293$ ; minimum  $r = -.690$ ; maximum  $r = .331$ ; fail-safe  $N = 168$ ;  $I^2 = 96.233$ . The results support previous studies that discourage its use (e.g., Koizumi & In'nami, 2012).

from a single study were coded and included as they were. We conducted Fisher's  $z$ -transformation of correlation coefficients, summed the transformed values to compute an average weighted effect size, and back-transformed the value for interpretation. We repeated this process for all meta-analyses.

The minimum number of correlations, not studies, to be included in the meta-analysis (Li, 2016) was set at three. Thus, for example, our analysis included three correlations from three studies each (in the case of the phonation time ratio) and 11 correlations from two studies (in the case of the error ratio).

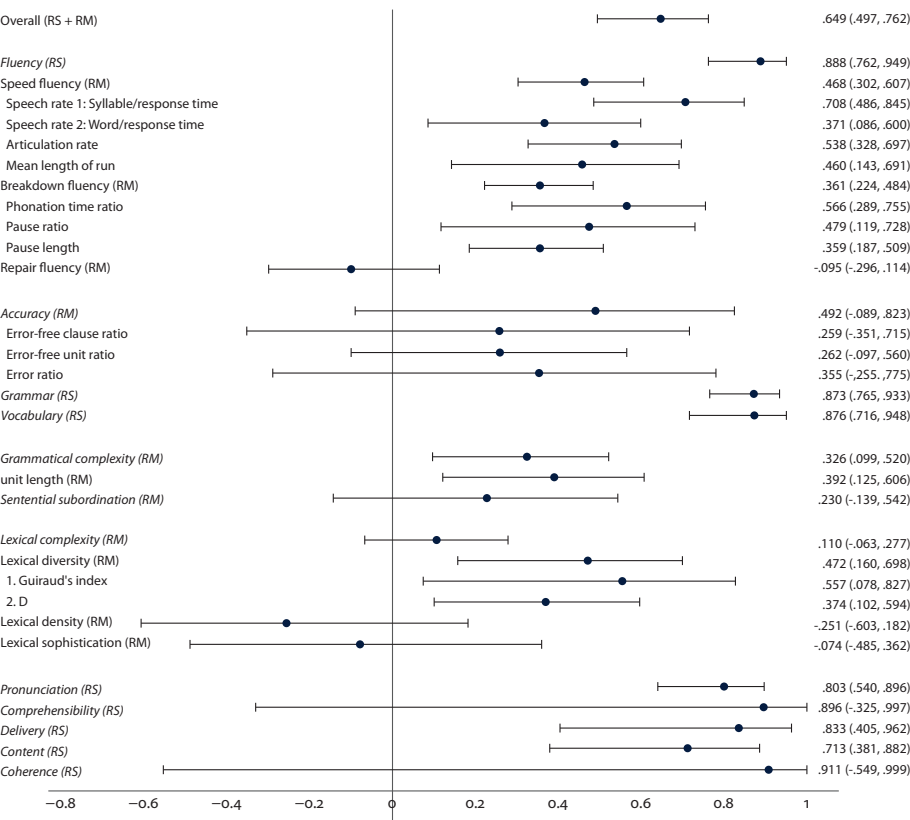
When there was no overlap between the 95% confidence intervals, we interpreted this as a statistically significant difference between the correlations that were being compared. We used  $I^2$  statistics to determine the percentages where "the variance in observed effects reflects variation in true effects, rather than sampling error" (Borenstein, 2019, p. 115), with larger  $I^2$  values indicating that more variation in true effects was reflected. We used Orwin's (1983) fail-safe  $N$  (calculated in Microsoft Excel) to examine publication bias and treated this value as the number of studies required to diminish the aggregated effect size from the criterion effect size ( $r = .01$ ), with a larger number indicating a lower likelihood of publication bias. In particular, we judged fail-safe  $N$ s that were  $5k + 10$  or more ( $k$  = the number of aggregated studies) to indicate an inconsequential level of publication bias (Rosenthal, 1979). Following Plonsky and Oswald (2014), we interpreted effect sizes of correlations of .25, .40, and .60 as weak, moderate, and strong, respectively.

## 5. Results

Table 2 shows that the aggregated correlation between the global assessment of L2 speaking and measures/scores of its internal features was statistically significant and large ( $r = .649$  [95% confidence interval = .497, .762],  $p < .001$ ). The  $I^2$  value indicates that 98% of the variance in observed effects consists of variation in the true overall correlation between L2 speaking and its correlates. Publication bias was not found with the fail-safe  $N$  of 3,288, which exceeded the threshold of 205 (calculated using the formula  $5 \times 39 + 10$ ). This means that we needed 3,288 studies to negate the summary effect ( $r = .649$ ) and suggests a limited impact of publication bias on our results. Results of more detailed analyses suggest trends similar to the overall meta-analysis results. For example, there were no indications of publication bias throughout the analyses. However, there were eight internal features for which the  $I^2$  values were small: D ( $I^2 = 0.000$ ), phonation time ratio ( $I^2 = 36.284$ ), content ( $I^2 = 44.181$ ), sentential subordination ( $I^2 = 43.556$ ), articulation rate ( $I^2 = 52.138$ ), pause length ( $I^2 = 57.273$ ), repair fluency ( $I^2 = 57.596$ ), and error ratio ( $I^2 = 59.769$ ). Further, some correlations were not statistically significant and/or had wide confidence intervals (e.g., comprehensibility: .896 [−.325, .997];

coherence: .911 [−.549, .999]). The low  $I^2$  values, nonsignificant correlations, and wide confidence intervals may have been the result of a small number of studies (e.g.,  $k = 3$  or 4; see e.g., von Hippel, 2015, for  $I^2$  and sample size effects) and required careful interpretation. These results are not emphasized in the interpretation below.

As shown in Table 2 and Figure 2, most of the features assessed using rating scales were significantly and strongly related with the overall assessment of L2 speaking, such as fluency (.888 [.762, .949]). Regarding features assessed using ratio measures, some had statistically significant correlations with L2 speaking that were strong (i.e., speech rate 1 [syllable/response time]: .708 [.486, .845]), moderate (e.g., speed fluency: .468 [.302, .607]), or weak (e.g., grammatical complexity: .326 [.099, .520]). These results suggest varied relationships between L2 speaking and each of the internal features.



**Figure 2.** Forest plot of L2 speaking and its internal features  
*Note.* RS = Rating scale. RM = Ratio measure.

Table 2. L2 speaking and its internal features

Feature	N of studies	N of independent participants	N of dependent participants	N of rs	R	95% CI		Min <i>r</i>	Max <i>r</i>	Fail-safe <i>N</i>	<i>I</i> <sup>2</sup>
Overall (RS + RM)	39	4602	31576	284	.649**	.497	.762	-.450	.990	3288	98.167
Fluency (RS)	11	2037	2043	12	.888**	.762	.949	.310	.990	2113	97.130
Speed fluency (RM)	15	1662	3805	35	.468**	.302	.607	-.418	.830	779	96.581
Speech rate 1: Syllable/response time	5	478	478	5	.708**	.486	.845	.520	.830	496	95.074
Speech rate 2: Word/response time	8	730	1066	14	.371*	.086	.600	-.418	.804	312	80.649
Articulation rate	4	455	455	4	.538**	.328	.697	.430	.633	251	52.138
Mean length of run	7	929	957	8	.460*	.143	.691	.106	.781	356	93.841
Breakdown fluency (RM)	8	936	3926	31	.361**	.224	.484	-.230	.707	302	79.489
Phonation time ratio	3	389	389	3	.566*	.289	.755	.510	.707	203	36.284
Pause ratio <sup>a</sup>	4	784	784	4	.479*	.119	.728	.294	.707	214	87.044
Pause length <sup>a</sup>	5	815	1210	6	.359**	.187	.509	.204	.547	187	57.273
Repair fluency (RM) <sup>a</sup>	6	287	1118	24	-.095	-.296	.114	-.430	.371	63	57.596
Accuracy (RM)	8	611	1906	36	.492	-.089	.823	-.105	.980	444	98.231
Error-free clause ratio	3	135	304	6	.259	-.351	.715	-.020	.540	77	76.242
Error-free unit ratio	4	231	442	9	.262	-.097	.560	.020	.570	105	77.699
Error ratio <sup>a</sup>	2	89	462	11	.355	-.285	.776	-.105	.560	74	59.769
Grammar (RS)	9	1673	1679	10	.873**	.765	.933	.370	.960	1602	92.742
Vocabulary (RS)	7	585	591	8	.876**	.716	.948	.340	.980	1264	91.920
Grammatical complexity (RM)	11	440	1798	46	.326*	.099	.520	-.397	.850	368	96.466
Unit length (RM)	9	358	707	19	.392*	.125	.606	-.013	.850	374	96.309
Sentential subordination (RM)	4	158	357	8	.230	-.139	.542	-.062	.517	91	43.556
Lexical complexity (RM)	8	1238	11328	49	.110	-.063	.277	-.450	.780	81	94.417

(continued)

Table 2. (continued)

Feature	N of studies	N of independent participants	N of dependent participants	N of rs	R	95% CI		Min <i>r</i>	Max <i>r</i>	Fail-safe N	I <sup>2</sup>
Lexical diversity (RM)	6	831	6346	24	.472*	.160	.698	-.091	.780	315	94.560
1. Guiraud's index	4	641	1218	8	.557*	.078	.827	.174	.780	264	92.490
2. D	4	619	640	5	.374*	.102	.594	.245	.440	157	0.000
Lexical density (RM)	4	632	1002	9	-.251	-.603	.182	-.450	.011	108	89.594
Lexical sophistication (RM)	3	620	2538	10	-.074	-.485	.362	-.370	.166	25	88.778
Pronunciation (RS)	13	1920	2162	20	.803**	.640	.896	.356	.990	1738	95.567
Comprehensibility (RS)	3	372	400	4	.896	-.325	.997	.630	.990	602	97.403
Delivery (RS)	3	163	303	4	.833*	.405	.962	.720	.923	449	68.619
Content (RS)	4	229	229	4	.713*	.381	.882	.356	.830	403	44.181
Coherence (RS)	3	397	397	3	.911	-.549	.999	.422	.980	660	96.670

Note RS = Rating scale. RM = Ratio measure.

a. The sign of the *r* value was reversed from negative to positive for consistency with other measures; thus a positive correlation indicates that learners with less of this feature are more likely to perform well in speaking tests (see Coding).

\*  $p < .05$ .

\*\*  $p < .01$ .

## 6. Discussion

To investigate the relative contributions of various internal features to L2 speaking proficiency and to address our two research questions, we systematically collected and meta-analyzed previous studies. Concerning Research Question 1, the synthesized mean correlation between L2 speaking (assessed using global ratings) and its internal features (assessed using analytic ratings or ratio measures) overall indicated a strong relationship ( $r = .649$ ).

A more detailed picture emerged when the internal features were analyzed separately (Research Question 2). The statistically significant correlations were all positive, but the strength of the correlations varied across features, ranging from strong (e.g., fluency:  $r = .888$ ) to weak (grammatical complexity:  $r = .326$ ). A clear pattern was observed between features assessed using rating scales and features assessed with ratio measures. Correlations with rating scales tended to be stronger than correlations with ratio measures. For example, Fluency, which was assessed using rating scales, was more strongly related to the overall assessment of L2 speaking ( $r = .888$ ) than speed, breakdown, and repair fluency ( $r = -.095$  to  $.708$ ), all of which were assessed using ratio measures. These findings may result from the tendency of rating scales to elicit wider aspects of the target constructs. For instance, (perceived) fluency includes dimensions of speed, breakdown, and repair fluency (e.g., Sato, 2012). These relative relationships can be predicted by imperfect correlations of fluency, pronunciation, vocabulary, grammar, and discourse (measured using rating scales) with the corresponding features (measured using ratio scales; Saito, Trofimovich, & Isaacs, 2017; Saito et al., 2018; Suzuki et al., 2021). The discussions below are grouped by internal features and focus more on other explanations.

### 6.1 Fluency

Fluency, as measured by rating scales, was significantly and strongly correlated with L2 speaking ( $r = .888$ ). The results are not surprising for two reasons. First, fluent oral production and underlying linguistic processing speed and smoothness have been considered highly important in speaking (e.g., De Jong, 2018; Segalowitz, 2010; Tavakoli & Wright, 2020), because articulating messages in an unnaturally disfluent manner may result in communication breakdown, even if the messages were conceptualized and formulated well.

Second, raters might have conceptualized fluency as equivalent to or synonymous with L2 overall proficiency or speaking proficiency. This view conceives of fluency broadly (Tavakoli & Wright, 2020). If untrained raters or even well-trained raters subconsciously adopted this broad view, this might have led to a stronger correlation in the meta-analysis.

When (utterance) fluency was divided into speed, breakdown, and repair fluency, the results suggested that the correlation between L2 speaking and speed fluency was the strongest ( $r = .468$  [.302, .607]), followed by breakdown fluency ( $r = .361$  [.224, .484]), and then repair fluency ( $r = -.095$  [-.296, .114]); the confidence intervals for speed and breakdown fluency overlapped. The findings of the relative importance of speed, breakdown, and then repair fluency in relation to L2 speaking concur with the existing literature (e.g., Iwashita et al., 2008; Saito et al., 2018). The absence of a significant correlation between repair fluency and L2 speaking may be due to the hesitation markers (e.g., repetitions) considered in repair fluency sometimes helping effective communication rather than hindering comprehension (De Jong, 2018).

Among speed fluency measures, although the confidence intervals all overlapped with each other, syllable/response time (speech rate 1) had the strongest correlation with L2 speaking ( $r = .708$  [.486, .845]), followed by syllable/phonation time (articulation rate;  $r = .538$  [.328, .697]), mean length of run ( $r = .460$  [.143, .691]), and word/response time (speech rate 2;  $r = .371$  [.086, .600]). Syllable/response time is usually preferred by researchers as a measure of speech rate over word/response time, as the former differentiates production volumes precisely (Ellis & Yuan, 2004). The superiority of syllable/response time seemed to be reflected in its stronger mean correlation with L2 speaking. Regarding two similar measures, De Jong (2018) argued that, in contrast to syllable/phonation time as a pure measure of speed fluency, syllable/response time assessed both speed fluency (specifically the articulation rate) and breakdown fluency. In support of this argument, the correlation between syllable/response time and L2 speaking was stronger than the syllable/phonation time correlation, probably because syllable/response time assesses fluency more broadly (Cucchiari et al., 2000) and has a better discriminatory power across different proficiency levels (Tavakoli, Nakatsuhara, & Hunter, 2020; Yan, Kim, & Kim, 2018).

Among breakdown fluency measures, phonation time/response time (phonation time ratio) had the strongest correlation with L2 speaking ( $r = .566$  [.289, .755]), followed by silent pause time/response time (pause ratio;  $r = .479$  [.119, .728]) and then silent pause time/silent pause (pause length;  $r = .359$  [.187, .509]). Similar correlations across the first and second measures are not surprising because the phonation time ratio is calculated using a formula of 1 minus the pause ratio (Tavakoli & Wright, 2020). L2 speaking may have a slightly stronger correlation of L2 speaking with the phonation time ratio than with the pause length because the phonation time ratio better distinguishes groups of different proficiency levels (Tavakoli et al., 2020).



### *Accuracy, grammar, and vocabulary*

Accuracy was moderately but not significantly correlated with L2 speaking ( $r = .492$  [ $-.089, .823$ ]). In contrast, grammar and vocabulary, both accuracy-related features, were significantly and strongly associated with L2 speaking ( $r = .873$  [ $.765, .933$ ] and  $.876$  [ $.716, .948$ ], respectively). This is consistent with previous studies that show a limited relationship between L2 speaking and accuracy (e.g., Gan, 2008; Iwashita et al., 2008) and strong relationships between L2 speaking and grammar and vocabulary (e.g., Sato, 2012).

There may be two reasons why grammar and vocabulary had stronger correlations with L2 speaking than accuracy did. First, accuracy is more narrowly defined than grammar and vocabulary. Accuracy typically refers to the degree to which grammatical, lexical, and/or phonological forms are used correctly, whereas grammar and vocabulary involve more holistic judgments of learners' usage, which include broad aspects, such as grammatical accuracy and range, and lexical accuracy and range. The difference in correlations of these different constructs with L2 speaking may indicate that the key contributors to speaking success may be broadly defined grammar and vocabulary rather than narrowly defined accuracy.

Second, the other reason for differences between accuracy and grammar and vocabulary is that the rating scales may allow the raters to vary weighting according to aspects whose importance changes depending on proficiency levels. The key issue in this case is error gravity. In ratio measures of accuracy, all grammatical errors are typically taken equally seriously. Reviewing various measures employed in previous studies, Foster and Wigglesworth (2016) argued that not all errors were equal and that the gravity (i.e., seriousness) of the error should be considered when calculating accuracy. They suggested classifying errors into three types according to the degree to which they impeded meaning, with the least minor errors being least penalized. Otherwise, it would be difficult to differentiate between those who spoke comprehensibly but made many minor errors and those who made a similar number of serious errors that hindered their speech. Nevertheless, the weighting of errors does not seem to have been widely performed in the field, and this seems to be the case as well for previous L2 speaking studies. Thus, the equal weighting of minor and major errors in accuracy ratio measures could have masked the relationship between accuracy and L2 speaking. As discussed above, grammar and vocabulary were assessed holistically, with rating scales often assigning less weight to minor errors. To provide these scores, the raters typically follow rating scales that often consider the gravity of error, with minor errors receiving less weight and major errors receiving more weight. Thus, the gravity of error may need to be considered to clarify how accuracy is related to L2 speaking.

The correlations between L2 speaking and the three accuracy measures were similar, and all the confidence intervals overlapped: errors/word (error ratio;  $r = .355 [-.285, .776]$ ), error-free unit/unit (error-free unit ratio;  $r = .262 [-.097, .560]$ ), and error-free clause/clause (error-free clause ratio;  $r = .259 [-.351, .715]$ ). Foster and Wigglesworth (2016) stated that error-free clause/clause was “the best tool in the measurement kit ... because it combines a reliably defined and valid unit with a finer-grained analysis” and that it outperformed error-free unit/unit (p. 104). Interestingly, relationships with L2 speaking did not differ greatly across the measures.

## 6.2 Grammatical complexity and lexical complexity

Grammatical complexity was found to have a weak but statistically significant correlation with L2 speaking ( $r = .326 [.099, .520]$ ), whereas lexical complexity was not significantly related to L2 speaking, with a negligible correlation ( $r = .110 [-.063, .277]$ ). Considering the correlations were small at best, the results suggest that the two types of complexity, in general, are not strongly related to L2 speaking. These results accord well with previous studies (e.g., Gan, 2012; Lu, 2012). Further, it was found that the relationship between L2 speaking and grammatical and lexical complexity (assessed using ratio measures;  $r = .326$  and  $.110$ , respectively) was weaker than the relationship between L2 speaking and grammar and vocabulary (assessed using rating scales;  $r = .873$  and  $.876$ ). This parallels the finding that accuracy ( $r = .492$ ) had a weaker relationship with L2 speaking than grammar and vocabulary did ( $r = .873$  and  $.876$ ), and also that speed, breakdown, and repair fluency ( $r = -.095$  to  $.708$ ) had a weaker relationship with L2 speaking than fluency did ( $r = .888$ ). This suggests that the ratio measures of grammatical and lexical complexity, accuracy, and three features of fluency capture the concepts more narrowly than the rating scales of corresponding aspects (i.e., grammar, vocabulary, and vocabulary).

Among the two types of grammatical complexity, unit length was significantly and weakly related to L2 speaking ( $r = .392 [.125, .606]$ ), but sentential subordination was not significantly related to L2 speaking, as shown by its negligible correlation ( $r = .230 [-.139, .542]$ ). Because the two confidence intervals overlapped, the aggregated mean correlations of unit length and sentential subordination were considered to be similar in size and not strongly related to L2 speaking scores. This result may be due to the lack of clear linearity of the two dimensions. That is, they do not necessarily increase in linear increments throughout L2 proficiency development, especially in the case of sentential subordination (Bulté & Housen, 2012; Norris & Ortega, 2009).

Among three types of lexical complexity, lexical diversity was significantly and moderately related to L2 speaking ( $r = .472$  [.160, .698]). Lexical density and lexical sophistication, on the other hand, were not significantly related, with marginal correlations ( $r = -.251$  [-.603, .182] and  $-.074$  [-.485, .362], respectively). Although the three confidence intervals overlapped, the aggregated mean correlation of lexical diversity seems to be stronger than the correlations of lexical density and sophistication. This finding resonates with previous studies (e.g., Lu, 2012; Yan et al., 2018) and may be explained by the same reason for the small correlations with unit length and sentential subordination – that is, the nonlinear relationship between lexical density and sophistication and L2 speaking, in contrast to lexical diversity, which tends to increase with the development of L2 proficiency. This difference in linearity in the relationships with L2 proficiency between lexical diversity and the other two lexical dimensions was also observed in Lu (2012), which compared four groups of learners of different proficiency levels and found lexical density and sophistication were similar across the levels.

Among lexical diversity measures, the Guiraud's index ( $r = .557$  [.078, .827]) was slightly more strongly correlated than D ( $r = .374$  [.102, .594]), although the confidence intervals overlapped. Previous studies of both L2 speaking and writing (Koizumi & In'nami, 2012; Zenker & Kyle, 2021) have shown that the Guiraud's index is affected more by text length than D is, and that the Guiraud's index can be explained by not only lexical diversity but also text length to some degree (although to a lower degree than the TTR). Because text length (the number of tokens) is usually correlated with L2 speaking (e.g.,  $r = .83$  in Jin & Mak, 2013;  $r = .578$  in Gan, 2012), this characteristic may have led the Guiraud's index to have a stronger correlation with L2 speaking.

### 6.3 Pronunciation, comprehensibility, delivery, content, and coherence

Pronunciation, delivery, and content were significantly and strongly related to L2 speaking ( $r = .803$  [.640, .896],  $.833$  [.405, .962], and  $.713$  [.381, .882], respectively). This accords well with previous studies that emphasize the importance of these features in L2 speaking (e.g., Iwashita et al., 2008; Sato, 2012). Comprehensibility and coherence had strong but nonsignificant correlations ( $r = .896$  [-.325, .997] and  $.911$  [-.549, .999], respectively). Because the number of correlations included was small (four and three, respectively) but the minimum correlations derived from primary studies were positive and moderate ( $r = .630$  and  $.422$ ), the relationships of L2 speaking with comprehensibility and coherence should be considered tentative.

## 6.4 Relative strengths of relationships between L2 speaking and internal features

The strengths of the correlations differed across the features. General internal features assessed using rating scales (i.e., fluency, grammar, vocabulary, pronunciation, delivery, and content) had similarly strong correlations with L2 speaking ( $r = .713$  to  $.888$ ). This suggests that L2 speaking is supported by various internal features. As mentioned above, following Hulstijn's (2015) model, the current study conceptualizes vocabulary, grammar, pronunciation, and fluency as core components that contribute to L2 speaking. The results support the predictions based on the model. While the current study did not include periphery components and could not compare the relative strengths of core and periphery correlations with L2 speaking, the strong correlations with the core components partially support the core-periphery model.

We also found that speed fluency, breakdown fluency, grammatical complexity (particularly unit length), and lexical diversity, all of which were assessed using ratio scales, had similarly moderate or strong degrees of correlations ( $r = .326$  to  $.472$ ). These measures elicit some but not all aspects of fluency, grammatical complexity, and lexical complexity. Given strong correlations between L2 speaking and grammar, vocabulary, and pronunciation, which are related to accuracy, we would argue that all of the CAF constructs are shown to be essential contributors to L2 speaking, which supports Housen et al. (2012).

## 7. Conclusion

This study meta-analytically examined the relationships between L2 speaking (assessed using global ratings) and various oral internal features (assessed using analytic ratings or measures). Regarding Research Question 1, we found that the synthesized correlation between L2 speaking and internal features overall was significant and strong ( $r = .649$ ). Regarding Research Question 2, we found that the degrees of correlations varied across individual internal features: fluency, grammar, vocabulary, pronunciation, delivery, and content, all of which were assessed using rating scales, were significantly and strongly correlated to L2 speaking ( $r = .713$  to  $.888$ ). Some internal features assessed using ratio measures had significant correlations with L2 speaking that were weak or moderate: grammatical complexity ( $r = .326$ , particularly unit length of  $.392$ ), breakdown fluency ( $r = .361$ ), speed fluency ( $r = .468$ ) and lexical diversity ( $r = .472$ ).

Three main limitations should be noted. First, a limited number of primary studies were included in the meta-analysis. Although we made every effort to obtain as many studies (published by October 2018) as possible, we could only identify 15 studies with 35 correlations for speed fluency; this constituted the largest number of studies across internal features. Future primary studies can examine the differences in each internal feature's contributions to L2 speaking more precisely by considering more internal features and measures (e.g., pause frequency and location in Tavakoli & Wright, 2020; phrasal grammatical complexity in Bulté & Housen, 2012); conducting detailed analyses; and considering the effects of moderator variables such as different speaking task types, proficiency levels, and publication status. For instance, lexical sophistication was not significantly correlated with L2 speaking in the meta-analysis ( $r = -.074$ ). We predict that the contribution of lexical sophistication will vary across speaking task types. In nonacademic speaking tests, learners may be able to complete the task without using sophisticated vocabulary, whereas academic tasks may require it and lexical sophistication derived from such tasks may be more related to L2 speaking scores.

Second, we used Pearson or Spearman correlations, which assume linearity, between L2 speaking and its internal features. Nonlinear relationships may explain low correlations in the meta-analysis. For example, speed fluency has been reported to increase as L2 speaking proficiency progresses until a certain level at which point it levels off (Tavakoli et al., 2020), which is similar to grammatical complexity, especially sentential subordination (Bulté & Housen, 2012).

Third, this study examined each feature and its relationship with L2 speaking independently. Thus, it would be necessary to conduct meta-analytic structural equation modeling (Jak, 2015) to examine how each feature jointly predicts L2 speaking. As the features are correlated (e.g., for intercorrelations between fluency measures, see De Jong, 2018), such an analysis would deepen our understanding of the construct of L2 speaking.

To the authors' knowledge, this is the first meta-analysis that covers a wide range of internal features and reports on how each feature is related to L2 speaking. The results have emphasized the importance of internal features in L2 speaking and have pointed to areas in need of further research. One practical implication of the results is that teachers and researchers should focus more on internal features that correlate substantially with L2 speaking (e.g., more on speed and breakdown fluency rather than repair fluency or accuracy); they should consider including these features when assessing L2 speaking (e.g., by including related descriptors in a rating scale) and using rating scales and ratio measures with strong correlations to L2 speaking in their research. Another implication is that, as ratio measures tend to assess narrower constructs than rating scales, researchers should carefully consider

the extent to which the constructs of interests are well represented when using the former type of measures. This may be relevant to the discussion on the constructs measured in automated speaking assessments, as suggested by a reviewer. Our findings can be refined by updating the meta-analysis. Our study provides a stepping stone for such further rigorous investigations.

## References

- Adams, M. L. (1980). Five cooccurring factors in speaking proficiency. In J. R. Firth (Ed.), *Measuring spoken language proficiency* (pp. 1–6). Georgetown University Press.
- Bachman, L., & Palmer, A. (2010). *Language assessment in practice*. Oxford University Press.
- Borenstein, M. (2019). *Common mistakes in meta-analysis and how to avoid them*. Biostat.
- Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 21–46). John Benjamins. <https://doi.org/10.1075/llt.32.02bul>
- Cao, H. (2014). *Disentangling fluency, comprehensibility and coherence: Toward a better understanding of oral proficiency profiles* (Doctoral dissertation). Retrieved on 12 January 2022 from [https://docs.lib.purdue.edu/open\\_access\\_dissertations/239](https://docs.lib.purdue.edu/open_access_dissertations/239)
- Clark, J. L. D., & Swinton, S. S. (1980). *The Test of Spoken English as a measure of communicative ability in English-medium instructional settings* (TOEFL Research Report, RR 80–33). <https://doi.org/10.1002/j.2333-8504.1980.tb01230.x>
- Cucchiari, C., Strik, H., & Boves, L. (2000). Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *Journal of the Acoustical Society of America*, 107(2), 989–999. <https://doi.org/10.1121/1.428279>
- De Jong, N. (2018). Fluency in second language testing: Insights from different disciplines. *Language Assessment Quarterly*, 15(3), 237–254. <https://doi.org/10.1080/15434303.2018.1477780>
- De Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R., & Hulstijn, J. H. (2013). Linguistic skills and speaking fluency in a second language. *Applied Psycholinguistics*, 34(5), 893–916. <https://doi.org/10.1017/S0142716412000069>
- Ellis, R., & Yuan, F. (2004). The effects of planning on fluency, complexity and accuracy in second language narrative writing. *Studies in Second Language Acquisition*, 26(1), 59–84. <https://doi.org/10.1017/S0272263104026130>
- Farnsworth, T. L. (2013). An investigation into the validity of the TOEFL iBT Speaking Test for international teaching assistant certification. *Language Assessment Quarterly*, 10(3), 274–291. <https://doi.org/10.1080/15434303.2013.769548>
- Fisher, Z., & Tipton, E. (2015). Robust variance meta-regression (Version 2.0) [Software]. Retrieved on 12 January 2022 from <http://cran.r-project.org/web/packages/robmeta/robmeta.pdf>
- Foster, P. (2020). Oral fluency in a second language: A research agenda for the next ten years. *Language Teaching*, 53(4), 446–461. <https://doi.org/10.1017/S026144482000018X>
- Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, 21(3), 354–375. <https://doi.org/10.1093/applin/21.3.354>
- Foster, P., & Wigglesworth, G. (2016). Capturing accuracy in second language performance: The case for a weighted clause ratio. *Annual Review of Applied Linguistics*, 36, 98–116. <https://doi.org/10.1017/S0267190515000082>

- Freed, B. (1995). What makes us think that students who study abroad become fluent? In B. F. Freed (Ed.), *Second language acquisition in a study abroad context* (pp. 123–148). John Benjamins. <https://doi.org/10.1075/sibil.9.o9fre>
- Fulcher, G. (2003). *Testing second language speaking*. Pearson Education.
- Gan, Z. (2008). Extroversion and group oral performance: A mixed quantitative and discourse analysis approach. *Prospect*, 23(3), 24–42.
- Gan, Z. (2012). Complexity measures, task type, and analytic evaluations of speaking proficiency in a school-based assessment context. *Language Assessment Quarterly*, 9(2), 133–151. <https://doi.org/10.1080/15434303.2010.516041>
- Housen, A., Kuiken, F., & Vedder, I. (2012). Complexity, accuracy and fluency. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 1–20). John Benjamins. <https://doi.org/10.1075/llt.32.01hou>
- Hulstijn, J. H. (2015). *Language proficiency in native and non-native speakers: Theory and practice*. John Benjamins. <https://doi.org/10.1075/llt.41>
- Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, 29(1), 24–49. <https://doi.org/10.1093/applin/amm017>
- Jak, S. (2015). *Meta-analytic structural equation modelling*. Springer. <https://doi.org/10.1007/978-3-319-27174-3>
- Jin, T., & Mak, B. (2013). Distinguishing features in scoring L2 Chinese speaking performance: How do they work? *Language Testing*, 30(1), 23–47. <https://doi.org/10.1177/0265532212442637>
- Koizumi, R. (2005b). Speaking performance measures of fluency, accuracy, syntactic complexity, and lexical complexity. *JABAET (Japan-Britain Association for English Teaching) Journal*, 9, 5–33.
- Koizumi, R. (2013). Vocabulary and speaking. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* [online edition]. John Wiley and Sons. <https://doi.org/10.1002/9781405198431.wbeal1431>
- Koizumi, R., & In'nami, Y. (2012). Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens. *System*, 40, 554–564. <https://doi.org/10.1016/j.system.2012.10.012>
- Koizumi, R., & Kurizaki, I. (2002). Nihonjin chugakusei no monorogu niokeru supikingu no tokucho [Speaking characteristics of monologues given by Japanese junior high school students]. *Bulletin of the Kanto-Koshin-Etsu English Language Education Society*, 16, 17–28. [https://doi.org/10.20806/katejo.16.o\\_17](https://doi.org/10.20806/katejo.16.o_17)
- Koizumi, R., & Yamanouchi, I. (2003). Nihonjin chugakusei no supikingu noryoku no hattatsu [Development in speaking ability among Japanese junior high school students: Using self-introduction task]. *Bulletin of the Kanto-Koshin-Etsu English Language Education Society*, 17, 33–44. [https://doi.org/10.20806/katejo.17.o\\_33](https://doi.org/10.20806/katejo.17.o_33)
- Kormos, J. (2006). *Speech production and second language acquisition*. Lawrence Erlbaum Associates.
- Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, 32(2), 145–164. <https://doi.org/10.1016/j.system.2004.01.001>
- Kyle, K., Crossley, S. A., & Jarvis, S. (2021). Assessing the validity of lexical diversity indices using direct judgements. *Language Assessment Quarterly*, 18(2), 154–170. <https://doi.org/10.1080/15434303.2020.1844205>
- Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. The MIT Press.



- Li, S. (2016). The construct validity of language aptitude: A meta-analysis. *Studies in Second Language Acquisition*, 38(4), 801–842. <https://doi.org/10.1017/S027226311500042X>
- Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, 96(2), 190–208. [https://doi.org/10.1111/j.1540-4781.2011.01232\\_1.x](https://doi.org/10.1111/j.1540-4781.2011.01232_1.x)
- Luoma, S. (2004). *Assessing speaking*. Cambridge University Press.  
<https://doi.org/10.1017/CBO9780511733017>
- Malvern, D., & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing*, 19(1), 85–104.  
<https://doi.org/10.1191/0265532202lt2210a>
- Milton, J., Wade, J., & Hopkins, N. (2010). Aural word recognition and oral competence in a foreign language. In R. Chacón-Beltrán, C. Abello-Contesse, & M. Torreblanca-López (Eds.), *Further insights into non-native vocabulary teaching and learning* (pp. 83–98). Multilingual Matters.  
<https://doi.org/10.21832/9781847692900-007>
- Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, 30(4), 555–578.  
<https://doi.org/10.1093/applin/amp044>
- Ockey, G. J., Koyama, D., Setoguchi, E., & Sun, A. (2015). The extent to which TOEFL iBT speaking scores are associated with performance on oral language tasks and oral ability components for Japanese university students. *Language Testing*, 32(1), 39–62.  
<https://doi.org/10.1177/0265532214538014>
- Orwin, R. (1983). A fail-safe N for effect size in meta-analysis. *Journal of Educational Statistics*, 8(2), 157–159. <https://doi.org/10.2307/1164923>
- Pietilä, P. (1999). L2 speech: Oral proficiency of students of English at university level. *Anglicana Turkuensia*, 19, 1–80.
- Plonsky, L., & Oswald, F. L. (2014). How big is “big”? Interpreting effect sizes in L2 research. *Language Learning*, 64(4), 878–912. <https://doi.org/10.1111/lang.12079>
- Révész, A., Ekiert, M., & Torgersen, E. N. (2016). The effects of complexity, accuracy, and fluency on communicative adequacy in oral task performance. *Applied Linguistics*, 37(6), 828–848.  
<https://doi.org/10.1093/applin/amu069>
- Rosenthal, R. (1979). The “file drawer problem” and tolerance for null results. *Psychological Bulletin*, 86(3), 638–641. <https://doi.org/10.1037/0033-2909.86.3.638>
- Saito, K., Ilkan, M., Magne, V., Tran, M., & Suzuki, S. (2018). Acoustic characteristics and learner profiles of low, mid and high-level second language fluency. *Applied Psycholinguistics*, 39(3), 593–617. <https://doi.org/10.1017/S0142716417000571>
- Saito, K., Trofimovich, P., & Isaacs, T. (2017). Using listener judgements to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics*, 38(4), 439–462. <https://doi.org/10.1093/applin/amv047>
- Saito, K., Webb, S., Trofimovich, P., & Isaacs, T. (2016). Lexical correlates of comprehensibility versus accentedness in second language speech. *Bilingualism: Language and Cognition*, 19(3), 597–609. <https://doi.org/10.1017/S1366728915000255>
- Sato, T. (2012). The contribution of test-takers' speech content to scores on an English oral proficiency test. *Language Testing*, 29(2), 223–241. <https://doi.org/10.1177/0265532211421162>
- Segalowitz, N. (2010). *Cognitive bases of second language fluency*. Routledge.  
<https://doi.org/10.4324/9780203851357>
- Segalowitz, N., & Freed, B. F. (2004). Context, contact, and cognition in oral fluency acquisition: Learning Spanish in at home and study abroad contexts. *Studies in Second Language Acquisition*, 26(2), 173–199. <https://doi.org/10.1017/S0272263104262027>

- Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, 30(4), 510–532. <https://doi.org/10.1093/applin/amp047>
- Suzuki, S., Kormos, J., & Uchihara, T. (2021). The relationship between utterance and perceived fluency: A meta-analysis of correlational studies. *The Modern Language Journal*, 105(2), 435–463. <https://doi.org/10.1111/modl.12706>
- Tavakoli, P., Nakatsuhara, F., & Hunter, A.-M. (2020). Aspects of fluency across assessed levels of speaking proficiency. *The Modern Language Journal*, 104(1), 169–191. <https://doi.org/10.1111/modl.12620>
- Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure, and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239–276). John Benjamins. <https://doi.org/10.1075/llt.11.15tav>
- Tavakoli, P., & Wright, C. (2020). *Second language speech fluency*. Cambridge University Press. <https://doi.org/10.1017/9781108589109>
- Ushigusa, S. (2008). *The relationships between oral fluency, multiword units, and proficiency scores* (Doctoral dissertation). Retrieved from UMI. (Order No. 3344157)
- von Hippel, P. T. (2015). The heterogeneity statistic  $I^2$  can be biased in small meta-analyses. *BMC Medical Research Methodology*, 15(35), 1–8. <https://bmcmmedresmethodol.biomedcentral.com/articles/10.1186/s12874-015-0024-z>. <https://doi.org/10.1186/s12874-015-0024-z>
- Xi, X., & Mollaun, P. (2006). Investigating the utility of analytic scoring for the TOEFL Academic Speaking Test (TAST). (TOEFL iBT Research Report, RR-06-07). <http://www.ets.org/Media/Research/pdf/RR-06-07.pdf>
- Yan, X., Kim, H. R., & Kim, J. Y. (2018). *Complexity, accuracy and fluency (CAF) features of speaking performances on Aptis across different levels on the Common European Framework of Reference (CEFR)*. ARAGs Research Reports online. British Council. Retrieved on 12 January from <https://www.britishcouncil.org/complexity-accuracy-and-fluency-caf-features-speaking-performances-aptis-across-different-levels>
- Zenker, F., & Kyle, K. (2021). Investigating minimum text lengths for lexical diversity indices. *Assessing Writing*, 47, 1–15. <https://doi.org/10.1016/j.asw.2020.100505>

## Appendix A. 39 studies included in the meta-analysis

- Baker-Smemoe, W., Dewey, D. P., Bown, J., & Martinsen, R. A. (2014). Does measuring L2 utterance fluency equal measuring overall L2 proficiency? Evidence from five languages. *Foreign Language Annals*, 47, 707–728. <https://doi.org/10.1111/flan.12110>
- Cao, H. (2014). *Disentangling fluency, comprehensibility and coherence: Toward a better understanding of oral proficiency profiles* (Doctoral dissertation). Retrieved from [https://docs.lib.purdue.edu/open\\_access\\_dissertations/239](https://docs.lib.purdue.edu/open_access_dissertations/239)
- Christoffersen, K. O. (2017). Comparing native speaker ratings and quantitative measures of oral proficiency in IELTS interviews. *ELIA*, 17, 233–250. <https://doi.org/10.12795/elia.2017.i17.10>
- Clark, J. L. D., & Swinton, S. S. (1979). *An exploration of speaking proficiency measures in the TOEFL context*. Research Reports, 4. Educational Testing Service. <https://doi.org/10.1002/j.2333-8504.1979.tb01176.x>
- Clark, J. L. D., & Swinton, S. S. (1980). *The Test of Spoken English as a measure of communicative ability in English-medium instructional settings* (TOEFL Research Report, RR 80–33). <https://doi.org/10.1002/j.2333-8504.1980.tb01230.x>

- Cox, T. L. (2013). *Investigating prompt difficulty in an automatically scored speaking performance assessment*. (Doctoral dissertation). Retrieved from on 12 January 2022 from <https://pdfs.semanticscholar.org/55a4/92347a2650363b7be03ee4a0b861041a92e5.pdf>
- Crossley, S., & McNamara, D. (2013). Applications of text analysis tools for spoken response grading. *Language Learning & Technology*, 17(2), 171–192. <https://doi.org/10.125/44329>
- Freed, B. (1995). What makes us think that students who study abroad become fluent? In B. F. Freed (Ed.), *Second language acquisition in a study abroad context* (pp. 123–148). John Benjamins. <https://doi.org/10.1075/sibil.9.09fre>
- Fujimori, C., & Koizumi, R. (2011). Supichi happyo no ALT to JTE niyoru zentaitekihyoka: Kyakkanteki sokutei oyobi seito sogohyoka tono kanren [Correlations between holistic ratings of speech presentation by the ALT and the JTE, and objective measures and students' peer evaluation]. *KATE Bulletin*, 25, 21–31. [https://doi.org/10.20806/katejournal.25.o\\_21](https://doi.org/10.20806/katejournal.25.o_21)
- Gan, Z. (2008). Extroversion and group oral performance: A mixed quantitative and discourse analysis approach. *Prospect*, 23(3), 24–42.
- Gan, Z. (2012). Complexity measures, task type, and analytic evaluations of speaking proficiency in a school-based assessment context. *Language Assessment Quarterly*, 9, 133–151. <https://doi.org/10.1080/15434303.2010.516041>
- Ginther, A., Dimova, S., & Yang, R. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing*, 27, 379–399. <https://doi.org/10.1177/0265532210364407>
- Henning, G. (1983). Oral proficiency testing: Comparative validities of interview, imitation, and completion methods. *Language Learning*, 33, 315–332. <https://doi.org/10.1111/j.1467-1770.1983.tb00544.x>
- Hsieh, C.-N. (2011). *Rater effects in ITA testing: ESL teachers' versus American undergraduates' judgements of accentedness, comprehensibility, and oral proficiency* (Unpublished doctoral dissertation). Michigan State University.
- Iwashita, N. (2006). Syntactic complexity measures and their relation to oral proficiency in Japanese as a Foreign Language. *Language Assessment Quarterly*, 3, 151–169. [https://doi.org/10.1207/s15434311laq0302\\_4](https://doi.org/10.1207/s15434311laq0302_4)
- Jin, T., & Mak, B. (2013). Distinguishing features in scoring L2 Chinese speaking performance: How do they work? *Language Testing*, 30, 23–47. <https://doi.org/10.1177/0265532212442637>
- Kahng, J. (2014). Exploring utterance and cognitive fluency of L1 and L2 English speakers: Temporal measures and stimulated recall. *Language Learning*, 64, 809–854. <https://doi.org/10.1111/lang.12084>
- Katagiri, K. (1999). Evaluation of L2 speaking: JTEs versus AETs. *Annual Review of English Language Education in Japan*, 10, 93–102.
- Kim, H.-J. (2006). Rater reliability in L2 oral proficiency tests. *English Teaching*, 61(3), 105–118.
- Kim, M. (2001). Detecting DIF across the different language groups in a speaking test. *Language Testing*, 18, 89–114. <https://doi.org/10.1177/026553220101800104>
- Kim, Y.-M. (2010). Spoken corpora of EFL learners: Collocations, vocabulary use, and their oral proficiency levels. *Multimedia-Assisted Language Learning*, 13(1), 9–28. <https://doi.org/10.15702/mall.2010.13.1.9>
- Koizumi, R. (2005a). Predicting speaking ability from vocabulary knowledge. *JLTA (Japan Language Testing Association) Journal*, 7, 1–20. [https://doi.org/10.20622/jltaj.7.o\\_1](https://doi.org/10.20622/jltaj.7.o_1)
- Koizumi, R. (2005b). Speaking performance measures of fluency, accuracy, syntactic complexity, and lexical complexity. *JABAET (Japan-Britain Association for English Teaching) Journal*, 9, 5–33. Retrieved on 12 January 2022 from [http://www7b.biglobe.ne.jp/~koizumi/JABAET\\_Speaking\\_Performance\\_Measures\\_Koizumi.pdf](http://www7b.biglobe.ne.jp/~koizumi/JABAET_Speaking_Performance_Measures_Koizumi.pdf)

- Koizumi, R., & Kurizaki, I. (2002). Nihonjin chugakusei no monorogu niokeru supikingu no tokucho [Speaking characteristics of monologues given by Japanese junior high school students]. *Bulletin of the Kanto-Koshin-Etsu English Language Education Society*, 16, 17–28. [https://doi.org/10.20806/katejo.16.o\\_17](https://doi.org/10.20806/katejo.16.o_17)
- Koizumi, R., & Yamanouchi, I. (2003). Nihonjin chugakusei no supikingu noryoku no hattatsu [Development in speaking ability among Japanese junior high school students: Using self-introduction task]. *Bulletin of the Kanto-Koshin-Etsu English Language Education Society*, 17, 33–44. [https://doi.org/10.20806/katejo.17.o\\_33](https://doi.org/10.20806/katejo.17.o_33)
- Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, 96, 190–208. <https://doi.org/10.1111/j.1540-4781.2011.01232.1.x>
- Malvern, D., & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing*, 19, 85–104. <https://doi.org/10.1191/0265532202lt2210a>
- Metruk, R. (2018). Comparing holistic and analytic ways of scoring in the assessment of speaking skills. *Journal of Teaching English for Specific and Academic Purposes*, 6(1), 179–189. <https://doi.org/10.22190/JTESAP1801179M>
- Naito, T. (1995). Speech niokeru analytic evaluation to holistic evaluation [Analytic and holistic evaluation in speech assessment]. *STEP BULLETIN*, 7, 146–151. [https://www.eiken.or.jp/center\\_for\\_research/list\\_1/archives/](https://www.eiken.or.jp/center_for_research/list_1/archives/)
- Negishi, J. (2011). *Characteristics of group oral interactions performed by Japanese learners of English*. (Publication No. 5722) (Doctoral dissertation) Waseda University. Waseda University Repository. Retrieved on 12 January 2022 from <http://hdl.handle.net/2065/37662>
- Pietilä, P. (1999). L2 speech: Oral proficiency of students of English at university level. *Anglicana Turkuensia*, 19, 1–80.
- Robin, R. M. (2012). Lexicalized aspectual usage in Oral Proficiency Interviews. *The Modern Language Journal*, 96, 34–50. <https://doi.org/10.1111/j.1540-4781.2012.01292.x>
- Sato, T. (2012). The contribution of test-takers' speech content to scores on an English oral proficiency test. *Language Testing*, 29, 223–241. <https://doi.org/10.1177/0265532211421162>
- Segalowitz, N., & Freed, B. F. (2004). Context, contact, and cognition in oral fluency acquisition: Learning Spanish in at home and study abroad contexts. *Studies in Second Language Acquisition*, 26, 173–199. <https://doi.org/10.1017/S0272263104262027>
- Ushigusa, S. (2008). *The relationships between oral fluency, multiword units, and proficiency scores* (Doctoral dissertation). Retrieved from UMI. (Order No. 3344157)
- Usuda, Y. (2002). Oral Production Test no hyoka to hatsuwaryo no kankei: Hatsuwa ryo wa hyokashakudo no hitotsu ni narieruka [The relationship between students' oral test scores and the quantity of speech they produce: Can the amount of speech be used as a criterion?]. *Hakodate Eibungaku: Journal of the English Literary Society of Hakodate*, 41, 45–60.
- Xi, X. (2007). Validating TOEFL® iBT Speaking and setting score requirements for ITA screening. *Language Assessment Quarterly*, 4, 318–351. <https://doi.org/10.1080/15434300701462796>
- Xi, X., Higgins, D., Zechner, K., & Williamson, D. (2012). A comparison of two scoring methods for an automated speech scoring system. *Language Testing*, 29, 371–394. <https://doi.org/10.1177/0265532211425673>
- Xi, X., & Mollaun, P. (2006). Investigating the utility of analytic scoring for the TOEFL Academic Speaking Test (TAST). (TOEFL iBT Research Report, RR-06-07). <http://www.ets.org/Media/Research/pdf/RR-06-07.pdf>

*Note.* The following was found in the last production stage and could not be included in the current meta-analysis.

Ahmadi, A., & Sadeghi, E. (2016). Assessing English language learners' oral performance: A comparison of monologue, interview, and group oral test. *Language Assessment Quarterly*, 13(4), 341–358. <https://doi.org/10.1080/15434303.2016.1236797>

Appendix B. Coding sheet

Construct	Example
L2 speaking	Large-scale proficiency tests (e.g., FSI global rating; IELTS speaking section; Oral English Proficiency Test [OEPT] at Purdue University; Oral Proficiency Interview [OPI]; Simulated Oral Proficiency Interview [SOPI]; Speaking Proficiency English Assessment Kit [SPEAK]; Test for English Majors Band 4 (TEM-4) speaking section; TOEFL speaking section; Test of Spoken English [TSE]), author-made speaking tests
Internal feature	
Fluency (RS)	Fluency rating scales
Speed fluency (RM)	Speech rate: 1. Syllable/response time (the number of syllables divided by response time length) 2. Word/response time (the number of words divided by response time length)  Articulation rate:  Syllable/phonation time (the number of syllables divided by phonation time length [i.e., response time minus time with filled pauses]) Mean length of run: the number of words divided by the number of runs [i.e., utterances between silent pauses (e.g., pauses of 0.25 seconds or more)]  Others:  Phonation time/syllable The number of words in the longest fluent speech run, not containing any silent or filled disfluencies Mean length of run in words containing no silent pauses Mean length of run in words containing no filled pause disfluencies
Breakdown fluency (RM)	Phonation time ratio: Phonation time/response time (phonation time length divided by response time length) Pause ratio: Silent pause time/response time (silent pause time length divided by response time length) Pause length: Silent pause time/silent pauses (silent pause time length divided by the number of silent pauses)

Construct	Example
	<p>Others:</p> <p>Silent (or filled) pause/phonation time (the number of silent [or filled] pause divided by phonation time length)</p> <p>Silent pause/response time</p> <p>Silent pause/word</p> <p>Relative frequency of long pauses</p> <p>Silent pause rate at a clause boundary</p> <p>Silent pause rate at an AS-unit boundary</p> <p>Filled pause rate within a clause</p> <p>Filled pause rate at a clause boundary</p> <p>Filled pause rate at an AS-unit boundary</p> <p>Silent pause rate within a clause</p> <p>Proportion of er, as well as the unfilled pauses produced by each participant</p>
Repair fluency (RM)	<p>Disfluency rate (e.g., the number of disfluencies [e.g., repetitions, self-corrections] divided by total response time or by the number of units or words)</p> <p>Others:</p> <p>Disfluency marker/"the number of pruned tokens per minute"</p> <p>Pruned token/unpruned token</p>
Accuracy (RM)	<p>Error-free clause ratio: Error-free clause/clause (the number of error-free clauses divided by the number of clauses)</p> <p>Error-free unit ratio: Error-free unit/unit (the number of error-free units divided by the number of units)</p> <p>Error ratio: Error/word (the number of errors divided by the number of words)</p> <p>Others:</p> <p>Error/unit</p> <p>Proportion of imperfective</p> <p>Proportion of perfective</p> <p>Proportion of correct imperfective</p> <p>Proportion of correct perfective</p> <p>Proportion of correct aspectual choices</p> <p>Target-like syllables per 10 syllables (pronunciation)</p>
Grammar (RS)	Grammar rating scales
Vocabulary (RS)	Vocabulary rating scales
Grammatical complexity (RM)	The number of words divided by the number of units
Unit length	<p>Others:</p> <p>The number of words divided by the number of utterances (word/utterance)</p>
Sentential subordination	<p>The number of clauses divided by the number of units</p> <p>Others:</p> <p>Dependent clause (i.e., subordinate clause)/unit</p> <p>Independent clause/unit</p> <p>S-nodes/unit</p> <p>Subordinate conjunctions/unit</p> <p>Coordinate conjunctions/unit</p>

Construct	Example
Other aspects of lexical complexity	Verb phrase/unit Word/clause Dependent clause/clause Independent clause/clause Verb phrase/clause Complex noun phrase/noun phrase
Lexical complexity (RM)	1. Guiraud's index: the number of word types divided by the root square of the number of tokens
Lexical diversity	2. D Others (all from Lu, 2012): Bilogarithmic TTR Uber index Lexical word variation Verb variation 1 (VV1) Squared VVI Corrected VV1 Verb variation 2 Noun variation Adjective variation Adverb variation Modifier variation
Lexical density	The number of lexical words (tokens) divided by the number of tokens (Lexical word/token) The number of lexical words divided by the square root of the number of tokens
Lexical sophistication	The number of sophisticated (i.e., less frequent) word types divided by the number of types (Sophisticated word type/type) Sophisticated lexical word type/lexical word type Sophisticated word type/token Sophisticated word token + (0.5 * Basic word token)/token Others (all from Lu, 2012): Verb sophistication 1 (VS1) Corrected VS1 Verb sophistication 2
Other aspects of lexical complexity	Unique words normalized by speech duration Unique words normalized by total word duration Others (all from Crossley & McNamara, 2013): Word polysemy (vocabulary depth) Word meaningfulness all word (MRC database) Word familiarity content words (MRC Database) Word imageability content words (MRC database)
Pronunciation (RS), Comprehensibility (RS), Delivery (RS), Content (RS), and Coherence (RS): rating scales of each feature	

Note. RS = Rating scale. RM = Ratio measure. Word = Pruned tokens unless noted. MRC = Medical Research Council. Units include T-units, c-units, and AS-units.



## L2 speaking and its external correlates

### A meta-analysis

Eun Hee Jeon, Yo In'nami and Rie Koizumi

University of North Carolina at Pembroke / Chuo University / Seisen University

The present study synthesizes the correlation coefficients between discourse-level L2 speaking and nine relevant correlates: L2 vocabulary knowledge, L2 grammar knowledge, working memory, L2 reading comprehension, L2 listening comprehension, L2 writing, language aptitude, metacognition, and anxiety. Fifty-one independent samples from 41 primary studies published between 1975 and 2017 were included in the study. Correlations weighted for sample size and, when possible, corrected for attenuation due to measurement errors were submitted to a series of nine quantitative meta-analyses. The results showed that L2 knowledge variables were strongly correlated with L2 speaking (.563 for vocabulary and .622 for grammar). The same results were found for companion L2 proficiency variables such as L2 reading comprehension (.822), L2 listening comprehension (.530), and L2 writing (.605). As for the language general variables, metacognition was much more strongly correlated with L2 speaking (.624) compared to working memory (.102). Lastly, language aptitude and anxiety each showed a positive and a negative correlation falling in the medium to large range: .481 for language aptitude and  $-.432$  for anxiety.

#### 1. Introduction

What does it mean to be a strong L2 speaker? What cognitive, linguistic, and affective variables constitute, correlate with, and influence L2 speaking abilities? How do these variables operate and coordinate with one another as we speak in L2? These questions have been investigated by Second Language Acquisition (SLA) research (e.g., Iwashita, Brown, McNamara, & O'Hagan, 2008; Norris & Ortega, 2000; Ortega, 2003), language testing research (e.g., De Jong, Steinel, Florijn, Schoonen, & Hulstijn, 2012), and speech production modeling research (Kormos, 2006; Levelt, 1989). While studies vary in their methodological details such as measurement characteristics (e.g., subjective vs. objective measurement; see De Jong et al. [2012] for a detailed discussion of subjective-subjective approach and subjective-objective

approach in L2 speaking component research) or traits of oral proficiency under investigation (e.g., accuracy, vocabulary, syntactic complexity, pronunciation, fluency, appropriacy), the general consensus is that L2 speaking proficiency is comprised of linguistic knowledge components and processing components. Linguistic knowledge refers to various types of knowledge representations such as phonological, lexical, morpho-syntactic, syntactic, and discourse knowledge. This knowledge, however, cannot be rendered into fluent, accurate, complex, and appropriate impromptu speech without efficient processing components such as attention and working memory at play.

Of the studies that take such a componential approach to L2 speaking, two main approaches (i.e., internal and external) can be found: (1) studies that investigate the relationship between the quality of subjectively or objectively measured L2 speech outcome and a trait or traits of the speech outcome (e.g., lexical quality of the very speech sample which was holistically assessed by a judge), and (2) studies that investigate the relationship between the quality of subjectively or objectively measured L2 speech outcome and a component (linguistic, cognitive, affective, conative, or otherwise) or components that are measured independently of the earlier mentioned L2 speech outcome. As the second of the two-part meta-analyses on L2 speaking, the present meta-analysis synthesizes the findings of the second group of studies. In the following section, we provide a review of each of the variables included in the present meta-analysis, focusing on L2 knowledge variables (L2 vocabulary knowledge and L2 grammar knowledge), L2 proficiency variables (L2 reading comprehension, L2 listening comprehension, and L2 writing), language-general variables (working memory, metacognition, and language aptitude), and anxiety.

## 2. Review of variables

### 2.1 Vocabulary knowledge

Along with grammar, L2 vocabulary knowledge is an essential component of L2 speaking. Both Levelt's (1989) L1 speech production model and Kormos's (2006) L2 speech production model propose that speech production comprises four main stages: conceptualization, formulation, articulation, and self-monitoring. The models further propose that the first three stages are subject to automatization and can run concurrently. In the conceptualization stage, the speaker conceives preverbal messages. In the next, formulation stage, the speaker retrieves necessary vocabulary and other knowledge to encode preverbal messages. Finally, in the articulation stage, the speaker articulates encoded utterances. In the case of L1 speech production, encoding and articulation are carried out with minimal efforts

as they are typically automatized, and therefore can run parallel while consuming little processing resources. In L2 speech production, however, these processes are more effortful and, therefore, can lead to more variable outcomes depending on the speaker's level of linguistic knowledge and processing efficiency (Kormos, 2006). Accordingly, Koizumi and In'nami (2013) noted that not only vocabulary size and depth, but also lexical processing efficiency should be incorporated in the theoretical frameworks of vocabulary knowledge.

The importance of these different dimensions of L2 vocabulary knowledge in L2 speech production is well-documented in the growing body of relevant research. Lexical knowledge seems to be directly related to fluency, a key aspect of speaking proficiency. Individual difference studies examining L2 speakers (e.g., Kormos & Denés, 2004; see also Chapter 9 for a detailed discussion on how lexical knowledge and fluency are related) have reported that less proficient L2 speakers often struggle with lexical retrieval, which negatively affects perceived fluency. Conversely, L2 speakers with higher lexical access efficiency not only showed stronger oral performance (Qian & Lin, 2020), but the higher initial values of lexical access efficiency led to higher gains in L2 production over time (Segalowitz & Freed, 2004). Studies also report that the speech production of fluent L2 speakers features higher accuracy and lexical diversity compared to their less proficient peers (Kormos & Denés, 2004). Given the multidimensional nature of vocabulary knowledge and their contributions to L2 speaking, our meta-analysis also takes an inclusive approach and includes all operationalizations of L2 vocabulary knowledge: e.g., size, depth, lexical access efficiency, etc.

## 2.2 Grammar knowledge

Grammar knowledge is another key component of linguistic knowledge. In the above-mentioned models of speech production proposed by Levelt (1989) and Kormos (2006), the speaker encodes in the formulation stage her preverbal messages using grammatical (e.g., morpho-syntactic, syntactic), lexical, and phonological knowledge. In line with this, De Jong and Van Ginkle (1992) defined grammatical knowledge as the type of linguistic knowledge that integrates basic elements of linguistic knowledge (e.g., lexical and phonological knowledge). As such, efficient morpho-syntactic (e.g., use of inflectional variants, use of relative pronouns, use of auxiliary verbs) and syntactic processing (e.g., use of knowledge of word order) is crucial for achieving speaking fluency (De Jong, Steinel, Florijn, Schoonen, & Hulstijn, 2013).

While there is a strong consensus on the importance of overall grammar knowledge as a contributor to L2 speaking, there is a less clear agreement on how different types of grammar knowledge, especially explicit vs. implicit knowledge, contributes

to overall speaking performance in L2. This problem is only compounded by the significant difficulty involved in the assessment of implicit vs. explicit knowledge of grammar using current forms of grammaticality judgment tests (GJTs) among L2 learners (Vafaei et al., 2017). Keeping this methodological caveat in mind, the results of some studies which attempted to examine the multifaceted nature of grammar knowledge are considered here; Ellis's (2013) factor analysis found that variables representing implicit grammar knowledge (imitation test, timed grammaticality judgment test) loaded on the same factor as oral narrative test while variables representing explicit grammar knowledge (untimed grammaticality judgement test, metalinguistic knowledge test) loaded on other factors. All variables, however, significantly correlated with oral narrative test. Gutiérrez (2013), on the other hand, found no significant correlations between explicit grammar knowledge-related variables (i.e., metalinguistic, metalingual) and L2 oral proficiency. As with vocabulary knowledge, our meta-analysis takes an inclusive approach and includes all variables related to grammar knowledge regardless of its type (e.g., implicit, explicit).

### 2.3 Three companion variables of L2 speaking: L2 reading comprehension, L2 listening comprehension, and L2 writing

Some L2 speaking studies included companion (as opposed to component) proficiency variables such as L2 reading comprehension, L2 listening comprehension, and L2 writing. In most cases, the primary purpose of including such companion proficiency variables was to examine the concurrent validity of a test under development (e.g., prototype tests under development for Test of English as a Foreign Language Internet-based Test [TOEFL iBT]) by correlating it with existing tests (e.g., TOEFL paper-based test [TOEFL PBT]), rather than to investigate a theoretical question about the relationship between companion proficiency variables. Apart from the theoretical motivation (or lack thereof) of these primary studies, we decided it would be useful to examine the relationship between L2 speaking and its companion proficiency variables as they represent different dimensions of L2 proficiency. For this reason, we included studies that included all measures of discourse-level L2 speaking, passage-level L2 reading comprehension, passage-level L2 listening comprehension, and discourse-level L2 writing.

## 2.4 Working memory

Successful L2 speaking, specifically, real-time oral performance, relies heavily on working memory. Working memory is a necessary resource as the speaker, having conceived her preverbal messages, accesses different linguistic knowledge representations (e.g., lexical, grammatical, discourse) and renders them available for message encoding and utterance articulation. It is also important to note that all of these speech production processes are subject to self-monitoring, another mechanism supported by working memory (e.g., in order to access linguistic knowledge representation and use it to detect errors in the encoding or utterance stages). In the case of unplanned, interactive speech events such as a conversation, the demands on the speaker's working memory become even higher as the speaker needs to comprehend aural input from the interlocutor(s), concurrently plan his next turn which ideally stays coherent with the on-going conversation and comply with the desired discourse styles of the given communicative context. Compared to L1 speech production, where much of the needed processes has been automatized and therefore demand little cognitive resources, the quality of L2 speech can heavily depend not only on linguistic knowledge itself, but also on processing components such as working memory. One quality of L2 speaking that has often been associated with working memory is fluency. Fehrer and Fry (2007a) reported that despite the high proficiency in L2, their study participants produced more hesitations in their L2 than in their L1, indicating that linguistic knowledge does not guarantee fluent L2 speech and that there is a higher cognitive load imposed by L2. Another study by the same authors (Fehrer & Fry, 2007b) showed that working memory is not only significantly linked to fluency, but also to syntactic complexity. As for the acceptable measures of working memory, we take an inclusive approach and include all measures tapping into all constructs related to working memory (e.g., processing, storage, processing, and storage).

## 2.5 Metacognition

Of the many dimensions of metacognition, perhaps one that is most pertinent to L2 speaking is self-monitoring. Kormos (2006) noted that self-monitoring can take place in all stages of speech production (i.e., formulation, encoding, and articulation) as the speaker attempts to preempt or repair errors. The speaker can detect an error in the preverbal message and correct it before it is encoded, the speaker can also detect errors in the encoding stage and correct it, and lastly, the speaker can detect and correct errors in the utterances as she speaks, or after she has made the utterance. As can be easily inferred from these scenarios, processing components such as working

memory and attention also play an important role alongside the metacognitive construct, self-monitoring. While it is evident that real time self-monitoring in L2 speaking is important, studying it is a challenging task. For example, as Kormos (2006) notes, covert repairs (i.e., noticing and repairing errors before the utterance is made) can only be reliably studied in highly controlled laboratory settings, which are not easily available to researchers. Verbal reports, a method which is often used to access metacognitive functions during a task, have also been criticized for being intrusive (Bowles, 2010). Perhaps for this reason, metacognition is only rarely studied as a component of L2 speaking (5 samples included in the present meta-analysis), all of which involved a self-report type data elicitation method inquiring about strategies related to L2 speaking.

## 2.6 Language aptitude

Unlike the variables reviewed above, language aptitude may not be so much a component of L2 speaking as a correlate or a contributor (or predictor) of L2 speaking proficiency. While the layman definition of aptitude, namely, a natural ability or “knack” for learning a language may seem deceptively simple, the theoretical and operational definition of language aptitude in Second Language Acquisition and language testing communities have seen much controversy and revision. Ehrman (1998) introduces some of the criticisms on the Modern Language Aptitude Test (MLAT) as follows: (1) MLAT is not a good measure for people who have not received formal education; (2) MLAT is the product of the audiolingual era and, therefore, is no longer appropriate for testing language aptitude given the highly communicative language learning settings of today. Perhaps the more important criticism and controversy were noted by Winke (2013) and Doughty (2019); they argued that it is problematic to rely on MLAT to comprehensively capture language aptitude as we have yet to reach the theoretical definition of language aptitude with some viewing it as a comprehensive and multidimensional construct encompassing cognitive, conative, affective, personality factors, and even language learning experience, while others considering cognitive aptitude only. With the advent of the High-Level Language Aptitude Battery (Hi-LAB), a new language aptitude battery partially borne out of this controversy, and with a hope to create more theoretically updated language aptitude tests, empirical studies have examined how the performances of the classic MLAT and Hi-LAB compare. Interestingly, Doughty’s (2019) study showed that both MLAT and Hi-LAB were good predictors of general professional proficiency and limited working proficiency. In keeping with the generally inclusive approach we take in this study, we included all studies that included any type of language aptitude measure along with a speaking proficiency measure.

## 2.7 Anxiety

Of the nine correlates of L2 speaking included in this study, anxiety is the only variable that is neither linguistic nor cognitive in nature. Nevertheless, anxiety, especially in the context of real time communication, plays an important role in speech production through its modulation of working memory. Vytal, Cornwell, Letkiewicz, Arkin, and Grillon (2013) investigated how verbal and spatial working memory is affected by anxiety while completing tasks which demand high, medium, and low levels of cognitive load among adults. The results showed that even not very demanding verbal memory tasks were negatively affected by anxiety. The authors further explained that verbal working memory processes share the same resources (e.g., neural circuitry, executive control) with anxious apprehension (i.e., anxiety). Increased anxiety, therefore, competes with verbal working memory for limited cognitive resources, resulting in decreased processing efficiency and accuracy (Eysenck & Calvo, 1992). The importance of anxiety is well acknowledged among L2 researchers, especially in relation with oral communication (Horwitz, Horwitz, & Cope, 1986). The type of anxiety which is most commonly associated with L2 speaking is state anxiety (Oya, Manalo, & Greenwood, 2004), and the empirical findings on the relationship between anxiety and L2 oral performance converge on Vytal et al.'s (2013) explanations; (1) state anxiety is related to higher rates of error (Oya et al., 2004); (2) oral performance produced in high-anxiety state was perceived by raters less expressive (Steinberg & Horwitz, 1986); (3) anxiety was significantly related with speech raters' judgment of fluency, syntactic complexity, and accentedness (MacIntyre & Gardner, 1994). It is also noteworthy that anxiety and the quality of L2 oral performance may form a feedback cycle; proficient L2 speakers have lower levels of anxiety, which then contributes to superior oral performance. Conversely, less proficient L2 speakers may have higher levels of anxiety and may face a twofold challenge posed by limited linguistic knowledge and reduced processing efficiency. Acknowledging the significance of anxiety as a correlate of L2 speaking, we included all studies which included measurements pertinent to all types of anxiety (e.g., trait anxiety, state anxiety, foreign language anxiety, strategy questionnaire about anxiety control).



### 3. Research question

The following research question was investigated in the present meta-analysis.

What are the strengths of association with discourse-level L2 speaking and the following nine correlates: L2 vocabulary knowledge, L2 grammar knowledge, L2 reading comprehension, L2 listening comprehension, L2 writing, working memory, metacognition, language aptitude, and anxiety?

### 4. Method

#### 4.1 Literature search and inclusion criteria

Major electronic databases (ERIC, LLBA, PsychINFO, PsycARTICLES, ProQuest), a search tool (Primo Search), 24 journals, and applied linguistics and linguistics section of a major university's library were either electronically or manually searched for all relevant studies. First, abstracts published between January 1975 and December 2017 were searched using various combinations of the following key terms: component\*, subcomponent\*, construct\*, subconstruct\*, correlate\*, oral communica\*, oral skill\*, speak\*, spee\*, spoken, phon\*, pronunciat\*, vocab\*, lexic\*, word\*, working memory, memory, grammar\*, syntac\*, morphosyntac\*, discours\*, speech produc\* communicative competence, associate\*, relat\*, aptitude, metacogni\*, reading, listening, writing. When abstracts indicated a correlational design, full texts were examined for information relevant to the present study (e.g., correlation tables, reliability reporting, descriptions of study features). After periodicals and electronic databases have been searched, we also visited a major university in the United States to manually search all books and monographs in the applied linguistics, linguistics, and education section of the library. As the search took place over the span of 2 years (from 2015 to 2017), it is difficult to report the exact number of sources identified as relevant at each step of the search. However, approximately 4,000 abstracts were initially reviewed, and in the end, a total of 40 studies with 49 independent samples including 6,998 participants were included in the final analysis. For the meta-analysis of each of the included correlates, different subsets of 49 independent samples were used.

#### 4.2 Measures of L2 speaking and its correlates included in the study

For L2 speaking and each of its nine correlates, in order for a measure to be considered eligible for the study, certain requirements had to be met. Table 1 summarizes the inclusion criteria and examples of included measures for L2 speaking and its correlates.

**Table 1.** Summary of included measures of L2 speaking and its correlates

Target constructs (Correlates of L2 Speaking)	Inclusion criteria for measures	Examples of included measures
L2 speaking	Must elicit a discourse-level oral response in L2	Planned presentation, unplanned Oral Proficiency Interview, monologue (e.g., picture description), interactive task (e.g., interview)
L2 vocabulary knowledge	Must assess at least one aspect of L2 vocabulary (e.g., size, depth, lexical organization, lexical access efficiency)	Vocabulary size test, derivation test, antonym test, collocation test
L2 grammar knowledge	Must assess explicit and implicit morpho-syntactic and syntactic knowledge	TOEFL PBT Structure and Expression subtest, timed grammaticality judgement test
Working memory	Must assess processing and/or capacity components or working memory	Reading span test, backward digit span test, phonological memory test
Metacognition	Must assess any aspect of metacognition (e.g., reflective thinking, strategies, self-appraisal of oral proficiency, metacognitive awareness)	Metacognitive awareness questionnaire
Language aptitude	Must assess any aspect of aptitude related to language learning (e.g., constructs measured by MLAT, pitch discrimination and tonal memory)	Modern Language Aptitude Test
L2 reading comprehension	Must assess passage-level L2 reading comprehension	TOEFL PBT Reading Comprehension subtest
L2 listening comprehension	Must assess passage-level L2 listening comprehension	TOEFL PBT Listening Comprehension subtest
L2 writing	Must assess discourse-level L2 writing	TOEFL iBT prototype measure
Anxiety	Must assess trait, state, and foreign language anxiety	State-Trait Anxiety Inventory, Foreign Language Anxiety of Reactions

### 4.3 Analytical procedures

In case multiple studies reported duplicate data or overlapping data, only one of them was included in the analysis. In addition, if the study reported multiple sets of data collected at different time points, in most cases, the earliest set was used. However, there was an exception to this rule; if the majority of study samples under

analysis featured a similar characteristic (e.g., study participants were young adults) and a single study reported longitudinal data collected over the time span of grade 9 through 12, for example, the grade 12 data were used to make the data in the study pool more comparable. When studies reported correlations between L2 speaking and subconstructs of a correlate (e.g., a correlation between L2 speaking and vocabulary size, a correlation between L2 speaking and vocabulary depth), we used an average value of the reported correlations. Approximately 24% of all studies were coded for various study features by more than one coder (i.e., the authors of the study). Inter-coder reliabilities were very high (99%) and all inconsistencies in coding were resolved through discussion.

#### 4.4 Meta-analytic procedures

Study names, sample sizes, and when reliability information was provided in the primary study, correlations which were corrected for attenuation using Spearman's (1904) formula were entered in Comprehensive Meta-Analysis Version 2 (Borenstein et al., 2005). When the primary study did not report reliability information, we used the average of the reliability indices of the study pool under analysis. When only a minority of studies in the pool of analysis reported reliability indices, correlations were not corrected. To be eligible for analysis, a minimum of five independent samples were required.

To address the potential file drawer problem, we computed the classic fail-safe  $N$  as well as Orwin's fail-safe  $N$ . We also examined funnel plots and used the trim and fill method when appropriate.

Weighted average correlations between L2 speaking and nine correlates were then computed along with their 95% confidence interval (CI). A correlation with its CI surpassing zero indicates that it is statistically significantly different from zero.

If there were both a clear theoretical motivation and a statistical indication, namely, a statistically significant  $Q$  test of homogeneity (Hedges & Olkin, 1985) and a large  $I^2$  statistic, moderator analyses were considered. However, in the end we agreed that there was no appropriate justification for moderator analyses and only carried out main analyses. For these, we used a random-effects model due to the assumption that there would be heterogeneity above and beyond the sampling error.

For the interpretation of the size of correlations, we referred to Cohen (1988): a correlation of .10 was considered small, 0.3 was considered moderate, and a correlation of .05 or larger was considered large.

## 5. Results

Table 2 summarizes the overall effect sizes, 95% CIs, Q test results and  $I^2$  values, the results of retrieval bias tests (fail-safe  $N$ , Orwin's fail-safe  $N$  [1983]), and adjustments (trim and fill) for asymmetries detected in the funnel plot for the four high-evidence variables. As for the retrieval bias tests, we followed Rosenthal's (1979) guideline of  $5k + 10$  when  $k$  equals the number of effect sizes under analysis. A value higher than  $5k + 10$  was considered to indicate inconsequential concern for retrieval bias. Below, we report on the analysis results for each correlate.

**Table 2.** Mean correlations between L2 speaking and nine correlates

Variable name	$k^a$	$r$ [95% CI]	Significant test of difference (Q test)	$I^2$	Classical fail-safe $N$	Orwin's fail-safe $N$	Adjusted effect estimate after trim and fill	Trimmed studies
Vocabulary knowledge <sup>b</sup>	15	.56 [.42, .68]	114.92**	87.82	1523	207	.66 [.62, .69]	3
Grammar knowledge <sup>b</sup>	15	.62 [.43, .76]	641.36**	97.82	6584	310	.62 [.43, .76]	0
Working memory	6	.10 [−.04, .24]	5.15	2.82	0	6	.06 [−.11, .22]	2
Language aptitude	10	.48 [.31, .62]	126.25*	92.87	722	65	.48 [.31, .62]	0
L2 reading comprehension <sup>b</sup>	8	.82 [.15, .98]	7248.97**	99.90	2538	215	.88 [.51, .98]	2
L2 listening comprehension	8	.53 [.33, .68]	341.32**	97.95	2592	104	.58 [.41, .72]	1
L2 writing	7	.61 [.40, .75]	310.16**	98.07	2224	91	.66 [.50, .78]	2
Anxiety <sup>b</sup>	9	−.43 [−.69, −.09]	174.17*	95.41	169	39	−.52 [−.57, −.47]	3
Metacognition <sup>b</sup>	5	.62 [.04, .89]	182.28**	97.81	228	64	0.62 [.04, .89]	0

*Note*

a. The number of effect sizes.

b. Correlations were corrected for attenuation.

5.1 L2 speaking and L2 vocabulary knowledge

Fifteen independent samples from eleven studies involving 1,123 study participants were included in the analysis (mean sample size = 66.06, *SD* = 28.78, range = 26–38). Figure 1 summarizes the study results for this correlate. Participants' age ranged from high schoolers to adults. Of the fifteen samples, eleven were Japanese L1 speakers, one sample each consisted of English L1 and Persian L1 speakers, and the rest had mixed L1s (e.g., Arabic, Chinese, French, German). As for the target language, English was the most popular (14 samples), with Spanish (1 sample) in the minority. The overall mean correlation was medium,  $r = .56$ , 95% CI [.42, .68] (Cohen, 1988) and statistically significant ( $p < .01$ ). The variability across studies was significant and large,  $Q(14) = 114.92$ ,  $p < .01$ ,  $I^2 = 87.82$ .

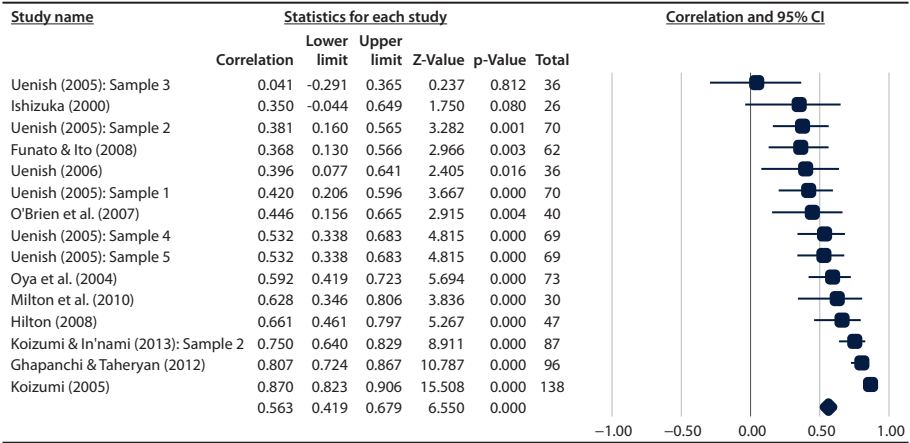
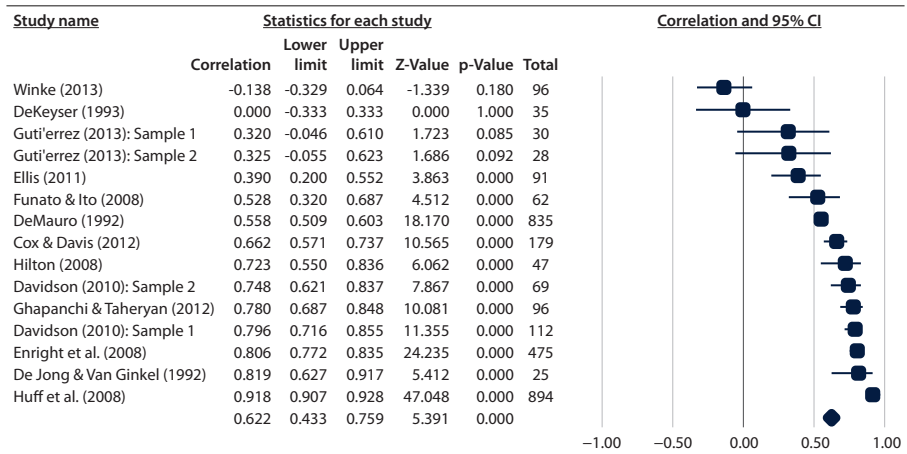


Figure 1. Overall average correlation (displayed by a diamond) and correlation with confidence interval for each study correlating L2 speaking and L2 vocabulary knowledge

5.2 L2 speaking and L2 grammar knowledge

Fifteen independent samples from thirteen studies involving 3,074 study participants were included in this analysis (mean sample size = 204.93, *SD* = 290.05, range = 25–894). See Figure 2 for a summary of the study results for this correlate. Participants' age ranged from secondary school to adults (college level). Of the fifteen samples, five were English L1 speakers, two were Dutch L1 speakers, one consisted of Persian L1 speakers, and the rest had mixed L1s (Albanian, Arabic, Bambara, Bedoussian, Chinese, French, Fulfulde, German, Hindi, Indonesian, Italian, Japanese, Korean, Latvian, Nepali, Polish, Portuguese, Russian, Spanish, Thai, Tunisian, Ukrainian, Vietnamese). As for the target languages involved, English

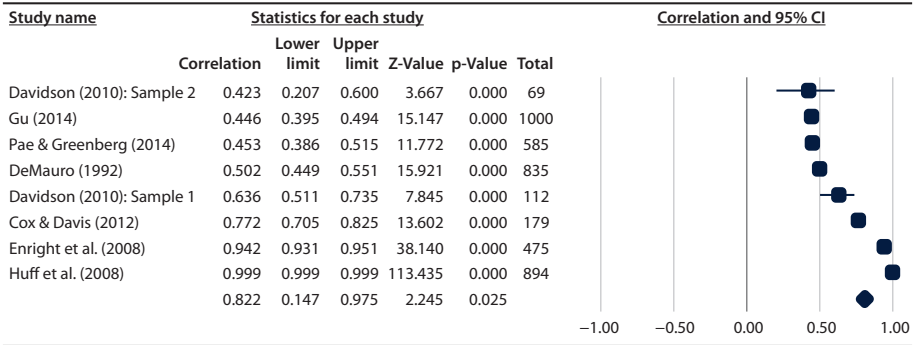


**Figure 2.** Overall average correlation (displayed by a diamond) and correlation with confidence interval for each study correlating L2 speaking and L2 grammar knowledge

was again the most popular (7 samples). Spanish (2 samples), Russian (2 samples), French (2 samples), and Chinese (1 sample) were other L2s. One remaining sample involved mixed L2s (English, Italian, French). The overall mean correlation was large,  $r = .62$ , 95% CI [.43, .76] (Cohen, 1988) and statistically significant ( $p < .01$ ). The variability across studies was significant and large,  $Q(14) = 641.362$ ,  $p < .01$ ,  $I^2 = 97.817$ .

### 5.3 L2 speaking and L2 reading comprehension

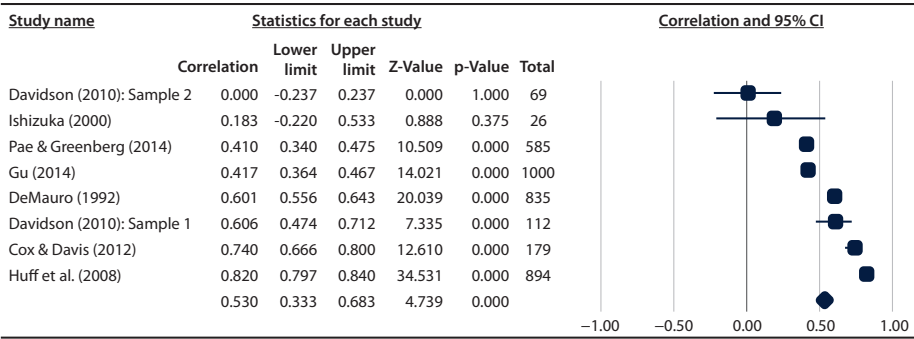
Eight independent samples from seven studies involving 4,149 study participants were included in this analysis (mean sample size = 518.63,  $SD = 370.69$ , range = 69–1000). See Figure 4 for a summary of the study results of this correlate. Participants' ages were in the adult range (DeMauro [1992]; Gu [2014]; Huff et al. [2008] did not report their participants' age but we assume they were adults given that they were TOEFL test takers around the world). Two samples were exclusively L1 English speakers while the rest of the samples featured diverse combinations of mixed L1s (e.g., Albanian, Arabic, Bambara, Bedoussian, Chinese, French, Fulfulde, German, Indonesian, Italian, German, Hindi, Japanese, Korean, Latvian, Nepali, Portuguese, Russian, Spanish, Thai, Tunisian, Ukrainian, Vietnamese). Once again, English was the most commonly featured L2 among the included samples (6 samples) with Russian in the minority (2 samples). The overall, corrected and weighted mean correlation was significant and large,  $r = .82$ , 95% CI [.15, .98] (Cohen, 1988) and statistically significant ( $p < .01$ ). The variability across samples was large and significant,  $Q(7) = 7248.97^{**}$ ,  $p < .01$ ,  $I^2 = 99.90$ .



**Figure 3.** Overall average correlation (displayed by a diamond) and correlation with confidence interval for each study correlating L2 speaking and L2 reading comprehension

5.4 L2 speaking and L2 listening comprehension

Eight independent samples from seven studies involving 3,700 study participants were included in the analysis (mean sample size = 462.50, *SD* = 410.13, range = 26–1000). See Figure 5 for a summary of the study results of this correlate. Participants’ ages were in the adult range (DeMauro [1992]; Gu [2014]; Huff et al. [2008] did not report their participants’ age but we assume they were adults given that they were TOEFL test takers around the world). Two samples and one sample consisted of exclusively English and Japanese L1 speakers, respectively, while the rest of the samples featured combinations of mixed L1s (e.g., Albanian, Arabic, Bambara, Bedoussian, Chinese, French, Fulfulde, German, Hindi, Indonesian, Italian, Japanese, Korean, Latvian, Nepali, Portuguese, Russian, Spanish, Thai, Tunisian, Ukrainian, Vietnamese). English was the target language among six samples with



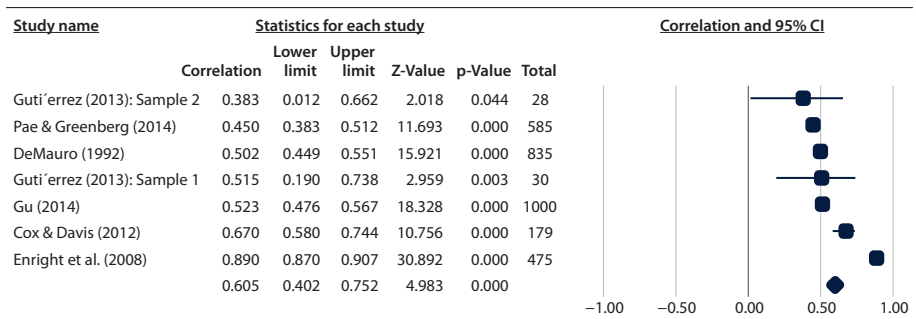
**Figure 4.** Overall average correlation (displayed by a diamond) and correlation with confidence interval for each study correlating L2 speaking and L2 listening comprehension



Russian in the minority (2 samples). The overall, corrected and weighted mean correlation was large,  $r = .53$ , 95% CI [.33, .68] (Cohen, 1988) and statistically significant ( $p < .01$ ). The variability across samples was large and significant,  $Q(7) = 341.32$ ,  $p < .01$ ,  $I^2 = 97.95$ .

## 5.5 L2 speaking and L2 writing

Seven independent samples from six studies involving 3,132 study participants were included in the analysis (mean sample size = 447.43,  $SD = 386.75$ , range = 28–1000). See Figure 6 for a summary of study results of this correlate. Participants' ages were in the adult range (DeMauro [1992] and Gu [2014] did not report their participants' age but we assume they were adults given that they were TOEFL test takers around the world). Two samples were exclusively L1 English speakers while the rest of the samples featured diverse combinations of mixed L1s (e.g., Albanian, Bambara, Bedoussian, Chinese, French, Fulfulde, German, Italian, Japanese, Korean, Nepali, Portuguese, Russian, Spanish, Tunisian, Ukrainian, Vietnamese). English was the only target language in all seven samples. The overall, corrected and weighted mean correlation was large,  $r = .61$ , 95% CI [.40, .75] (Cohen, 1988) and statistically significant ( $p < .01$ ). The variability across samples was large and significant,  $Q(6) = 310.16$ ,  $p < .01$ ,  $I^2 = 98.07$ .



**Figure 5.** Overall average correlation (displayed by a diamond) and correlation with confidence interval for each study correlating L2 speaking and L2 writing

5.6 L2 speaking and working memory

Six independent samples from five studies involving 227 study participants were included in this analysis (mean sample size = 37.83, *SD* = 30.06, range 12–96). See Figure 3 for a summary of study results for this correlate. All participants were adults with diverse L1 backgrounds: Catalan (2 samples), English (1 sample), Portuguese (1 sample), Chinese (1 sample), mixed L1s of English and German (1 sample). English was once again the most frequently represented L2 (4 samples) with Chinese (1 sample) and some mixed L2s represented in the study pool (1 mixed sample of English and German). Most of the studies did not report reliability statistics of their measures. For this reason, the correlations reported here were not corrected for attenuation. Therefore, it must be noted that the reported correlation may seem smaller compared to the correlations that were corrected for attenuation as in the case of L2 vocabulary knowledge and L2 grammar knowledge. The overall mean correlation was small,  $r = .10$ , 95% CI  $[-.04, .24]$  (Cohen, 1988) and statistically not significant ( $p = .15$ ) The variability across studies was also not significant and small,  $Q(5) = 5.15$ ,  $p = .40$ ,  $I^2 = 2.82$ .

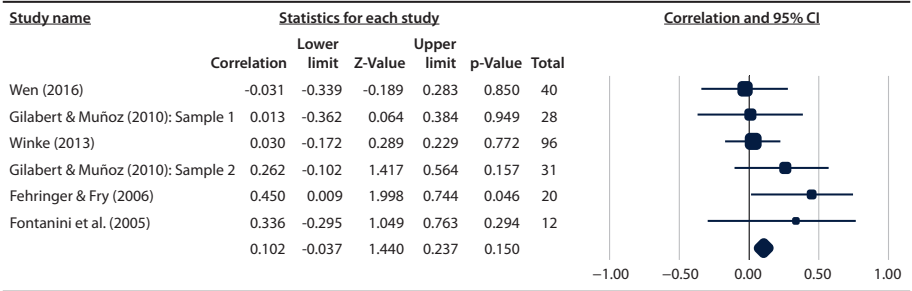
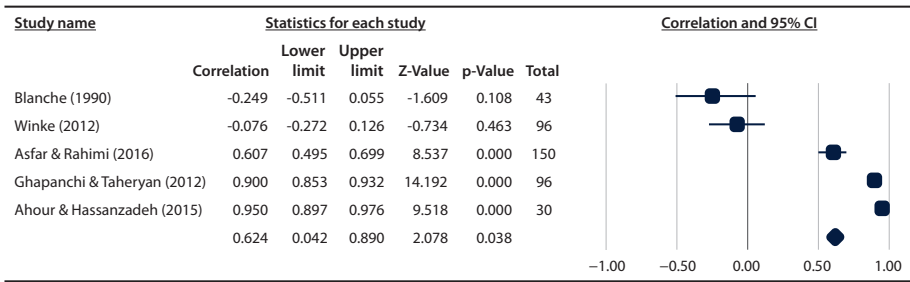


Figure 6. Overall average correlation (displayed by a diamond) and correlation with confidence interval for each study correlating L2 speaking and working memory

5.7 L2 speaking and metacognition

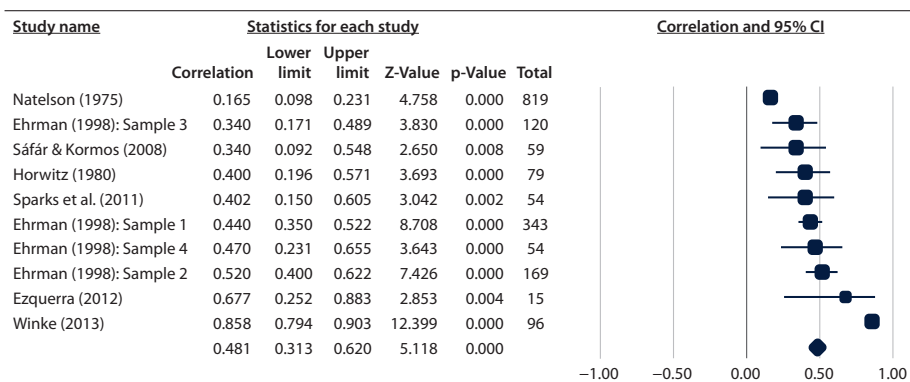
Five independent samples from five studies involving 415 study participants were included in this analysis (mean sample size = 82.50, *SD* = 66.71, range = 30–169). See Figure 8 for a summary of study results of this correlate. Participants’ L1s included English (2 samples), Farsi (2 samples) and one sample featured bilingual speakers in Gileki and Farsi. As for the L2s, three samples featured English, and one sample each featured French and Chinese. The overall, corrected and weighted mean correlation was large,  $r = .62$ , 95% CI  $[.04, .89]$  (Cohen, 1988) and statistically significant ( $p < .01$ ). The variability across samples was large and significant,  $Q(4) = 182.28$ ,  $p < .01$ ,  $I^2 = 97.81$ .



**Figure 7.** Overall average correlation (displayed by a diamond) and correlation with confidence interval for each study correlating L2 speaking and metacognition

## 5.8 L2 speaking and language aptitude

Ten independent samples from seven studies involving 1,808 study participants were included in this analysis (mean sample size = 180.80,  $SD = 242.55$ , range = 15–819). See Figure 7 for a summary of study results of this correlate. English was the most common L1 among the study participants with a few samples with L1s such as Hungarian (Sáfár & Kormos, 2008) and unidentified, non-English mother tongues (Horwitz, 1980). As for the L2s, one sample each involved Chinese, French, and Spanish exclusively, and the rest of the samples featured mixed L2s (e.g., Arabic, Chinese, Japanese, Korean, Swahili, Indonesian, unidentified Northern European languages, unidentified Western European languages, Russian, Thai). Due to lack of reliability reporting in the studies included for the analysis, correlations were not corrected for attenuation although they were weighted for sample size. The overall mean correlation was moderate to large,  $r = .48$ , 95% CI [.31, .62] (Cohen, 1988) and statistically significant ( $p < .01$ ). The variability across samples was large and significant,  $Q(9) = 126.25$ ,  $p < .01$ ,  $I^2 = 92.87$ .



**Figure 8.** Overall average correlation (displayed by a diamond) and correlation with confidence interval for each study correlating L2 speaking and language aptitude

5.9 L2 speaking and anxiety

Nine independent samples from eight studies involving 671 study participants were included in this analysis (mean sample size = 74.56, *SD* = 54.59, range = 30–169). See Figure 9 for a summary of study results of this correlate. Participants’ L1s included English (4 samples), Chinese (2 samples), Farsi (1 sample), Dutch (1 sample), and Japanese (1 sample). As for the L2, four samples featured English, three featured French, and two remaining samples involved mixed L2s (Spanish, French, German). We would like to note that unlike the other correlates, the direction of correlations between L2 speaking and anxiety is in the negative (higher levels of anxiety is associated with poorer oral performance). The overall, corrected and weighted mean correlation was in the moderate to large range,  $r = -.43$ , 95% CI  $[-.69, -.09]$  (Cohen, 1988) and statistically significant ( $p < .05$ ). The variability across samples was large and significant,  $Q(8) = 174.17^*$ ,  $p < .05$ ,  $I^2 = 95.41$ .

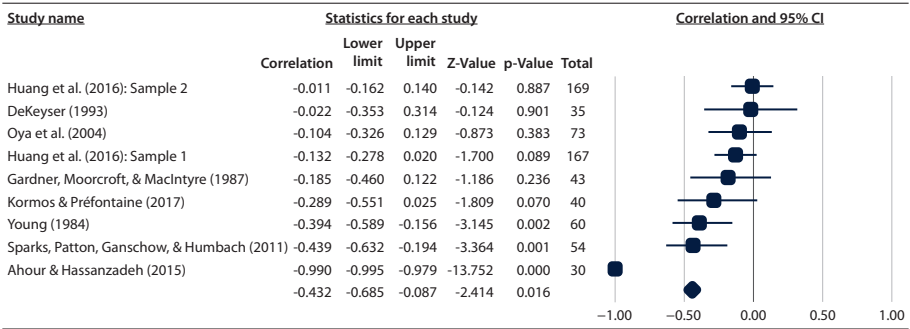


Figure 9. Overall average correlation (displayed by a diamond) and correlation with confidence interval for each study correlating L2 speaking and anxiety

6. Discussion

The purpose of this study was to examine the relationship between discourse-level L2 speaking and nine correlates representing four domains: L2 knowledge (L2 vocabulary knowledge, L2 grammar knowledge), language-general, cognitive resources (working memory, metacognition), L2 proficiency (L2 reading comprehension, L2 listening comprehension, L2 writing), and two individual difference variables (language aptitude and anxiety). To this end, the study meta-analyzed correlations between L2 speaking and each of its nine correlates as reported in select primary studies. The results indicated that L2 linguistic knowledge-and L2 proficiency-related variables consistently showed a moderate to strong relationship

with L2 speaking. Language-general variables such as cognitive resource variables and individual difference variables, on the other hand, showed more varied strengths of relationship with L2 speaking. In this section, we discuss in detail the findings related to the nine correlates belonging to the aforementioned four domains in turn.

## 6.1 L2 speaking and L2 knowledge variables: L2 vocabulary and L2 grammar

Two key constructs representing L2 knowledge, namely, L2 vocabulary knowledge and L2 grammar knowledge, were included in this study. As noted earlier, the overall correlations for these variables were in the medium to large range (Cohen, 1988):  $r = .56$ , 95% CI [.42, .68] for L2 vocabulary knowledge and  $r = .62$ , 95% CI [.43, .76] for L2 grammar knowledge. In other words, 32% and 39% of the variance in L2 speaking was accounted for by L2 vocabulary and L2 grammar knowledge, respectively. Although the overall correlation for L2 grammar knowledge was found to be slightly higher, the 95% CIs of the two correlates largely overlapped, suggesting that they were equally important correlates of L2 speaking.

While many primary studies synthesized here involved different types of L2 vocabulary knowledge (e.g., size, depth, lexical processing efficiency, productive vocabulary, receptive vocabulary) and indicated increasingly diverse target constructs, vocabulary size and receptive vocabulary knowledge were clearly the most popular choices. This likely reflects the trend among the established vocabulary knowledge tests (e.g., Vocabulary Levels Test, vocabulary test items from TOEFL PBT, DIALANG vocabulary subtest) which also focus on vocabulary size and receptive knowledge. We would like to note, however, that the dominant use of receptive vocabulary tests among the primary studies summarized here may have had suppressing effects on the strength of the overall correlation due to the following reason; productive vocabulary knowledge which, being the type of knowledge needed to efficiently search for words to encode preverbal message, is likely to be more strongly associated with L2 speaking performance than receptive knowledge (Koizumi, 2005). Due to the small number of studies which involved productive vocabulary knowledge, however, we could not examine the potential moderating effects of vocabulary knowledge types on the relationship between vocabulary knowledge and L2 speaking. Including measures of receptive and productive vocabulary knowledge simultaneously in future research would help us better understand the nature of their relationship with L2 speaking.

As expected, L2 grammar knowledge was found to be an important correlate of L2 speaking. Similar to the assessment practice of L2 vocabulary knowledge within L2 speaking research, most researchers of the primary studies included

in the present meta-analysis adopted a traditional measure of grammar knowledge such as sentence completion, error detection, and error correction both in the form of researcher-made tests or commercial tests (e.g., the Structure and Written Expression subtest of TOEFL PBT, DIALANG grammar subtest). Except for Ellis (2011), which included both explicit (untimed grammaticality judgment test, metalinguistic knowledge test) and implicit grammar knowledge test (timed grammaticality judgment test), all studies only included measures of explicit grammar knowledge. Due to the insufficient number of effect sizes representing implicit grammar knowledge, we could not examine whether there was a significant moderating effect of the type of grammar knowledge measured (i.e., explicit vs. implicit grammar knowledge) on the relationship between grammar knowledge and L2 speaking. A review of individual correlations yielded by some studies does provide, however, a complex picture which is not readily explicable, and warrants further research. For example, Ellis's (2011) study, which included both explicit and implicit measures of L2 grammar knowledge, found that the correlations between L2 speaking and three different types of grammar tests did not substantially vary; the correlations reported by the researcher for untimed GJT and metalinguistic knowledge test, both of which were designed to measure explicit grammar knowledge were .36 and .27, respectively. After we corrected them for attenuation for the purpose of this study, the correlations were .43 and .31. Correlations for timed GJT which was designed to measure implicit grammar knowledge were in a comparable range: .36 as reported in the study, and .43 after we corrected it for attenuation. As L2 speaking is a performance that typically takes place under time pressure and often in an unplanned manner, it is likely that implicit, rather than explicit, grammar knowledge plays a more important role. Why this assumption was not reflected in the correlations reported by Ellis (2011) is unclear and clearly calls for further research which includes different dimensions of grammar knowledge.

## 6.2 L2 speaking and language-general cognitive resources: Working memory and metacognition

Working memory and metacognition were included as variables representing language-general constructs supporting L2 speaking. In contrast with metacognition which showed a robust and positive association with L2 speaking ( $r = .62$ , 95% CI [.04, .89]), the overall correlation was nonsignificant and small for working memory ( $r = .10$ , 95% CI [-.04, .24]). While the overall correlation for working memory was computed without correction for attenuation due to lack of reliability reporting among studies, and therefore, may have been downward-biased, the result is not a far cry from the current status of the research domain. As Linck et al. (2014) noted, the renewed interest in working memory in cognitive psychology and bilingualism

following Daneman and Carpenter (1983) has extended to L2 research, especially in the realm of individual difference research (Miyake & Freedman, 2021). The findings on the relationship between working memory and L2 performance, however, have been inconsistent as can be seen in Linck et al. (2014) where the 95% CI of 53 out of 79 individual correlations included zero. It is also worth noting that two of those individual correlations were in the negative direction.

Another noteworthy observation is that while a significant amount of L2 studies have investigated the relationship between working memory and various criterion measures ranging from narrow constructs (e.g., phonological discrimination, productive vocabulary knowledge) to comprehensive constructs (e.g., passage-level reading comprehension, overall speaking proficiency) (Linck et al., 2014), there have been very few studies that belong to the latter category, especially concerning L2 speaking. Considering how speaking proficiency is conceptualized outside a laboratory setting, this research practice remains wanting for inclusion of comprehensively assessed, discourse-level L2 speaking tasks which closely resemble speaking tasks of the target language use domain.

Lastly, the potential task effects of different working memory measures are nontrivial. Linck et al. (2014), in their meta-analysis of working memory and its relationship with a broad range of L2 comprehension and production tasks, concluded that complex span tasks tapping executive control and storage were more highly correlated with L2 tasks than simple span tasks tapping storage capacity only. In the present study, both types of working memory measures were represented in a balanced manner as can be seen in Table 3. While we could not run a moderator analysis on test task effects due to a low number of comparable effect sizes, a review of Table 3 reveals that it is often the case that within the same study, complex tasks tend to correlate higher with L2 speaking than simple tasks.

Unlike working memory, the overall correlation between metacognition and L2 speaking was robust and positive ( $r = .62$ ). However, as can be seen in Table 3, a more complex picture emerged at a closer look at the individual correlations summarized. Out of the five corrected correlations included in the present study, two were in the negative direction:  $r = -.25$  in Blanche (1990) and  $r = -.08$  in Winke (2013). In Blanche (1990), metacognition in the form of participants' self-assessment of grammatical competence was correlated with their L2 speaking performance. The results showed that the participants with the highest proficiency were highly accurate in their assessments, the average-level participants tended to underestimate their ability, and the weakest participants overestimated their ability, potentially leading to an overall correlation in the negative direction. Winke's (2013) interpretation of the negative correlation between language learning strategy use and L2 speaking is also in line with this observation in that highly skilled learners tend to successfully use fewer, task-specific strategies, while less skilled



**Table 3.** Working memory measures, take type, and correlations with L2 speaking

Study	WM measure	Task Type	Correlation with L2 Speaking
Fehringer & Fry (2006)	Backward Digit Span	Simple or complex	0.37
	Aural Working Memory Span	Complex	0.44
	Adult Memory and Information Processing Battery (AMIPB)	Complex	0.54
Fontanini, Weissheimer, Bergsleithner, Perucci, & D'Ely (2005)	Operation Word Span	Complex	0.53
	Reading Span	Simple	0.43
	Syntax Span	Complex	0.27
	Speech Span	Complex	0.64
Gilabert & Muñoz (2010): Sample 1	Reading Span	Simple	0.01
Gilabert & Muñoz (2010): Sample 2	Reading Span	Simple	0.26
Wen (2016)	Nonword Span Test	Simple	0.03
	Speaking Span Test in L1	Complex	−0.07
	Speaking Span Test in L2	Complex	−0.05
Winke (2013)	Phonological Working Memory Span Test	Complex	0.03

participants use a wide range of random strategies. The overall negative correlations seem to have resulted from such a curvilinear pattern of relationships between strategy use and L2 speaking proficiency. If this is indeed the case, the investigation into metacognition in L2 speaking would benefit from a qualitative and longitudinal observation which tracks the relationship between metacognition and L2 speaking proficiency as L2 learners progress.

**Table 4.** Metacognition measures and correlations with L2 speaking

Study	Metacognition measure	Correlation (corrected for attenuation)
Blanche (1990)	Self-appraisal of oral proficiency	−0.25
Winke (2012)	Strategy Inventory for Language Learning	−0.08
Asfar & Rahimi (2016)	The reflective thinking skills questionnaire	0.61
Ghapanchi & Taheryan (2012)	Metacognitive Awareness Inventory in Listening and Speaking Strategies	0.9
Ahour & Hassanzadeh (2015)	Strategy Inventory for Language Learning	0.95

### 6.3 The relationship between L2 speaking and three proficiency variables: L2 reading comprehension, L2 listening comprehension, and L2 writing

All three companion proficiency variables included in the study showed a strong correlation with L2 speaking:  $r = .82$ , 95% CI [.15, .98] for L2 reading comprehension,  $r = .53$ , 95% CI [.33–.68] for listening comprehension, and  $r = .61$ , 95% CI [.40, .75] for L2 writing. We would like to note that these results should be interpreted with the following detail in mind; of the three correlations, only that for L2 reading comprehension was corrected for attenuation due to the lack of reliability reporting in the primary studies involving L2 listening comprehension and L2 writing. As such, it is possible that had the correlations for L2 listening comprehension and L2 writing been corrected, they would have been rendered larger than the current values. In any case, the 95% confidence intervals associated with the three correlations largely overlapped, indicating that L2 reading comprehension, L2 listening comprehension, and L2 writing were equally important correlates of L2 speaking.

The consistently strong correlations between L2 speaking and the other three proficiency variables render support for the multicomponential language ability structure (Bachman, 1998; In'nami & Koizumi, 2012), which, applied to L2 speaking, assumes that it is composed of “a number of distinct but related component abilities” (Bachman, 1998, p. 177).

### 6.4 L2 speaking and language aptitude

In this study, language aptitude and L2 speaking were found to have a positive and a moderate to large correlation ( $r = .48$ , 95% CI [.31, .62]). Except for Ezquerra (2012), all included primary studies used MLAT to measure language aptitude. As comprehensively reviewed in Wen et al. (2017), the language aptitude research has seen much advancement in the recent years both in theory (e.g., Sparks & Ganschow, 2001) and testing (e.g., Cognitive Ability for Novelty in Acquisition of Language-Foreign [CANAL-F], Hi-LAB, LLAMA). Despite the availability of more updated measurement options, the MLAT still seems to be the most preferred choice, a trend which also resonated with another, earlier meta-analysis by Li (2015) which examined the relationship between language aptitude and L2 grammar acquisition. As for the size of overall correlation, the present study and Li (2015) show that language aptitude is more strongly correlated with a comprehensive proficiency (or skill), namely, discourse-level L2 speaking performance ( $r = .48$ ) compared to L2 grammar ( $r = .31$ ), a component of L2 proficiency.

## 6.5 L2 speaking and anxiety

In line with Teimouri et al. (2019) which found a negative overall correlation between anxiety and L2 achievement, anxiety was shown to have a moderate to strong, inverse relationship with L2 speaking in the present study:  $r = -.43$  [95% CI:  $-.69, -.09$ ]. As noted earlier, oral communication in L2 is a notoriously anxiety-provoking task (Kormos & Préfontaine, 2017). Given that the L2 speaking tasks included in the primary studies synthesized here are impromptu in nature and allow no rehearsal opportunities (e.g., oral proficiency interview, picture description, TOEFL iBT independent speaking task, TOEFL iBT integrated speaking task), the present study finding is hardly surprising. Among the primary studies reviewed here, anxiety was most often assessed using a self-report type measure (e.g., Foreign Language Classroom Anxiety Scale [Horwitz et al., 1986], affective strategy subsection of Strategy Inventory for Language Learning (SILL) version 7.0, [Oxford, 1989], State-Trait Anxiety Inventory, [Spielberger et al., 1983]) where the respondent reports on a Likert scale how she feels about carrying out a certain oral communication task (state anxiety) or about her general level of anxiety in life (trait anxiety). Also noted was a self-report measure where the respondent is asked about her affective coping strategies when confronted with an anxiety-provoking L2 task. Considering that anxiety is inherently a subjective emotion, the predominant use of self-report measures to assess language-related anxiety seems judicious.

## 7. Conclusion

The present study meta-analyzed primary study findings on the relationship between important external correlates of L2 speaking and discourse-level L2 speaking performance. Our results showed that two key variables representing L2 knowledge, namely, L2 vocabulary and grammar knowledge, had a moderate to large correlation with discourse-level L2 speaking performance. On the other hand, an interesting and rather unexpected finding arose regarding two language-general variables, i.e., working memory and metacognition. While working memory only weakly correlated with L2 speaking, metacognition showed a strong correlation with L2 speaking, although they are both language-general variables that support speaking performances across languages. As discussed earlier, we conjecture that these differences may be related to the current state of the research domain of working memory and metacognition. In contrast with working memory, metacognition has not been popularly investigated in relation with L2 speaking. As such, the range of subconstructs and their operationalization is much more limited in the case of metacognition, and this may have contributed to the observed differences between

working memory and metacognition in their relationship with L2 speaking. Future research examining our conjectures would advance our understanding of how language-general variables contribute to L2 speaking performances. Another cognitive variable, language aptitude, also showed a moderate to large correlation with L2 speaking. As most of the primary studies used MLAT to measure language aptitude, we await more test data from studies using newly developed language aptitude to refine our findings. Next, the three companion proficiency variables, namely L2 reading comprehension, L2 listening comprehension, and L2 writing all correlated highly with L2 speaking, probably because they are largely supported by the same L2 knowledge base although we must not disregard the role of other variables at play. Lastly, anxiety, which was the only affective variable included in this study showed a small to medium size, negative correlation with L2 speaking performance. Considering that L2 speaking is an especially anxiety-provoking event for many L2 learners and that anxiety can and does affect cognitive functions (e.g., taxing working memory), it is imperative we continue to investigate how variables of different nature on the surface may actually interact as they support L2 speaking.

## References

*Note.* Studies that were included in the meta-analysis are marked with an asterisk (\*).

- \*Ahour, T., & Hassanzadeh, Z. (2015). An investigation of the relation between self-esteem, indirect strategy use and Iranian intermediate EFL learners' oral language proficiency. *Theory and Practice in Language Studies*, 5(2), 442–451. <https://doi.org/10.17507/tpls.0502.28>
- \*Asfar, H. S., & Rahimi, M. (2016). Reflective thinking, emotional intelligence, and speaking ability of EFL learners: Is there a relation? *Thinking Skills and Creativity*, 19, 97–111. <https://doi.org/10.1016/j.tsc.2015.10.005>
- \*Blanche, P. (1990). Using standardized achievement and oral proficiency tests for self-assessment purposes: the DLIFLC study. *Language Testing*, 7(2), 202–229. <https://doi.org/10.1177/026553229000700205>
- Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2005). *Comprehensive meta-analysis version 2* [software]. Biostat.
- Bowles, M. A. (2010). *The think-aloud controversy in second language research*. Routledge. <https://doi.org/10.4324/9780203856338>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.
- \*Cox, T. L., & Davis, R. S. (2012). Using automatic speech recognition technology with elicited oral response testing. *CALICO Journal*, 29(4), 601–618. <https://doi.org/10.11139/cj.29.4.601-618>
- Daneman, M., & Carpenter, P. A. (1983). Individual differences in integrating information between and within sentences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 9(4), 561–584. <https://doi.org/10.1037/0278-7393.9.4.561>
- \*Davidson, D. E. (2010). Study abroad: When, how long, and with what results? New data from the Russian front. *Foreign Language Annals*, 43(1), 6–26. <https://doi.org/10.1111/j.1944-9720.2010.01057.x>

- De Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R., & Hulstijn, J. H. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition*, 34(1), 5–34.  
<https://doi.org/10.1017/S0272263111000489>
- \*De Jong, J. H. L., & van Ginkle, L. W. (1992). Dimensions in oral foreign language proficiency. In L. Verhoeven, & J. H. A. L. De Jong (Eds.), *The construct of language proficiency* (pp. 187–205). John Benjamins. <https://doi.org/10.1075/z.62.19jon>
- \*DeKeyser, R. (1993). The effects of error correction on L2 grammar knowledge and oral proficiency. *The Modern Language Journal*, 77(4), 501–514.  
<https://doi.org/10.1111/j.1540-4781.1993.tb01999.x>
- \*DeMauro, G. (1992). Examination of the relationships among TSE, TWE and TOEFL scores. *Language Testing*, 9(2), 149–161. <https://doi.org/10.1177/026553229200900203>
- Doughty, C. J. (2019). Cognitive language aptitude. *Language Learning*, 69(s1), 101–126.  
<https://doi.org/10.1111/lang.12322>
- \*Ehrman, M. (1998). The Modern Language Aptitude Test for predicting learning success and advising students. *Applied Language Learning*, 9(1–2), 31–70.
- \*Ellis, N. (2011). Implicit and explicit SLA and their interface. In C. Sanz & R. P. Leow (Eds.), *Implicit and explicit language learning: Conditions, processes, and knowledge in SLA and bilingualism* (pp. 35–47). Georgetown University Press.
- \*Enright, M. K., Bridgeman, B., Eignor, D., Lee, Y.-W., & Powers, D. E. (2008). Prototyping measures of listening, reading, speaking, and writing. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 145–186). Routledge.
- Eysenck, M. W., & Calvo, M. G. (1992). Anxiety and performance: The processing efficiency theory. *Cognition and Emotion*, 6(6), 409–434. <https://doi.org/10.1080/02699939208409696>
- \*Fehring, C., & Fry, C. (2007a). Frills, furbelows and activated memory: syntactically optional elements in the spontaneous language production of bilingual speakers. *Language Sciences*, 29(4), 497–511. <https://doi.org/10.1016/j.langsci.2006.09.001>
- Fehring, C., & Fry, C. (2007b). Hesitation phenomena in the language production of bilingual speakers: The role of working memory. *Folia Linguistica*, 41(1–2), 37–72.  
<https://doi.org/10.1515/flin.41.1-2.37>
- \*Fontanini, I., Weissheimer, J., Bergsleithner, J. M., Perucci, M., & D'Ely, R. (2005). Memória de trabalho e desempenho em tarefas de L2. [Working memory and performance on L2 tasks]. *Rev. Brasileira de Linguística Aplicada*, 5(2), 189–230.  
<https://doi.org/10.1590/S1984-63982005000200009>
- \*Funato, S., & Ito, H. (2008). An empirical study on basic requirements for Japanese EFL learners to achieve oral fluency in English. *Annual Review of English Language Education in Japan*, 19, 41–50. [https://doi.org/10.20581/arele.19.o\\_41](https://doi.org/10.20581/arele.19.o_41)
- \*Gardner, R. C., Moorcroft, R., & MacIntyre, P. D. (1987). *The role of anxiety in second language performance of language dropouts* (Research Bulletin No 657). The University of Western Ontario.
- \*Ghapanchi, Z., & Taheryan, A. (2012). Roles of linguistic knowledge, metacognitive knowledge and metacognitive strategy use in speaking and listening proficiency of Iranian EFL learners. *World Journal of Education*, 2(4), 64–75. <https://doi.org/10.5430/wje.v2n4p64>
- \*Gilbert, R., & Muñoz, C. (2010). Differences in attainment and performance in a foreign language: The role of working memory capacity. *International Journal of English Studies*, 10(1), 19–42. <https://doi.org/10.6018/ijes/2010/1/113961>

- \*Gu, L. (2014). At the interface between language testing and second language acquisition: Language ability and context of learning. *Language Testing*, 31(1), 111–133. <https://doi.org/10.1177/0265532212469177>
- \*Gutiérrez, X. (2013). Metalinguistic knowledge, metalingual knowledge, and proficiency in L2 Spanish. *Language Awareness*, 22(2), 176–191. <https://doi.org/10.1080/09658416.2012.713966>
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press.
- \*Hilton, H. (2008). The link between vocabulary knowledge and spoken L2 fluency. *Language Learning Journal*, 36(2), 153–166. <https://doi.org/10.1080/09571730802389983>
- \*Horwitz, E. K. (1980). The relationship of conceptual level to the development of communicative competence in French (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign.
- Horwitz, E. K., Horwitz, M. B., & Cope, J. (1986). Foreign language classroom anxiety. *The Modern Language Journal*, 70(2), 125–132. <https://doi.org/10.1111/j.1540-4781.1986.tb05256.x>
- \*Huang, H.-T. D., Hung, S.-T. A., & Hong, H.-T. V. et al. (2016). Test-taker characteristics and integrated speaking test performance: A path-analytic study. *Language Assessment Quarterly*, 13(4), 283–301. <https://doi.org/10.1080/15434303.2016.1236111>
- \*Huff, K., Powers, D. E., Kantor, R. N., Mollaun, P. N., Nissán, S., & Schedl, M. (2008). Prototyping a new test. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp.187–225). Routledge.
- \*Ishizuka, H. (2000). Goi chishiki no fukasa to speaking nouryoku no soukan. [Correlations between depth of vocabulary knowledge and speaking ability: Can native-like fluency be enhanced by vocabulary knowledge?]. *STEP (Society for Testing English Proficiency) Bulletin*, 12, 13–25. Retrieved on 12 January 2022 from [https://www.eiken.or.jp/center\\_for\\_research/pdf/bulletin\\_archives/vol\\_12.pdf](https://www.eiken.or.jp/center_for_research/pdf/bulletin_archives/vol_12.pdf)
- Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, 29(1), 24–49. <https://doi.org/10.1093/applin/amm017>
- \*Koizumi, R. (2005). Predicting speaking ability from vocabulary knowledge. *ILTA (Japan Language Testing Association) Journal*, 7, 1–20. [https://doi.org/10.20622/jltaj.7.o\\_1](https://doi.org/10.20622/jltaj.7.o_1)
- \*Koizumi, R., & In'nami, Y. (2013). Vocabulary knowledge and speaking proficiency among second language learners from novice to intermediate levels. *Journal of Language Teaching and Research*, 4(5), 900–913. <https://doi.org/10.4304/jltr.4.5.900-913>
- Kormos, J. (2006). *Speech production and second language acquisition*. Lawrence Erlbaum Associates.
- Kormos, J., & Denés, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, 32(2), 145–164. <https://doi.org/10.1016/j.system.2004.01.001>
- \*Kormos, J. & Préfontaine, Y. (2017). Affective factors influencing fluent performance: French learners' appraisals of second language speech tasks. *Language Teaching Research*, 21(6), 699–716. <https://doi.org/10.1177/1362168816683562>
- Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. The MIT Press.
- Li, S. (2015). The associations between language aptitude and second language grammar acquisition: A meta-analytic review of five decades of research. *Applied Linguistics*, 36(3), 385–408. <https://doi.org/10.1093/applin/amu054>
- Li, S. (2016). The construct validity of language aptitude: A meta-analysis. *Studies in Second Language Acquisition*, 38(4), 801–842. <https://doi.org/10.1017/S027226311500042X>

- Linck, J. A., Osthus, P., Koeth, J. T., & Bunting, M. F. (2014). Working memory and second language comprehension and production: A meta-analysis. *Psychonomic Bulletin and Review*, 21, 861–883. <https://doi.org/10.3758/s13423-013-0565-2>
- MacIntyre, P., & Gardner, R. C. (1994). The subtle effects of language anxiety on cognitive processing in the second language. *Language Learning*, 44(2), 284–305. <https://doi.org/10.1111/j.1467-1770.1994.tb01103.x>
- \*Milton, J., Wade, J., & Hopkins, N. (2010). Aural word recognition and oral competence in English as a foreign language. In R. Chacón-Beltrán, C. Abello-Contesse, & M. Torreblanca-López (Eds.), *Insights into non-native vocabulary teaching and learning* (pp. 83–98). Multilingual Matters. <https://doi.org/10.21832/9781847692900-007>
- Miyake, A., & Friedman, N. P. (2012). The nature and organization of individual differences in executive functions: Four general conclusions. *Current Directions in Psychological Science*, 21(1), 8–14. <https://doi.org/10.1177/0963721411429458>
- \*Natelson, E. R. (1975). The predictive validity of each of the five parts of the Modern Language Aptitude Test: The effect of previous language training on MLAT Scores; The correlation of MLAT scores and achievement in specific language groups (Unpublished doctoral dissertation). Georgetown University.
- Norris, J., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, 50(3), 417–528. <https://doi.org/10.1111/0023-8333.00136>
- \*O'Brien, I., Segalowitz, N., Freed, B., & Collentine, J. (2007). Phonological memory predicts second language oral fluency gains in adults. *Studies in Second Language Acquisition*, 29(4), 557–581. <https://doi.org/10.1017/S027226310707043X>
- Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, 24(4), 492–518. <https://doi.org/10.1093/applin/24.4.492>
- Orwin, R. (1983). A fail-safe N for effect size in meta-analysis. *Journal of Educational Statistics*, 8(2), 157–159. <https://doi.org/10.2307/1164923>
- Oxford, R. L. (1989). *Strategy inventory for language learning*. Various versions. Oxford Associates.
- \*Oya, T., Manalo, E., & Greenwood, J. (2004). The influence of language contact and vocabulary knowledge on the speaking performance of Japanese students of English. *The Open Applied Linguistics Journal*, 2(1), 11–21. <https://doi.org/10.2174/1874913500902010011>
- \*Pae, H. K., & Greenberg, D. (2014). The relationship between receptive and expressive subskills of academic L2 proficiency in nonnative speakers of English: A multigroup approach. *Reading Psychology*, 35, 221–259. <https://doi.org/10.1080/02702711.2012.684425>
- Qian, D. D., & Lin, L. H. F. (2020). The relationship between vocabulary knowledge and language proficiency. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp.66–80). Taylor and Francis.
- \*Sáfár, A., & Kormos, J. (2008). Revising problems with foreign language aptitude. *IRAL*, 46, 113–136. <https://doi.org/10.1515/IRAL.2008.005>
- Segalowitz, N., & Freed, B. F. (2004). Context, contact and cognition in oral fluency acquisition: Learning Spanish in at home and study abroad contexts. *Studies in Second Language Acquisition*, 26(2), 173–199. <https://doi.org/10.1017/S0272263104262027>
- Sparks, R., & Ganschow, L. (2001). Aptitude for learning a foreign language. *Annual Review of Applied Linguistics*, 21, 90–111. <https://doi.org/10.1017/S026719050100006X>
- \*Sparks, R. L., Patton, J., Ganschow, L., & Humbach, N. (2011). Subcomponents of second language aptitude and second language proficiency. *The Modern Language Journal*, 95(2), 253–273. <https://doi.org/10.1111/j.1540-4781.2011.01176.x>



- Spearman, C. (1904). 'General intelligence,' objectively determined and measured. *The American Journal of Psychology*, 15(2), 201–293. <https://doi.org/10.2307/1412107>
- Spielberger, C. D., Gorsuch, R. L., Lushene, R., Vagg, P. R., & Jacobs, G. A. (1983). *Manual for the state-trait anxiety inventory*. Consulting Psychologists Press.
- Steinberg, F. S., & Horwitz, E. K. (1986). The effect of induced anxiety on the denotative and interpretive content of second language speech. *TESOL Quarterly*, 20(1), 131–136. <https://doi.org/10.2307/3586395>
- Teimouri, Y., Goetze, J., & Plonsky, L. (2019). Second language anxiety and achievement: A meta-analysis. *Studies in Second Language Acquisition*, 41(2), 363–387. <https://doi.org/10.1017/S0272263118000311>
- \*Uenishi, K. (2006). Nihonjin eigo gakushusha no spikingu ryoku ni kansuru jisshoteki kenkyu: Goi hatsuon tonon kanrensei ni shoten wo atete [An empirical study regarding speaking ability of Japanese learners of English: Focus on relationships with vocabulary and pronunciation]. *Eigo to eigo kyoiku tokubetsu go: Ozasa Toshiaki taikan kinen ronshu* [English and English Language Education (Special ed.): Ozasa Toshiaki's retirement festschrift] (pp. 135–144). English Association, Faculty of School Education, Hiroshima University.
- \*Uenishi, K. (2005). An empirical study on factors predicting the speaking ability of Japanese EFL learners (Unpublished doctoral dissertation, Hiroshima University, Japan).
- Vytal, K. E., Cornwell, B. R., Letkiewicz, A. M., Arkin, N. E., & Grillon, C. (2013). The complex interaction between anxiety and cognition: Insight from spatial and verbal working memory. *Frontiers in Human Neuroscience*, 28(7), 1–11. <https://doi.org/10.3389/fnhum.2013.00093>
- \*Wen, Z. (2016). Phonological and executive working memory in L2 task-based speech planning and performance. *The Language Learning Journal*, 44(4), 418–435. <https://doi.org/10.1080/09571736.2016.1227220>
- Wen, Z., Biedron, A., & Skehan, P. (2017). Foreign language aptitude theory: Yesterday, today and tomorrow. *Language Teaching*, 50(1), 1–31. <https://doi.org/10.1017/S0261444816000276>
- \*Winke, P. (2013). An investigation into second language aptitude for advanced Chinese language learning. *The Modern Language Journal*, 97(1), 109–150. <https://doi.org/10.1111/j.1540-4781.2013.01428.x>
- \*Young, D. J. (1986). The relationship between anxiety and foreign language oral proficiency ratings. *Foreign Language Annals*, 19(5), 439–445. <https://doi.org/10.1111/j.1944-9720.1986.tb01032.x>



## Discussion, limitations, and future research

Eun Hee Jeon, Yo In'nami and Rie Koizumi

University of North Carolina at Pembroke / Chuo University /  
Seisen University

This chapter synthesizes the results of the six meta-analyses (Chapters 3, 5, 6, 8, 10, and 11) included in this volume and interprets the findings in reference to the theories and models of L2 proficiency and assessment. We first review influential theories and models of L2 proficiency starting with Spearman's (1904, 1927) and Oller's (1979) unitary view of language proficiency, leading to the most recent, Hulstijn's (2015, 2019) core-periphery model. We then discuss the findings of 4 meta-analyses which reported on the correlations between each of the four L2 skills and its external correlates. Next, we discuss the findings of 2 meta-analyses which reported on the correlations between productive L2 skills (writing and speaking) and their internal correlates. The chapter then discusses the limitations and future research directions.

### 1. Introduction

In the history of second language (L2) research, various attempts have been made to understand the notion of L2 proficiency. Researchers have proposed models of L2 proficiency for different purposes: some for language testing (e.g., Bachman & Palmer, 1996; Oller, 1979; de Jong & Verhoeven, 1992) and others for more theoretical inquiries (e.g., Canale & Swain, 1980; Hulstijn, 2015). As Davidson (1988) noted, the history of language proficiency model building can be thought of as a feedback cycle composed of (1) a confirmatory, top-down imposition of a construct structure (i.e., proficiency model) and (2) an exploratory, bottom-up validation of the proposed model(s). Given this, the purpose of this volume was twofold; Chapters 1, 2, 4, 7, and 9 provide an overview of the historical models and theoretical inquiries of L2 proficiency; in addition, Chapters 3, 5, 6, 8, 10, 11 aimed to participate in the second part of the aforementioned cycle through a collection of quantitative meta-analyses that examine the relationship between L2 skills that make up L2 proficiency and their components. More specifically, this volume aimed to achieve the following goals; first, we hoped that this volume

would help us examine the key claims of select models and frameworks of L2 proficiency (e.g., Bachman & Palmer [1996]; Canale & Swain [1980]; Carroll [1972]; Hulstijn [2015]); second, we hoped to offer information on how various constructs of L2 proficiency have been defined and measured by empirical researchers; lastly, we hoped that the volume would provide a review on which variables have been commonly investigated across different L2 skills or have been particularly popular in certain L2 skills.

In pursuing these goals, we organized the meta-analyses of correlation coefficients based on four L2 skills: reading, writing, listening, and speaking. Given the increasingly popular use of the integrated-skills approach to both language testing and instruction (e.g., the move from TOEFL PBT to TOEFL iBT), the division of L2 proficiency in the traditional, four skills may seem antiquated to some. However, we would like to note that this is still the approach that many primary researchers take in their investigation of L2 proficiency. As such, to maximize the number of studies eligible for the meta-analyses of this volume, we deemed the four-skills approach was the best option for our purpose.

Lastly, we would like to note that there was no masterplan or detailed methodological guideline imposed on the authors of the six meta-analyses included in this volume. While some general methodological principles were agreed upon and shared across the meta-analyses, we felt that certain micro-level decisions may be domain-specific, and therefore, would be best left to the domain experts (i.e., authors of the meta-analysis). Some differences in the methodological details notwithstanding, we feel that the key findings emerging from the six meta-analyses can be synthesized to advance our understanding of L2 proficiency.

Before we proceed to reviewing influential models of L2 proficiency and discussing the key findings of the individual meta-analyses included in this volume, it would be useful to remind the readers that this volume included two different types of meta-analysis of correlation coefficients. The first group of meta-analyses (Chapters 3, 6, 8, and 11) examined the strength of association between a L2 skill and what we called in this volume the “external” correlates of an L2 skill under investigation. As explained in Chapter 10, external correlates are various linguistic, cognitive, affective, or conative variables which are assessed independently of the associated L2 skill. In contrast, the second group of meta-analyses (Chapters 5 and 10) examined the relationship between the L2 skill and what we called in this volume the “internal” correlates which are inherent features of the spoken or the written discourse produced by study participants. For ease of interpretation and comparison of studies, we first synthesize the findings of the four meta-analyses which dealt with external correlates. We then discuss the findings of the two meta-analyses which involved internal correlates.

## 2. Relevant theories and models of L2 proficiency

First of all, we would like to note that it is beyond the scope of this chapter to extensively review influential theories and models of L2 proficiency. Many excellent detailed overviews documenting the development of L2 proficiency as a construct already exist (e.g., Bachman & Cohen, 1998; Chalhoub-Deville, 1997; De Jong & Verhoeven, 1992; Hulstijn, 2015; McNamara, 1996), to which we direct interested readers. As such, we provide below in Table 1 a summary of the historical development of the construct of L2 proficiency as it relates to the findings of the meta-analyses included in this volume.

**Table 1.** Chronological overview: Development of theories and models of L2 proficiency

Author(s)	Key contributions
Spearman (1904, 1927)	Proposed that a unitary construct, general intelligence ( <i>g</i> ), explains all intellectual behavior including verbal behavior.
Lado (1961)	Proposed the multi-component approach to L2 proficiency
Carroll (1961)	Emphasized the importance of processing efficiency in addition to linguistic knowledge
Hymes (1972)	Added communicative competence to the L2 proficiency model
Canale & Swain (1980)	Added strategic competence to the L2 proficiency model
Oller (1983)	Proposed the unitary competence hypothesis in assessing L2 proficiency in the 1970s.
Bachman & Palmer (1982, 1983), Carroll (1983); Vollmer & Sang (1983)	Criticized the unitary construct view of L2 proficiency using empirical evidence
Bachman (1990)	Added pragmatic competence including socio-cultural knowledge associated with the L2 to communicative competence
Hulstijn (2015, 2019)	Proposed the core-periphery model of L2 proficiency

*Note.* See Bachman (2007) for further development of conceptualizing L2 proficiency as a combination of ability and context.

Regarding how the construct of L2 proficiency developed over the past century, a few key observations can be made. Of note among those is the move from the simple, unitary view of L2 proficiency to the multi-component view. At the beginning of the 20th century, Spearman (1904, 1927) proposed a drastically parsimonious *g*-theory. This theory claimed that the unitary construct of general intelligence (i.e., *g*) accounts for all intellectual behaviors including verbal abilities. However, this view became less popular as the multi-component approach which sees L2 proficiency as a complex construct subsuming multiple variables of different nature gained traction among researchers (e.g., Bachman, 1990; Carroll, 1961; Lado, 1961).

As can be seen in Table 1 and as reflected in the wide range of variables included in the meta-analyses of this volume, components of L2 proficiency has substantially diversified; from their initial focus on linguistic knowledge, subsequent proficiency models evolved to emphasize the importance of processing efficiency of linguistic knowledge (e.g., Carroll, 1961; Hulstijn, 2015, 2019) as well as variables related to social aspects of language communication (e.g., Bachman, 1990; Canale & Swain, 1980; Hymes, 1972).

Most recently, Hulstijn (2015, 2019) introduced the core-periphery model of L2 proficiency which grouped variables of different nature into core and periphery components. Based on a series of large-scale studies conducted by the researchers at the University of Amsterdam, Hulstijn (2015, 2019) concluded that linguistic knowledge (i.e., lexical and grammatical knowledge and processing efficiency) was significantly associated with all of the four L2 skills consistently across different populations. He, thus, proposed that linguistic knowledge and the processing efficiency of such knowledge would make up the core components of L2 proficiency. In addition to core components, this model proposed that L2 proficiency also includes interaction abilities that are language-general and peripheral. Such abilities include strategic competence, metalinguistic knowledge (e.g., explicit knowledge of grammar), and the knowledge of various written and oral discourse. After reviewing models of L2 proficiency, Wallace (2022) noted that one distinguishing feature of the core-periphery model was Hulstijn's proposal: The core component is expected to influence L2 proficiency more strongly than the peripheral component. Below, we base our discussion on the core-peripheral model.

### **3. Synthesis of findings of the 4 meta-analyses examining the relationship between L2 skills and their external correlates**

The four meta-analyses included in this volume consistently provided supporting evidence for the importance of core components of L2 proficiency. As can be seen in Table 2, L2 vocabulary and grammar knowledge, the two variables which represent linguistic knowledge, were included in all of the four meta-analyses which examined the relationship between L2 reading, writing, listening, and speaking and their external correlates. This indicates that they are acknowledged as key linguistic knowledge variables among primary researchers regardless of the L2 skill of their interest. Linguistic knowledge variables other than vocabulary and grammar knowledge featured in the meta-analyses were more varied across studies, representing the more diverse and perhaps more fine-grained approach to linguistic knowledge taken by the primary studies of the research domain; in addition to vocabulary and

Table 2. Summary of key findings yielded by four meta-analyses using external correlates

Variable type	Criterion variables	Chapter 3			Chapter 6			Chapter 8			Chapter 11	
		L2 reading			L2 writing			L2 listening			L2 speaking	
		<i>k<sup>a</sup></i>	<i>r</i> [95% CI]	<i>k<sup>a</sup></i>	<i>k<sup>b</sup></i>	<i>r</i> [95% CI]	<i>k<sup>a</sup></i>	<i>k<sup>b</sup></i>	<i>r</i> [95% CI]	<i>k<sup>a</sup></i>	<i>r</i> [95% CI]	
Linguistic knowledge												
	Vocabulary	51	.724 [.636, .794]	20	53	.489 [.427, .546]	66	182	.562 [.511, .609]	15	.563 [.419, .679]	
	Grammar	26	.697 [.517, .818]	20	71	.532 [.456, .592]	25	35	.517 [.424, .599]	15	.622 [.433, .759]	
	Decoding	29	.586 [.453, .694]	6	6	.526 [.444, .600]	0	0	NA	0	NA	
	Phonological awareness	20	.611 [.515, .693]	0	0	NA	11	24	.359 [.281, .433]	0	NA	
	Orthographic knowledge	6	.590 [.341, .761]	10	17	.535 [.409, .641]	0	0	NA	0	NA	
	Morphological knowledge	14	.635 [.556, .703]	0	0	NA	5	6	.525 [.304, .693]	0	NA	
Processing efficiency												
	Oral reading fluency	8	.640 [.521, .741]	0	0	NA	0	0	NA	0	NA	
Language-general: Cognitive												
	Working memory	19	.334 [.234, .427]	9	23	.340 [.181, .482]	33	123	.297 [.225, .366]	6	.102 [−.037, .237] <sup>NS</sup>	
	Metacognition	11	.330 [.083, .539]	10	25	.189 [−.089, .439] <sup>NS</sup>	24	64	.275 [.194, .353]	5	.624 [.042, .890]	
	Language aptitude	0	NA	4	4	.281 [.028, .500]	10	36	.105 [.023, .185]	10	.481 [.313, .620]	

(continued)



Table 2. (continued)

Variable type	Criterion variables	Chapter 3			Chapter 6			Chapter 8			Chapter 11	
		L2 reading		<i>k<sup>a</sup></i>	L2 writing		<i>k<sup>a</sup></i>	L2 listening		<i>k<sup>a</sup></i>	L2 speaking	
		<i>k<sup>a</sup></i>	<i>r</i> [95% CI]		<i>k<sup>b</sup></i>	<i>r</i> [95% CI]		<i>k<sup>b</sup></i>	<i>r</i> [95% CI]		<i>k<sup>a</sup></i>	<i>r</i> [95% CI]
Language-general: Conative												
	Motivation	0	NA	29	73	.338 [.225, .442]	4	17	.106 [−.076, .282] <sup>NS</sup>	0	NA	
Language-general: Affective												
	Anxiety	0	NA	12	17	.357 [.217, .483]	14	26	.439 [.340, .530]	9	−.432 [−.685, −.087]	
	Attitude	0	NA	5	12	.384 [.242, .510]	3	20	.098 [−.242, .417] <sup>NS</sup>	0	NA	
Proficiency: L1												
	L1 reading comprehension	34	.483 [.361, .589]	0	0	NA	0	0	NA	0	NA	
	L1 writing	0	NA	20	40	.426 [.330, .514]	0	0	NA	0	NA	
Proficiency: L2												
	L2 reading comprehension	0	NA	27	41	.588 [.516, .651]	0	0	NA	8	.822 [.147, .975]	
	L2 writing	0	NA	0	0	NA	0	0	NA	7	.605 [.402, .752]	
	L2 listening comprehension	20	.812 [.638, .907]	0	0	NA	0	0	NA	8	.530 [.333, .683]	
	L2 speaking	0	NA	16	24	.605 [.533, .668]	0	0	NA	0	NA	

Note. In coding, a negative sign was converted to a positive one in L2 writing and L2 listening, but not in L2 speaking. <sup>NS</sup> Statistically not significant. *k<sup>a</sup>* Number of studies. *k<sup>b</sup>* Number of correlations. "NA" indicates one of the following situations: (1) the meta-analysis did not report the mean correlation between the correlate and the target L2 skill due to lack of includible primary studies; or (2) we were able to identify some primary studies that are admissible to the meta-analysis, but removed them from the analysis due to lack of strong theoretical motivation. L2 reading and L2 speaking meta-analyses present correlations corrected for attenuation fully and partially, respectively (see each chapter for details).

grammar, the reading meta-analysis included decoding, phonological awareness, orthographic knowledge, and morphological knowledge. The L2 writing and listening meta-analyses also used subsets of some of those variables, but, interestingly, the speaking meta-analysis did not include any additional linguistic variables other than vocabulary and grammar knowledge. An obvious explanation for the wider range of linguistic knowledge variables, especially that of decoding, orthographic knowledge, and L2 transcription included in the reading and writing meta-analyses is their distinct relevance to the written language and the influential theories and models that have shaped the empirical research in the respective research domain. For example, the Simple View of Reading (Hoover & Gough, 1990) and the Simple View of Writing (Juel, Griffith, & Gough, 1986) highlighted the importance of decoding and transcription in L2 reading and writing, respectively.

The two meta-analyses on L2 listening and speaking, both of which are skills pertaining to oral communication, exhibited an interesting contrast in the range of linguistic knowledge variables included; while the primary studies pertaining to L2 listening comprehension included phonological awareness and morphological knowledge in addition to vocabulary and grammar, L2 speaking studies did not. This trend captured in the speaking meta-analysis is interesting in comparison with the writing meta-analysis, a study on another productive skill, but which involved a wider range of externally measured linguistic knowledge variables. One possible factor may be difficulty in assessing L2 speaking, as compared to other skills (Luoma, 2004), which makes it challenging for researchers to conduct L2 speaking studies and explore relationships between L2 speaking and its external correlates. This may be evidenced by a smaller number of primary studies collected for the whole meta-analyses: 87 in reading, 103 in writing, 118 in listening, and only 40 in speaking. It would be informative, however, to expand the range of externally measured linguistic knowledge variables in this research domain to make our observation more comparable across the four skill areas.

### 3.1 Strengths of association between correlates and target L2 skills found across the four meta-analyses

As for the strength of the association between linguistic knowledge variables and the L2 skills included in the meta-analyses, the results were invariably significant and robust. In all meta-analyses, L2 vocabulary and grammar showed a significant correlation with the target L2 skill. All correlations were moderate to strong in size. It is also worth noting that in all four meta-analyses, the 95% confidence intervals of vocabulary and grammar overlapped, indicating their equally important status as a correlate of L2 skills. There may be different explanations for this finding. One

possible explanation is lexicogrammar, namely, the idea that lexical knowledge and grammar knowledge are in fact mutually dependent rather than separate from each other (Paquot et al., 2021). Although vocabulary and grammar have been traditionally separated in theory and measurement practices, recent research in language testing (e.g., Ruegg et al., 2011), corpus linguistics (Biber et al., 1999; Römer, 2009, 2017), and L2 acquisition (e.g., Goldberg, 2006) offers increasing evidence that they may indeed be a unitary construct. For example, Ruegg et al. (2011) reported that the test taker's lexical scores on the Kanda English Proficiency Test were better predicted by grammar than by the scores on lexical range, frequency, or accuracy. According to corpus linguistics research, such a phenomenon may be a natural consequence of "lexicogrammatical co-selection phenomena" (Paquot et al., 2021, p. 223) where certain lexical items are much more frequently associated with certain grammatical environments than others. The implications for L2 acquisition, particularly, the usage-based and emergentist approaches are, then, that the development of lexical and grammatical knowledge is essentially one, rather than separate. Having said that, whether the overlapping 95% confidence intervals of vocabulary and grammar knowledge as observed in the four meta-analyses are due to the presence of a unitary construct, or a chance occurrence despite the presence of two separate constructs is unclear. Continued construct validity research into lexis and grammar will help us better understand the relationship between the two key linguistic knowledge variables: lexical and grammatical knowledge.

Some interesting, additional findings were noted regarding the correlations between vocabulary and each L2 skill. Here, the correlation with L2 reading was found to be much larger than the ones with the other three skills (i.e., .724 in reading > .489 in writing, .562 in listening, and .563 in speaking, basically with no 95% CI overlaps). This finding indicates that although vocabulary is a key linguistic knowledge variable which is influential to all L2 skills, it should be highlighted as a target of instruction and assessment when it comes to reading comprehension.

Another interesting finding was noted in the correlations between phonological awareness and each L2 skill. Here, the correlation with L2 reading was larger than the one with L2 listening (e.g., .611 > .359, with no CI overlaps). This was an especially interesting finding as phonological awareness develops in the pre-literate stage to segment speech sounds, and therefore, is most often associated with oral communication rather than the written language. Although we have yet to find an explanation for the relatively stronger correlation found in L2 reading, we conjecture that the large correlation found between phonological awareness and L2 reading may be symptomatic of the importance of vocabulary knowledge in reading. Wang et al. (2004) reported in their study that, even in L2 Chinese which features a logographic writing system, learners often rely not only on semantic but also on phonetic radicals to infer meaning of new words. It is possible to conjecture then,

that writers with strong phonological awareness are likely to have stronger vocabulary knowledge, and as a result, may become strong writers.

Results on the other, less frequently included linguistic knowledge variables also provide further evidence for the importance of core components of L2 proficiency; all correlations were significant, and were moderate to large in size (Plonsky & Oswald, 2014).

### 3.2 On processing efficiency

In the core-periphery model of L2 proficiency (Hulstijn, 2015, 2019), linguistic knowledge and the processing speed of such knowledge together make up the core components. As indicated in our meta-analysis chapters on external correlates, however, primary studies very rarely include processing efficiency variables or measure the processing efficiency of linguistic knowledge. Of the four meta-analyses, the reading meta-analysis was the only one to include oral reading fluency, a processing efficiency variable. Also in the reading meta-analysis, decoding, a linguistic knowledge variable, was sometimes assessed using processing efficiency measures which combined speed and accuracy of the performance. Compared to decoding, the processing efficiency of vocabulary knowledge was much less frequently measured among primary studies, making it difficult to separately examine the role of explicit lexical knowledge and that of its processing speed. As for grammar, the situation was a little better as primary studies are increasingly adopting grammaticality judgment task (GJT) to tap into the processing efficiency of grammar knowledge. This trend made it possible to run a moderator analysis on the studies that used GJTs vs. explicit grammar knowledge tests. However, caution must be taken as we interpret the results of the moderator analysis as it remains to be known whether all GJTs were speeded enough to keep the participants from accessing explicit grammar knowledge. Nonetheless, as reported in Chapter 3, the results showed that sentence completion tests which measure explicit knowledge of morpho-syntax and syntax correlated significantly higher with L2 reading comprehension than GJTs as a whole. In the framework of the core-periphery model (Hulstijn, 2015, 2019), explicit knowledge of grammar is considered as a peripheral component. The findings of the meta-analyses suggest then, that even as a peripheral component, explicit knowledge of vocabulary and grammar still plays an important role in L2 performance across L2 skills.

The writing meta-analysis also showed that explicit linguistic knowledge is a lot more frequently measured than processing efficiency of linguistic knowledge among primary studies on L2 writing. As the authors noted, while some studies did involve measures assessing processing efficiency of certain variables (e.g., GJTs to measure grammar knowledge processing efficiency), such measures were much

less frequently used compared to explicit knowledge measures. Among the L2 listening studies included in the meta-analysis, this trend was largely sustained with only very few studies involving processing efficiency type measures (e.g., Aural Vocabulary Test, GJTs) while the majority of the tests assessed explicit linguistic knowledge. Lastly, the primary studies included in the speaking meta-analysis almost exclusively used explicit knowledge measures to assess the correlates of L2 speaking. In sum, the measurement of processing efficiency clearly remains a limitation across all research domains of L2 skills.

### 3.3 Relationship between language-general correlates and L2 proficiency

Certain variables included in our meta-analyses are essential to language processing but are not specifically tied to the target language. These variables were therefore labelled as language-general variables (Hulstijn, 2015, 2019) in the meta-analyses and were categorized into three groups based on their nature: cognitive, conative, and affective.

The first group, language-general, cognitive variables include working memory, metacognition, and language aptitude. The meta-analyses on four L2 skills and their external variables showed that working memory and metacognition are popularly investigated, as indicated by their inclusion in all four studies. On the other hand, language aptitude was included in writing, listening, and speaking meta-analyses only.

As can be seen in Table 2, as for working memory, all but one correlation (with L2 speaking) were significant and were small to medium in size ( $r = .297$  to  $.334$ ). Additionally, all but one correlation (with L2 writing) for metacognition were significant, but their sizes ranged widely from small to large ( $r = .275$  to  $.624$ ). Lastly, language aptitude correlated very weakly with L2 listening, weakly with L2 writing, and moderately with L2 speaking ( $r = .105$ ,  $.281$ , and  $.481$ , respectively).

As for the language-general, conative variable, motivation was included in two meta-analyses examining L2 writing and L2 listening and showed small-to-medium size correlations with L2 writing ( $r = .338$  and  $.106$ , nonsignificant, respectively). Correlations between the two language-general affective variables, anxiety and attitude and each skill ranged from very weak to moderate in size ( $r = .098$  to  $.439$ ).

While the size of correlations does vary across correlates, one finding is that, generally, the relationships between linguistic knowledge, processing efficiency, and each L2 skill were stronger than the ones between language-general variables. For example, the correlation between vocabulary and L2 reading and between oral reading fluency and L2 reading ( $r = .724$  and  $.640$ , respectively) was stronger than that between working memory and L2 reading ( $r = .334$ ). The results were overall

consistent with the core-periphery model, suggesting language-specific knowledge (i.e., core components consisting of linguistic knowledge and processing efficiency) plays a more important role in explaining L2 proficiency than language-general abilities (i.e., peripheral components). Recall that the core-periphery model was initially proposed based on the results from a series of primary (not meta-analytic) studies conducted by Hulstijn and others. As the model was overall supported in our meta-analyses which included a much larger, systematically retrieved body of studies, the core-peripheral model finds stronger evidence in this volume.

Regarding the relationship between L1 and L2 skills, the strengths of correlations were moderate ( $r = .483$  between L1 reading comprehension and L2 reading comprehension;  $r = .426$  between L1 writing and L2 writing). Correlations were stronger among L2 skills, ranging from  $.530$  between L2 listening and L2 speaking to  $.822$  between L2 reading and L2 speaking. These results suggest an interconnected nature of L1 and L2 skills.

#### **4. Synthesis of findings of the two meta-analyses examining the relationship between L2 skills and their internal correlates**

Tables 3 and 4 summarize findings from internal correlates in relation to L2 skills. As explained in Chapters 5 and 10, internal correlates are (a) features of L2 writing or speaking that are predicted to be correlated with (b) L2 writing or speaking, and (a) and (b) are measured using the same L2 writing or speaking performances (e.g., lexical complexity of the spoken discourse produced by the study participant). Internal correlates can be divided into correlates measured by counting certain features and computing values (i.e., objectively measured correlates, or objective measures; e.g., the number of words per unit, called unit length) and correlates measured using rating scales and human judgment (i.e., subjectively measured correlates, or subjective measures; e.g., vocabulary and grammar). Results of internally and objectively measured correlates can be seen in Table 3, where those of internally and subjectively measured correlates can be seen in Table 4.

Although the strengths of correlations varied depending on correlates, correlations overall were weaker between objectively measured correlates and L2 writing or speaking than between subjectively measured correlates and L2 writing or speaking. For example, lexical complexity, an objectively measured correlate, was more weakly correlated with L2 writing than was vocabulary ( $r = .295$  and  $.888$ , respectively), a subjectively measured correlate. The same pattern was found in L2 speaking ( $r = .111$  and  $.876$ ). There may be two underlying reasons for this observation. First, subjective measures using rating scales can capture and assess broader

aspects of the target constructs. In the above example, lexical complexity consisted of diversity, density as well as sophistication, and these three aspects were measured using an objective measure. In contrast, when vocabulary was assessed using a subjective measure, these three aspects as well as others (that were not specifically measured in an objective measure of lexical complexity) were measured as well. The broader range of the target construct captured by subjective measures and the likely higher levels of individual variability present in the data could have resulted in a stronger correlation for subjective measures. Second, descriptions in rating scales or subjective judgment by humans have some weighting on particular aspects of the construct, depending on learners' proficiency levels. For example, rating scales or human raters may focus more on lexical diversity at novice and intermediate levels but they may shift their focus to lexical sophistication at an advanced level. Because development of lexical complexity may be non-linear, there is a possibility that human raters using rating scales can better capture subtle changes in L2 proficiency by changing weighting across different stages of learning.

Similarly, when comparing correlations across L2 writing and speaking, all of the strong correlates were associated with subjective measures: vocabulary ( $r = .888$  and  $.876$ , respectively), grammar ( $r = .837$  and  $.873$ ), coherence ( $r = .668$  and  $.911$ , nonsignificant), and content ( $r = .927$  and  $.713$ ). In contrast, all of the moderate correlates were consistently associated with objective measures: accuracy ( $r = .477$  and  $.492$ , nonsignificant), fluency ( $r = .570$  and  $.468$ ), and the weak correlate were associated with objective measures: syntactic complexity ( $r = .271$  and  $.326$ ). In some objective measures, the degrees of correlations varied across L2 writing and speaking (with some 95% CI overlaps): lexical complexity (weak and very weak;  $r = .295$  and  $.110$ , nonsignificant), lexical diversity (weak and moderate;  $r = .285$  and  $.472$ ), lexical density (very weak and weak;  $r = .134$  and  $-.251$ , nonsignificant), and lexical sophistication (weak and very weak;  $r = .317$  and  $-.074$ , nonsignificant). The finding that lexical complexity and its aspects (i.e., diversity, density, and sophistication), which are all objectively measured, showed varied correlations may suggest that the impact of lexical features may differ across the two L2 skills. Furthermore, although some measures had nonsignificant correlations, it was found that the effects of objective and subjective measures on L2 skills were similar, other than lexical complexity and its aspects, when the same measures were used across L2 writing and speaking.

In relation to Hulstijn's (2015, 2019) model, it is unclear whether some internal correlates can be thought of as core or peripheral components (e.g., mechanics and content). However, other correlates that are common across L2 writing and speaking seem to be clearly core components: accuracy, fluency (both objective measures), vocabulary, and grammar (both subjective measures). Tables 3 and 4 shows that all these measures had moderate or strong correlations with L2 skills, providing further support Hulstijn's (2015, 2019) model regardless of the type of measurements used (subjective vs. objective measures)

**Table 3.** Summary of correlations between objectively measured correlates and L2 writing and speaking

Correlate type	Correlates (objective)	Chapter 5			Chapter 10		
		L2 writing			L2 speaking		
		<i>k<sup>a</sup></i>	<i>k<sup>b</sup></i>	<i>r</i> [95% CI]	<i>k<sup>a</sup></i>	<i>k<sup>b</sup></i>	<i>r</i> [95% CI]
Linguistic knowledge	<i>Accuracy</i>	27	128	.477 [.373, .570]	8	36	.492 [−.089, .823] <sup>NS</sup>
	<i>Lexical Complexity</i>	37	282	.295 [.234, .354]	8	49	.110 [−.063, .277] <sup>NS</sup>
	<i>Diversity</i>	26	92	.285 [.224, .343]	6	2	.472 [.160, .698]
	<i>Density</i>	6	13	.134 [.031, .234]	4	9	−.251 [−.603, .182] <sup>NS</sup>
	<i>Sophistication</i>	29	177	.317 [.259, .374]	3	10	−.074 [−.485, .362] <sup>NS</sup>
	<i>Syntactic complexity</i>	32	245	.271 [.125, .405]	11	46	.326 [.099, .520]
	<i>Global</i>	29	110	.289 [.143, .423]	0	0	NA
	<i>Phrasal</i>	9	38	.218 [.066, .359]	0	0	NA
	<i>Specific</i>	17	97	.248 [.099, .386]	0	0	NA
	<i>Unit length</i>	0	0	NA	9	19	.392 [.125, .606]
	<i>Sentential subordination</i>	0	0	NA	4	8	.230 [−.139, .542] <sup>NS</sup>
	<i>Cohesion</i>	21	201	.198 [.096, .296]	0	0	NA
Processing efficiency	<i>Fluency (Speed fluency)</i>	30	73	.570 [.463, .661]	15	35	.468 [.302, .607]
	<i>Breakdown fluency</i>	0	0	NA	8	31	.361 [.224, .484]
	<i>Repair fluency</i>	0	0	NA	6	24	−.095 [−.296, .114] <sup>NS</sup>

Note. <sup>NS</sup> Statistically not significant. *k<sup>a</sup>* Number of studies. *k<sup>b</sup>* Number of correlations.

**Table 4.** Summary of correlations between subjectively measured correlates and L2 writing and speaking

Correlate type	Correlates (subjective)	Chapter 5			Chapter 10		
		L2 writing			L2 speaking		
		<i>k<sup>a</sup></i>	<i>k<sup>b</sup></i>	<i>r</i> [95% CI]	<i>k<sup>a</sup></i>	<i>k<sup>b</sup></i>	<i>r</i> [95% CI]
Linguistic knowledge	<i>Vocabulary</i>	13	17	.888 [.825, .929]	7	8	.876 [.716, .948]
	<i>Grammar</i>	9	15	.837 [.737, .901]	9	10	.873 [.765, .933]
	<i>Accuracy</i>	4	9	.781 [.361, .938]	0	0	NA
	<i>Language use</i>	7	9	.920 [.862, .954]	0	0	NA
	<i>Mechanics</i>	9	13	.766 [.524, .894]	0	0	NA
	<i>Coherence</i>	3	3	.668 [.450, .811]	3	3	.911 [−.549, .999] <sup>NS</sup>
	<i>Cohesion</i>	3	11	.688 [.259, .891]	0	0	NA
	<i>Organization</i>	14	21	.878 [.812, .921]	0	0	NA
	<i>Pronunciation</i>	0	0	NA	13	20	.803 [.640, .896]
	<i>Comprehensibility</i>	0	0	NA	3	4	.896 [−.325, .997] <sup>NS</sup>
	<i>Delivery</i>	0	0	NA	3	4	.833 [.405, .962]
Processing efficiency	<i>Fluency</i>	0	0	NA	11	12	.888 [.762, .949]
Content	<i>Content</i>	8	10	.927 [.892, .951]	4	4	.713 [.381, .882]

Note. <sup>NS</sup> Statistically not significant. *k<sup>a</sup>* Number of studies. *k<sup>b</sup>* Number of correlations.



## 5. Limitations and future research

We describe limitations and future directions that we believe would help researchers better understand the relationship between L2 proficiency and its correlates. First, as with meta-analytic studies in general, it would be useful to update the results in the meta-analysis chapters by adding more primary studies. Although each chapter author attempted to search, locate, and include relevant studies exhaustively, there could have been studies that had been missed. This was to some extent unavoidable given that a wide range of correlates was targeted in each meta-analysis. Moreover, new primary studies have been published since the search end dates of our meta-analyses which varied across studies from July 2017 (Chapter 4, for the reading meta-analysis) to August 2020 (Chapter 6, for the writing meta-analysis of external correlates). Furthermore, broadening the scope of a meta-analysis allows one to compare the strengths of correlates. However, managing such a meta-analysis is extremely time consuming and labor intensive. Although the authors of the meta-analysis chapters attempted to strike the balance between them, it would be necessary to add more studies to examine the findings in more detail.

Second, not all correlates have been researched equally frequently. This led the number of studies (and correlations) for some correlates to be small. For example, L2 reading meta-analysis included 51 studies for vocabulary as opposed to six studies for orthographic knowledge. This suggests that orthographic knowledge has been less often researched than vocabulary, although there were always possibilities that some studies failed to be identified. Further, in a small-sample meta-analysis, it would be challenging to conduct moderator variable analyses and interpret the results since the number of studies within each moderator becomes small. To address the small number of studies in meta-analysis, van Lissa (2020) recommended using a random forest algorithm. It is a machine learning technique suited for examining the relative importance of predictors explaining the outcome variable. van Lissa (2020) reported that the algorithm worked well in most cases with 20 studies except for some conditions (see his study for details). Obtaining 20 studies might not be possible for some under-researched or under-documented variables, but it might be interesting to add more studies over time and meta-analyze them using the proposed algorithm.

Third, as mentioned in Section 1 (Introduction) of this chapter, methodological details across the meta-analyses were not the same, as chapter authors were allowed to make decisions as described above. This was partly because it was challenging to agree upon every detail before conducting a meta-analysis. For example, studies differed in the type of reliability indices reported (that can be used to correct for measurement error; e.g., inter-rater agreement percentage, Cronbach's alpha, test-retest reliability, among others) and the frequency with which these indices

were reported. As a result, correction for measurement error was fully applied to the reading meta-analysis (Chapter 3), partially applied to the speaking meta-analysis (Chapter 11), and not applied to the remaining meta-analyses. Wiernik and Dahlke (2020) argue that one consequence of not correcting for measurement error is that correlations artificially shrink toward zero. Correction formulas are available in their article but they vary under different conditions. Careful considerations are required to apply the formulas. Thus, methodological differences varied across the meta-analyses and this might have led to difficulty in comparing findings across the meta-analyses in a straightforward manner. Although there might not be a quick solution for this, researchers need to consider how best to conduct and compare multiple meta-analyses.

Lastly, the meta-analyses in this volume synthesized correlation coefficients in which linear relationships are assumed, so other methods to examine nonlinear relationships would also be needed. Furthermore, relationships between variables were not considered. One way to factor in such multivariate relationships is to use meta-analytic structural equation modeling (Cheung, 2015). This would enable meta-analytic researchers to test models created based on theory and to further investigate the structures and correlates of L2 proficiency.

## 6. Conclusion

In this chapter, we have summarized and compared some of the findings – those that included the same variables – from each meta-analysis. The results from the relationship between L2 skills and their external correlates showed that linguistic knowledge (such as vocabulary and grammar) and processing efficiency were correlated more strongly with L2 skills than other variables. This provided support for the core-periphery model of L2 proficiency by Hulstijn (2015, 2019) that proposed linguistic knowledge and processing efficiency were more strongly related with L2 skills than other variables.

The strengths of association between correlates and target L2 skills were consistent across the meta-analyses. Peripheral, language-general variables were less strongly related to L2 skills, suggesting their weaker but important role in understanding and producing L2 language. Finally, objectively and subjectively measured correlated were both related to L2 skills, with the latter exhibiting a stronger relationship.

In addition to updating these findings above, future studies would benefit from analysis with methods that allow for considering the small number of studies and correlations and for correcting for psychometric artifacts.

## References

- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.
- Bachman, L. F. (2007). What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment. In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. E. Turner, & C. Doe (Eds.), *Language testing reconsidered* (pp. 41–71). University of Ottawa Press. <https://doi.org/10.2307/j.ctt1ckpccf.9>
- Bachman, L. F., & Cohen, A. D. (1998). *Interfaces between second language acquisition and language testing research*. Cambridge University Press.
- Bachman, L. F., & Palmer, A. S. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly*, 16(4), 449–465. <https://doi.org/10.2307/3586464>
- Bachman, L. F., & Palmer, A. S. (1983). The construct validation of the FSI Oral Interview. In J. W. Oller, Jr. (Ed.), *Issues in language testing research* (pp. 154–169). Newbury House.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford University Press.
- Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Pearson.
- Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1(1), 1–47. <https://doi.org/10.1093/applin/1.1.1>
- Carroll, J. B. (1961). *Fundamental considerations in testing for English language proficiency of foreign students. Testing the English proficiency of foreign students*. Center for Applied Linguistics.
- Carroll, J. B. (1972). Fundamental considerations in testing for English language proficiency of foreign students. In H. B. Allen & R. N. Campbell (Eds.), *Teaching English as a second language: A book of readings* (pp. 313–320). McGraw-Hill.
- Carroll, J. B. (1983). Psychometric theory and language testing. In J. W. Oller, Jr. (Ed.), *Issues in language testing research*. Newbury House.
- Chalhoub-Deville, M. (1997). Theoretical models, assessment frameworks and test construction. *Language Testing*, 14(1), 3–22. <https://doi.org/10.1177/026553229701400102>
- Cheung, M. W.-L. (2015). *Meta-analysis: A structural equation modeling approach*. Springer.
- Davidson, F. G. (1988). An exploratory modeling survey of the trait structures of some existing language test datasets. (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (8815771)
- De Jong, J. H. A. L., & Verhoeven, L. (1992). Modeling and assessing language proficiency. In L. Verhoeven & J. H. A. L. De Jong (Eds.), *The construct of language proficiency: Applications of psychological models to language assessment* (pp. 3–19). John Benjamins. <https://doi.org/10.1075/z.62.03jon>
- Goldberg, A. (2006). *Constructions at work: The nature of generalization in language*. Oxford University Press.
- Hoover, W., & Gough, P. (1990). The simple view of reading. *Reading and Writing*, 2, 127–160. <https://doi.org/10.1007/BF00401799>
- Hulstijn, J. H. (2015). *Language proficiency in native and non-native speakers: Theory and research*. John Benjamins.
- Hulstijn, J. H. (2019). An individual-differences framework for comparing nonnative with native speakers: Perspectives from BLC theory. *Language Learning*, 69(s1), 157–183. <https://doi.org/10.1111/lang.12317>

- Hymes, D. (1972). On communicative competence. In J. B. Pride & A. Holmes (Eds.), *Sociolinguistics: Selected readings* (pp. 269–293). Penguin.
- Juel, C., Griffith, P. L., & Gough, P. B. (1986). Acquisition of literacy: A longitudinal study of children in first and second grade. *Journal of Educational Psychology*, 78(4), 243–255. <https://doi.org/10.1037/0022-0663.78.4.243>
- Lado, R. (1961). *Language testing*. Longmans, Green, and Co.
- Luoma, S. (2004). *Assessing speaking*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511733017>
- McNamara, T. F. (1996). *Measuring second language performance*. Longman.
- Oller, J. (1979). *Language tests at school: A pragmatic approach*. Longman.
- Oller, J. W., Jr. (1983). Evidence for a general language proficiency factor: An expectancy grammar. In Oller, J. W., Jr. (Ed.), *Issues in Language Testing Research* (pp. 3–10). Newbury House.
- Paquot, M., Gries, S. T., & Yoder, M. (2021). Measuring lexicogrammar. In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 223–232). Routledge.
- Plonsky, L., & Oswald, F. (2014). How big is “big”? Interpreting effect sizes in L2 research. *Language Learning*, 64(4), 878–912. <https://doi.org/10.1111/lang.12079>
- Römer, U. (2009). The inseparability of lexis and grammar: Corpus linguistic perspectives. *Annual Review of Cognitive Linguistics*, 7(1), 140–162. <https://doi.org/10.1075/arcl.7.06rom>
- Römer, U. (2017). Language assessment and the inseparability of lexis and grammar: Focus on the construct of speaking. *Language Testing*, 34(4), 477–492. <https://doi.org/10.1177/0265532217711431>
- Ruegg, R., Fritz, E., & Holland, J. (2011). Rater sensitivity to qualities of lexis in writing. *TESOL Quarterly*, 45(1), 63–80. <https://doi.org/10.5054/tq.2011.240860>
- Spearman, C. (1904). “General Intelligence,” objectively determined and measured. *The American Journal of Psychology*, 15(2), 201–292. <https://doi.org/10.2307/1412107>
- Spearman, C. (1927). *The abilities of man: Their nature and measurement*. Macmillan.
- Van Lissa, C. J. (2020). Small sample meta-analyses: Exploring heterogeneity using MetaForest. In R. van de Schoot & M. Miočević, (Eds.), *Small sample size solutions: A guide for applied researchers and practitioners* (pp. 186–202). <https://doi.org/10.4324/9780429273872-16>
- Vollmer, K., & Sang, F. (1983). Competing hypotheses about second language ability: A plea for caution. In J. W. Oller, Jr. (Ed.), *Issues in language testing research* (pp. 29–79), Newbury House.
- Wallace, M. P. (2022). Individual differences in second language listening: Examining the role of knowledge, metacognitive awareness, memory, and attention. *Language Learning*, 72(1), 5–44. <https://doi.org/10.1111/lang.12424>
- Wang, M., Liu, Y., & Perfetti, C. A. (2004). The implicit and explicit learning of orthographic structure and function of a new writing system. *Scientific Studies of Reading*, 8(4), 357–379. [https://doi.org/10.1207/s1532799xssr0804\\_3](https://doi.org/10.1207/s1532799xssr0804_3)
- Wiernik, B. M., & Dahlke, J. A. (2020). Obtaining unbiased results in meta-analysis: The importance of correcting for statistical artifacts. *Advances in Methods and Practices in Psychological Science*, 3(1), 94–123. <https://doi.org/10.1177/2515245919885611>



# Index

## A

- accuracy 115, 137–139, 310–311, 325–326
- affective feature 243–244, 261–263
- age 46, 118, 170–173
- anxiety 244, 262–263, 345, 356, 362
- aptitude 240–241, 259–260, 344, 355, 361, 363
  - language analytic ability 241, 259
  - phonetic coding 241, 259
  - rote memory 241, 259
- automatic scoring 102–103

## C

- cognitive abilities 2–4, 168, 217
- cognitive information processing model 213
- Cognitive Revolution 1–3
- coherence 312–313, 327
- cohesion 116–117, 140–142
- components of reading 13
- components of writing 116
- comprehensibility 312
- content 312–313, 327
- Core-Periphery Model 371–372, 377, 379, 383
- cross-linguistic transfer 14–17, 19, 21
- crossover effect 16–22

## D

- decoding 6, 11–13, 15–16
- delivery 312–313, 327–328
- differentiating listening from reading 222
- discrete-point testing 97
- Dual Coding Theory 214, 221, 229

## F

- fluency 115–116, 139–140, 309–310, 323–324, 328–329

## G

- generalizability 100
- global rating 312
- grammar knowledge 37–38, 52, 60, 64–65, 237, 307, 341–342, 350–351, 357–358
- grammatical complexity 311–312, 326, 328–329

## I

- individual differences in writing 87
- inhibition hypothesis 162, 170
- interactive speaking/listening ability 221–222

## L

- language setting 46, 61, 67
- language learning context 136, 138, 143, 173, 175–176, 188, 190
- L1–L2 language distance 46, 73, 114, 140, 170
- L1–L2 script distance 46, 49–50, 52–53, 55–56
- L1 reading comprehension 29–31, 34, 38, 43, 47, 53–54, 60–61, 65
- L2 listening ability 213–214, 216–219
- L2 listening comprehension 39, 54–55, 60, 65–66, 342, 352, 356, 361, 363
- L2 proficiency 1
- L2 pronunciation 294–295
- L2 reading comprehension 5, 29–32, 34, 43–45, 47–49, 51–52, 54–73, 342, 351–352, 361
- L2 speaking 339–363

- L2 speaking pedagogy 293

- L2 writing 88–89, 342, 353, 361
- language-general correlates 73, 378
- lexical complexity 114, 136–137, 311–312, 326–327
- linguistic knowledge 93, 161, 189, 196, 235–236, 237–240, 256–259, 279, 340, 371–373, 375–379, 381, 383
- linguistic processing in reading 6

## M

- measurement characteristic 53, 64
- metacognition 10, 30, 42, 58–59, 70–71, 169, 219, 343–344, 347, 349, 354–355, 358–360
- metacognitive awareness 241–242, 260
- morphological knowledge 41, 58, 69–70, 92, 240, 258–259, 279

## O

- objective measurement 339
- oral reading fluency 42–43, 59, 71–72
- orthographic knowledge 40–41, 57, 68–69

## P

- phonological awareness 35–36, 49, 60–62
- processing efficiency 377–378

## R

- rating scale 313, 323
- ratio measure 318, 323
- real-world spoken language 228–229

S

scoring rubric 120–121  
Simple View of Reading  
    11, 33, 39  
Simple View of Writing 167  
speaking 188–189, 308–309,  
    339–340, 347  
syntactic complexity 113–114,  
    133–135

T

testing language 40, 46, 56–57,  
    68  
Threshold Hypothesis 160, 173,  
    190, 196

V

vocabulary 256–258, 310–311,  
    325, 328  
    vocabulary depth 238,  
        256–257  
    vocabulary knowledge 8,  
        36–37, 50, 62–64, 340–341,  
        350, 357  
    vocabulary size 238,  
        256–257

W

working memory 39–40, 56,  
    67–68, 242–243, 260–261,  
    343, 358–360

writing assessment 98–100  
writing processes 89–90, 92,  
    94, 96, 98

This edited volume is a collection of theoretical and empirical overviews of second language (L2) proficiency based on four skills: reading, writing, listening, and speaking. Each skill is reviewed in terms of how it has been conceptualized, measured, and studied over the years in relation to relevant (sub-) constructs of the language skill under discussion. This is followed by meta-analyses of correlation coefficients that examine the relationship between the L2 skill in question and its component variables. Unlike most meta-analyses that have a limited range of variables under investigation, our meta-analyses are much larger in scope to better clarify such relationships. By combining theoretical and empirical approaches, the book is helpful in deepening the understanding of how subcomponents or various variables are related to a particular L2 skill.

ISBN 978 90 272 1117 0



9 789027 211170

**John Benjamins Publishing Company**