Methodological Issues in Literacy Research Across Languages: Evidence From Alphabetic Orthographies

Timothy C Papadopoulos, Valéria Csépe, Mikko Aro, Marketa Caravolas, Irene-anna Diakidoy, Thierry Olive

PREPRINT

Papadopoulos, T., Csépe, V., Aro, M., Caravolas, M., Diadikoy, I-A., & Olive, T. (2021). Methodological issues in literacy research across languages. Reading Research Quarterly. 56(S1), S351–S370. https://doi.org/10.1002/rrq.407.

Abstract

Research on literacy has become universal and is essential for researchers of various disciplines, educators, and psychologists. This paper examines the most important methodological challenges that arise when conducting literacy research across languages, some of which have long been acknowledged in the relevant literature. Specifically, we focus on challenges related to research on word reading, spelling, passage comprehension and writing, ranging from the target skills, constructs and assessment issues to the matching of the samples and measurement and factorial invariance issues. We conclude that although theoretical and applied issues are addressed in the literature, to date, this happens only with limited relevance for reading and writing research across languages. The discussion provides some relevant evidence from a neuroscience perspective to promote useful insights and greater methodological rigor in literacy research across languages.

Keywords: literacy research across languages; methodological challenges; reading, spelling, writing, passage comprehension

Methodological Issues in Literacy Research Across Languages: Evidence from alphabetic orthographies

Literacy research across languages is essential for researchers of various disciplines, educators, and psychologists. However, for the broad range of research to be informative and useful, the data collected in various languages has to be comparable (see Verhoeven & Perfetti, 2017). This comparability hinges on diverse issues such as the definition of constructs (e.g., Authors, 2012), the precision of assessment and research methods (e.g., Authors, 2013), the measurement and factorial invariance of the predictor and outcome measures (e.g., Authors, 2012), or even challenges at the level and complexity of statistical analysis and deriving conclusions (e.g., Authors, 2003). The present position paper aims to review some of the most relevant methodological issues involved in literacy research across languages and provide guidelines for addressing these issues.

Literacy relates to reading, spelling, reading comprehension, and text composition. In learning to read and write, children learn to encode language into their writing system and decode printed words to speech to derive meaning (see Alves, Limpo, & Joshi, 2020). Much of the history of literacy research on European alphabets[1], to which we restrict ourselves in the present review, shows that the field has been driven by data acquired in cross-linguistic studies. Cross-linguistic research focuses on the development of these fundamental literacy skills in different languages, varying primarily in orthographic consistency. It also investigates various

[1] We focus on "European alphabets"; the term refers to Western and Centric European, Roman alphabets and English (an outlier orthography among European alphabets). Although, we recognize that significant cross-linguistic research is also carried out in non-European alphabets or non-alphabetic scripts (e.g., Arabic, Korean, Chinese; e.g., Katzir, Shaul, Breznitz, & Wolf, 2004; Kuo & Anderson, 2006; McBride, 2016), we hope this paper to stimulate discussion and further research on the methodological issues derived from non-alphabetic orthographies too.

relations among fundamental components, or precursor skills, and between literacy skills themselves.

Most of the cross-linguistic research relates to reading (e.g., Authors, 2008; Landerl et al., 2019; Ziegler et al., 2010) and spelling (e.g., Moll et al., 2014), and to a lesser extent to reading comprehension (e.g., Authors, 2019) or writing (e.g., Strömqvist et al., 2002). The basic assumption underlying our understanding of the processes involved in learning to read, write, spell, and comprehend texts is that these core processes are similar in all alphabetic languages.

Regarding *reading* and *spelling* development, we accept that the relative contribution of underlying linguistic or cognitive skills is expected to vary as a function of *orthographic depth*, that is, the consistency of print-to-speech correspondences (Katz & Frost, 1992; Authors, 2003; 2009). Similarly, to the extent that there is a specific set of core functions and skills that underlie *reading comprehension*, such as vocabulary, semantic, or syntactic processes, then all factors contributing to the development of reading comprehension in one alphabetic language should function in the same way in other alphabetic languages regardless of the languages' orthographic depth (McClung & Pearson, 2019). Finally, given the strong relationship between reading and writing, *writing* is argued to rely on highly similar sets of processes (Kim, 2020). These relate to various linguistic levels (e.g., orthographic, semantic, syntactic, or pragmatic), the writer's characteristics and the social context in which writing occurs (Berninger, Swanson, & Griffin, 2015; Fitzgerald & Shanahan 2000). Nevertheless, this interaction has not been studied as thoroughly, particularly in cross-linguistic experiments.

Literacy Research Across Languages

Considering all of the above, writing systems can show minor but significant variations in development, based on the consistency and the complexity with which print reflects speech in

alphabetic languages (Share, 2008) or as a function of task characteristics and writing contexts, for writing and comprehension (Schmalz, Marinus, Coltheart, & Castles, 2015). This variability in alphabetic languages is reflected in theories of reading development that have been proposed, including among others, the Orthographic Depth Hypothesis (Katz & Frost, 1992), the parallel distributed processing and connectionist division of labour models (Harm & Seidenberg, 1999; Seidenberg, 2011), the Dual Route Cascaded (DRC) model (Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001) or the Psycholinguistic Grain-Size Theory (Ziegler & Goswami, 2005). All these theories share two fundamental assumptions: (a) that orthographic information interacts with lexical knowledge to produce acceptable word pronunciation and access to meaning; (b) that the encoding of the *implicit structure of orthography* is one of the prime factors of *reading fluency,* the primary index of reading achievement over reading accuracy (International Literacy Association, 2018). Particularly concerning cross-linguistic comparisons, the focus on orthographic consistency has helped to explain the mechanism by which reading fluency is accomplished. In consistent orthographies, children achieve high accuracy levels after a few months of reading instruction because grapheme-phoneme decoding strategy use becomes more efficient. In contrast, in inconsistent orthographies, fluency may be achieved with the inclusion of multiple unit size mapping strategies to become more reliable and readily available (Authors, 2018). This demands the joint contribution of other underlying skills such as rapid automatized naming (RAN) and orthographic processing skills (Georgiou et al., 2008) (see next sections).

These theories aim to guide reading and writing research approaches across alphabetic languages. After all, any reading model's validity must be tested across languages that differ in orthographic depth to confirm whether it is general or language-specific (Georgiou, Das, & Hayward, 2009). However, in this quest, research has overlooked broader and deeper

methodological issues emerging from studies carried out across alphabetic languages. For instance, Schmalz et al. (2015) argue that if the orthographic depth is the starting point for any discussion that centres on reading development across languages, it has to be more precisely defined. The authors suggest that since the degree of complexity and unpredictability[2] of print-to-speech correspondences affects skilled reading and reading acquisition in different ways, differences across orthographies may be attributed erroneously to orthographic depth post-hoc. There is always the possibility that observed differences are the result of other language-level differences that remained uncontrolled. For this reason, we need to consider some of the methodological issues emerging from cross-linguistic studies.

Methodological Challenges

Literacy research across languages presents several unique challenges that concern not only whether the investigated constructs, models or theories are relevant in the research context, as described above. They are also associated with issues related to different phases of an investigation.

Before research is conducted, one of the central questions is whether the data acquisition measures are equivalent or invariant across the language groups in focus, representing the same ability or skill construct. Relevance (as defined by the unit of analysis; see, for example, Durgunoğlu & Öney, 1999; Authors, 2010), reliability and validity, translation, and the normalization of the measures (e.g., Landerl et al., 2019; Moll et al. 2014) or the use of single or multiple indicators for measuring a particular skill or ability (e.g., Authors, 2012; Parrila,

[2] *Complexity* refers to aspects such as syllabic complexity, morphological complexity, orthographic density, or even the proportion of mono- versus polysyllabic words in a language. *Unpredictability* refers to the non-lexical procedure which uses knowledge of print-to-speech regularities to assemble a word's pronunciation, that is, from orthography to phonology. For more information, see Schmalz et al. (2015).

Aunola, Leskinen, Nurmi, & Kirby, 2005; Soodla, Torppa, Kikas, Lerkkanen, & Nurmi, 2019), are usually critical matters in cross-linguistic research. Likewise, the matching of the samples, in terms of relevant demographic variables (gender, age of school entry, family background; see Berman & Verhoeven, 2002; Authors, 2013; Parrila et al., 2005), cognitive abilities (e.g., Authors, 2010) or their representativeness of the target population, facilitates the control of potentially confounding extraneous variables (Soodla et al., 2019).

During a study, aspects such as the testing conditions, the timing and place of data collection, the equivalence of instructions and the guidelines for scoring could also affect the equivalence of the participants' responses. For example, data regarding time and place of data collection may be less apparent and provided in online supplements when publication outlets impose strict length requirements on papers (e.g., Authors, 2013). Nevertheless, it is important not only to provide such information but to consistently ensure that any unforeseen performance differences are not the result of the conditions in which students are tested. The administration equivalence also, which refers to dimensions such as the respondents' familiarity with the test instruments as a result of educational practices (Parrila et al., 2005) or their psychological reactions, such as the level of anxiety (Anthony & Lonigan, 2004), is rarely addressed in the relevant literature (see Soodla et al., 2019, for a relevant argument). Such dimensions could threaten the validity of the findings with serious implications for research and instruction.

After the study is completed, equivalence has traditionally focused on issues of measurement and factorial invariance. The measured constructs must be interpreted in a conceptually similar manner across different language groups to yield comparative data. Testing for the equivalence of measures allows the detection of major threats of construct validity, construct underrepresentation, and construct irrelevant variance (Authors, 2012; Tate, 2003) and,

therefore, controls the possibility of unreliable results due to measurement bias (Milfont & Fischer, 2010). Asil and Brown (2016) argue that there are reasonable grounds to doubt the invariance of responses to tests despite careful adaptation processes. Particularly, regarding reading comprehension, they suggest that several factors relevant to language, culture, cognitive, and economic development are likely to influence these skills. Therefore, unless measurement invariance holds across languages, examining differences in the performance level across languages is meaningless.

Another relevant issue concerns the variety of statistical approaches used to investigate the role or contribution of different predictor variables to literacy skills across languages. One family of measures relies on hierarchical regression analysis, with language as a predictor (Step 1) and language X skill interaction term as a predictor (Step 3), after controlling for the effects of the cognitive skills (Step 2) (e.g., Patel, Snowling, & de Jong, 2004). Multi-group or multilevel analysis is another, more advanced way to compare the strength of predictions across languages (e.g., Authors, 2013; Georgiou et al., 2008; Georgiou, Torppa, Manolitsis, Lyytinen, & Parrila, 2012). Finally, various approaches focusing on *regularity, consistency*, and *entropy* are used to quantify and statistically control variation in orthographic transparency, as outlined by Borleffs and colleagues (2017) in their review. Regardless of the method of analysis chosen, it is evident from the above that controlling several issues at different phases helps establish comparability or equivalence at every stage of the research process and minimizes or controls results bias.

The present paper

The present paper intends to contribute to the ongoing discussion about a greater methodological rigour in literacy research conducted across European languages. We discuss the results from different studies across four major areas of basic and advanced literacy skills: word

reading, spelling, reading comprehension, and writing. Each literacy skill is considered from the perspective of target skills and constructs, assessment issues, and guidelines for further research.

In doing so, we propose the Triple Foundation Model (Authors, 2015) as a testable framework for cross-linguistic investigations of the impact of orthographic input on early, word-level literacy development for learners of alphabetic orthographies. This model expands on Byrne's (1998) "dual foundation model," which was elaborated based on research into reading acquisition in English. The model proposed that phoneme awareness (PA), broadly defined as the ability to perceive and manipulate the sounds of spoken words (Authors, 2012), and letter knowledge are two reciprocally related prerequisite skills for reading and spelling development. The Triple Foundation Model recognizes that a third foundational cognitive skill, rapid automatized naming (RAN) – defined as the ability to name as fast as possible highly familiar symbols (Kirby, Georgiou, Martinussen, & Parrila, 2010; Authors, 2016) – also needs to be included. Consequently, acquiring word reading and writing skills depends crucially on three core cognitive constructs: the ability to have conscious awareness of and the ability to manipulate oral sublexical units (e.g., phonemes, syllables, morphemes), eventually, those which correspond to the basic symbols of one's orthography; the ability to learn the available set of writing symbols of one's orthography (e.g., letters, diphthongs, multi-level graphemes); and the ability to establish and use quick and efficient connections between the linguistic units and their corresponding orthographic units, potentially measured by RAN[3].

[3] A project that used the Triple Foundation Model as a framework for cross-linguistic investigations of word-level literacy development was the European Initial Training Network, ELDEL, which has been described in several publications (Authors, 2012, 2013, 2017, 2019). The interested reader is invited to examine several parallel measures that are available at www.eldel-mabel.net, which were adapted across five languages from the ELDEL project. A publication on the methods used to create this battery is in preparation.

Our review does not focus on other relevant issues, such as vocabulary or metalinguistic awareness. We restrict the scope to domains directly related to aspects of literacy, not general language development. The discussion also touches on evidence from a neuroscience perspective to gain useful insights into reading development across alphabetic languages. Data relevant to typical and atypical reading development and factors affecting both the process and the outcome at the text processing level are also addressed. In the end, we underscore the need for a full understanding of the differences in literacy research that may exist in different languages and how these differences may affect study results and that conducting cross-linguistic studies requires a lot of a priori thinking.

Word Reading Research Across Languages

Target skills and constructs

In alphabetic orthographies, beginning readers need to understand that there are predictable relationships between written letters and speech sounds. Mastering these language-specific grapheme-phoneme correspondences (GPC) and phonemic assembly allows independent reading and advanced skill levels in all alphabetic orthographies (Share, 2008). In most transparent orthographies, basic GPC, at the level of single graphemes (or even letters), are sufficient for accurate decoding of new words encountered in texts. In most irregular orthographies, such as English, more complex and context-specific G-P correspondences and exceptions to these pronunciation rules have to be learned for accurate reading. Thus, the accurate pronunciation of a word often requires learning of larger representational units, such as rimes or even whole words (Ziegler & Goswami, 2005).

Cross-linguistic studies have shown that children learning to read in a writing system with unreliable and complex GPC rules, such as English, acquire fundamental reading skills

more slowly than typical development in more regular orthographies (e.g., Seymour et al., 2003). This finding is very consistent: there is no cross-language comparison on early word-level reading that provides different findings.

English-based models of reading acquisition typically reflect this need for context- or word-specific information even at the early stages of reading by emphasizing dual routes for word recognition (e.g., Zorzi, 2010) or different divisions of labour between the activation of phonological and semantic codes during word identification (e.g., Seidenberg, 2011; Smith, Monaghan & Huettig, 2021). This notion of dual routes for words is reflected in the tradition of focusing separately on decoding skills, assessed with pseudoword reading tasks, and word recognition skills, assessed with familiar or exception word reading tasks. Although the development from novice to expert reading reflects a change from early serial, algorithmic processing of small units towards rapid, direct retrieval mechanisms (Share, 2008), it is still an open question of how the characteristics of different orthographies modulate this shift.

Regarding the predictors of reading development across languages, research has often focused on three skills: phonological awareness, rapid automatized naming, and letter naming. It is worth noting that the relation between word reading and different phonological unit levels might vary across languages (Georgiou et al. 2008). RAN has been a powerful predictor of learning to read across different alphabetic orthographies (e.g., Moll et al., 2014). Letter knowledge has also been a stronger predictor of initial reading skills in English than in more transparent orthographic systems (e.g., Spanish and Czech; Authors, 2013). However, whether these two skills are primary precursors of reading performance across languages has been inconclusive. First, RAN assesses a wide range of cognitive skills and the two measures used, namely speed and accuracy, reflect different processes, including the speed of perception of the

object to be named (Kirby et al., 2010, Authors et al., 2016). Second, as Ziegler and his colleagues (2010) reported, in Grade 2 participants representing five European languages at different positions along the transparency continuum (Finnish, Hungarian, Dutch, Portuguese, and French), phonological awareness was strongly associated with reading performance. However, this relationship was modulated by the orthography's transparency, being stronger in less transparent orthographies. Third, in a follow-up study focusing on a somewhat different set of languages (English, French, German, Dutch and Greek) from Grade 1 to Grade 2, Landerl and her colleagues (2019) did not find a universal model of predictive patterns between RAN, phonological awareness, and reading. The authors used phonologically matched items for assessing phonological awareness, and identical assessment procedures for RAN and reading, with language-specific items. While RAN was a consistent predictor of reading fluency in all five orthographies, the association between phonological awareness and reading was mostly interactive. The authors concluded that RAN taps into universal cognitive mechanisms involved in reading. In contrast, the relationship between phonological awareness and reading is more complex and depends on factors such as orthographic complexity, developmental stage, and task requirements.

With regard to poor reading performance or dyslexia, it is generally agreed that it is best explained by a multifactorial model (e.g., Peterson & Pennington, 2012), considering several individual and environmental factors. Three subcomponents of phonological processing skills are typically deficient in dyslexic individuals: phonological awareness, phonological working memory and lexical access (see Landerl, 2019 for a recent review). Tasks measuring the ability to manipulate sublexical phonological segments require phonological awareness but also phonological working memory. Lexical access is typically assessed with RAN tasks. As Landerl

(2019) notes, although problems in these phonological processing skills have been shown for dyslexia across a variety of alphabetic orthographies, their relative weight as a background deficit might vary depending on the features of the writing system. It has been suggested that in transparent orthographies, simple and regular GPC might be protective of the effects of phonological processing problems. It has also been suggested that RAN deficits reflecting problems in lexical access are more relevant as an underlying impairment of dyslexia in regular orthographies, as fluency problems emerge as the primary manifestation of poor reading in transparent orthographies.

Although most of the results speak for a complex relationship between the underlying processes and reading development, the available studies do not give conclusive answers to the question of language-universal and language-specific predictors of reading development. This is partly due to several methodological differences, such as variation in the age of assessment, reading instruction methods, school entry age, measures, or criteria for participant selection. The following section discusses some of these issues with a focus on word-level reading.

Assessment Issues

Several factors challenge cross-linguistic studies focusing on word-level reading. Literacy, for example, as expressed primarily by the home literacy environment (e.g., Manolitsis, Georgiou, Stephenson, & Parrila, 2009) and teaching, have a different status between countries due to historical or cultural reasons (Dockrell et al., 2021). Besides, the educational systems have marked differences regarding school entry age or the role of early education. Between school systems, there are also differences in teacher qualifications and the support available for learning problems. These socio-cultural and educational differences are best controlled in cross-linguistic studies carried out in multilingual contexts within the same country, where such differences are

minimal. Examples of such approaches come from comparisons of English and Welsh readers (e.g., Spencer & Hanley, 2003). Similarly, van Daal and Wass (2017) studied reading development in Scandinavian countries with closely related languages and societal contexts but varying orthographic regularity.

With regard to cross-linguistic investigations of word (or pseudoword) reading, one of the most serious challenges to validity arises from the selection of measures and materials. The basic question is how to ensure that the measures and reading materials are as equal as possible to allow conclusions about similarities and differences of typical and atypical reading development across languages involved.

When assessing word-level reading, there have been three different approaches to ensure the comparability of reading items. First, some studies have used the translation of equivalent words (e.g., Spencer & Hanley, 2003). Although this might be a good control for item familiarity or frequency, translations are often not comparable about word length or other factors, such as neighbourhood density or morphological complexity. Second, some studies have tried to ensure comparable familiarity with word items by selecting them in each language from age-appropriate reading materials (e.g., Seymour et al., 2003). The problems with this approach also relate to problems in the comparability of word length and complexity. Third, some studies have used materials based on cognates, words having a common etymological origin (such as English *garden, night* and German *Garten, Nacht*) to ensure similarity of the materials (e.g., Rau et al., 2016). This approach is understandably applicable only in languages having shared roots, such as Germanic languages or languages that share loan words with other languages. In practice, matching the materials perfectly becomes difficult, especially when orthographies in focus belong to different language families.

Using pseudowords to assess decoding skills allows one to bypass some of the structural differences between languages that are hard to control for with word items. When creating a comparable set of pseudowords, one needs to ensure that the pseudoword structures used conform to each language's phonotactic rules. Nevertheless, also this approach comes with complications. For example, Seymour et al. (2003) used very simple pseudoword structures varying from CV, VC to CVCV, VCVC in complexity to compare beginning readers' decoding skills in thirteen languages. Even with these simplest structures, the resulting pseudowords did not necessarily represent natural word structures in all languages involved. Further, the frequency distribution of various syllable and word structures across languages could not be controlled for.

Another methodological challenge of cross-linguistic studies relates to the scoring of reading accuracy. This is usually relatively straightforward since there is usually only one correct pronunciation. However, this is not the case with pseudowords. Whereas any pseudoword has only one possible pronunciation in a fully regular orthography, there might be many plausible pronunciations in irregular orthographies. Using an analogy with an existing word or pronunciation based on a large unit, such as rime, might result in a completely different outcome than assembling pronunciation based on single GP correspondences. Authors (2003) used a lenient scoring principle to take this into account, allowing all permissible pronunciations based on these alternative grain sizes to measure knowledge of orthographic conventions instead of phonemic recoding skill.

Since the development of reading accuracy is more protracted in irregular writing systems, and it reaches ceiling before the end of Grade 1 in languages with a transparent orthography, the cross-linguistic comparisons of reading development tend to focus mostly on

reading fluency in recent years (e.g., Georgiou et al., 2020; Moll et al., 2014). In reading research, fluency is typically operationalized as a measure combining reading accuracy and rate (e.g. words read correctly within a time limit), since an objective assessment of the third aspect of reading fluency, prosody is hard to achieve. Nevertheless, comparisons on reading rate are meaningful only when there are no large differences in reading accuracy.

Guidelines for further research

On the one hand, the cross-linguistic studies within alphabetic languages have shown basic similarities in the principles of learning to read (see Verhoeven & Perfetti, 2017), its prerequisites (e.g., Authors, 2012) and deficits related to dyslexia (e.g., Ziegler, Perry, Ma-Wyatt, Ladner, & Schulte-Körne, 2003). On the other hand, they have shown that the differences between languages and their orthographies play a role in the developmental rate of reading and the relative role of different language skills as prerequisites of development or the potential deficits underlying dyslexia (e.g., Landerl et al., 2019; Authors, 2003). These differences are also typically reflected in the practices related to literacy instruction or assessment. In regular orthographies, reading instruction is often based on synthetic phonics, and the identification of dyslexia relies primarily on fluency measures emphasizing reading rate (e.g., Authors, 2017). Respectively, in irregular writing systems, reading instruction may also emphasize whole words (see Perfetti & Harris, 2017).

As summarized earlier, comparable measures of global word reading skills can be very hard to develop across languages. Since comparability is hard to achieve, one way forward would be to quantify and control statistically the variation observed between languages (see Borleffs et al., 2017, emphasizing *entropy*). However, approaches based on onset mappings only

miss many irregularities and probably underestimate language differences. Developing more precise methods for quantifying language differences is a future enterprise for research.

How to deal with the comparability issue also relates to the goal of cross-linguistic research. When the goal is to explicitly compare the rate of development or attainment of the skill in different orthographies, the equivalence of the measures on item level is a more critical methodological challenge. However, when the research questions relate to developmental associations between language skills and literacy development, attention should be paid more to the measures' ecological validity and measurement invariance. The age-expected level of literacy is not universally dependent on grade level or age but also relates to language, orthography, and the curricular goals of the various educational systems. For the same reason, the cut-off criteria for dyslexia refer to a certain deviation of age-expected skill level compared to language-specific normative data instead of absolute values of accuracy or speed universally. The relevant question is whether the measures reflect the same skill constructs across languages at the age-appropriate level in the particular orthographies. This would emphasize designing language-specific measures that reflect the frequency distribution of the linguistic features (e.g., phonemes, phonotactics, syllable structures) and age of acquisition in respective languages.

As shown in this review, the focus on the development of fluent, expert-level reading has been a lot scarcer as opposed to research on early reading acquisition (Share, 2008). The models accounting for fluency development emphasize a rather dichotomous shift from early serial, alphabetic processing to later direct recognition of words based on word-specific orthographic knowledge. Although this account might be relevant in describing the development of fluency in English, it might be more problematic in describing development in many other orthographies. There is a large variation between languages about, for example, syllable structures, length of the

common words in a text and their morphological complexity, and productiveness of word-formation. This variation suggests that a clearer focus on the role of sub-lexical processing at an intermediate level – units larger than single graphemes but smaller than words – would be useful for understanding reading fluency development and its problems. It seems plausible to expect that expert-level reading requires combinatorial skills utilizing knowledge of these sub-lexical units. To date, the focus on sub-lexical units in reading research has been rare. Cross-linguistic comparisons would serve well in identifying the differences in grain sizes of orthographic or morphological units relevant in fluency development in different languages and in developing means of support for reading fluency development.

Spelling Research Across Languages

Target skills and constructs

Learning to write words correctly is an essential part of becoming literate. It is a skill that requires both accurate retrieval and correct recoding of written symbols representing spoken words (e.g., Treiman & Kessler, 2014). Since reading relies on recognition and spelling on recall, spelling takes longer to learn than word reading, which seems to be true across languages and orthographies (e.g., Verhoeven & Perfetti, 2017). Spelling is defined as the ability to produce the letters of words in the correct order, according to orthographic conventions; this skill is relevant mostly to languages using alphabetic orthographies. Cross-linguistic studies of spelling acquisition remain relatively rare for several reasons, some of which are considered here.

Important differences exist across countries and sometimes across national and international organizations in the terminology relating to spelling difficulties. For example, in the English-speaking context, spelling impairment is seen as an integral part of the *dyslexia* profile, along with reading impairment (e.g., DSM 5, 2013; National Reading Panel, 2000). In contrast,

in many European countries, reading and spelling difficulties are seen as the separable disorders of dyslexia and *dysorthographia* (e.g., see overview in Authors, 2019). In yet other contexts, the term *dysgraphia* with reference to a specific disorder of writing skills, broadly defined, and which seems to have poor fine motor skills *and/or* weak language skills at its core (e.g., the US-based *LDA-America*; Australian *Dyslexia SPELD Foundation*). The International Dyslexia Association (IDA) uses the term dysgraphia in reference to those who "may have only impaired handwriting, only impaired spelling (without reading problems)", or both impairments, which are the consequence of deficient orthographic coding (storing spelling representations in memory) and planning sequential finger movements.

These differences in nomenclature may reflect the fact that spelling (and writing) difficulties manifest more noticeably in some languages, especially those with relatively high letter-sound consistency, such as Czech, Slovak, Finnish, Greek, Hungarian, and even French. For example, in English, correlations between reading and spelling accuracy persist into adulthood, and both skills tend to be impaired in individuals with dyslexia. In contrast, in languages with consistent orthographies, the persisting indicators of word-level difficulties tend to manifest differently for reading (speed) and spelling (accuracy) after the second grade. Therefore, spelling disorders may seem more obvious or may perdure as the only literacy problem. The important point for researchers designing or interpreting cross-linguistic research is that such terminological differences highlight the need to clearly define how spelling skills and their impairment are conceptualized in each country under study and ensure that similar populations with similar impairment profiles are recruited to the study.

An important theme running through this paper concerns the importance of orthographic consistency and its influence on the developmental trajectories of literacy acquisition in typical

and disordered learner groups. Spelling-sound and sound-spelling consistency can be calculated in different ways that may take the surrounding context and the frequency of various units into account. Moreover, it can be considered alongside other closely related constructs, such as orthographic depth, regularity, and grapheme complexity, all of which may play some role in reading and spelling performance (e.g., Schmalz, Beyersmann, Cavalli, & Marinus, 2016). Arguably, in research focusing on the earliest spelling development stages, unconditional phono-graphemic consistency is probably *the* most influential measure. That is, in early spelling attempts, children's letter choices are probably most affected by the context-free probability with which a given sound will be represented by a given letter, weighted by its frequency of occurrence. Importantly, the influence of the context of adjacent sounds and letters in a word, as well as grapheme complexity, begin to exert an influence on English children's spelling by their second year of schooling (e.g., Authors, 2005). The effects of these constructs have been explored in only a handful of other languages to date among beginner spellers (e.g., see Lété, Peereman & Fayol, 2008 for similar developmental patterns in French). Thus, how orthographic consistency is defined and calculated, and whether other orthographic attributes are considered, needs to be determined based on the developmental window and the skill under investigation.

Assessment Issues

The evidence base confirming the critical precursors of alphabetic spelling ability, namely phoneme awareness, letter knowledge, and RAN comprised in the Triple Foundation Model, is well established across many alphabetic orthographies (e.g., Authors, 2012, 2013; Furness & Samuelsson, 2011; Wimmer & Meyringer, 2002). We return to some potential measurement issues regarding these three abilities below.

The measurement of spelling ability itself across alphabetic orthographies is not straightforward. Across orthographies, almost invariably, some spellings deviate from the *alphabetic principle* due to inherent linguistic differences between language families and due to cultural/historical influences on the codification and updating of different orthographies. Thus, deviations from the one-phoneme-to-one-grapheme mapping principle may reflect morphophonological, morphological, and supraphonological processes and etymological artefacts that give rise to graphotactic constraints, exceptions, loan words, and so on (Desrochers, Manolitsis, Gaudreau, & Georgiou, 2018). The extent to which these occur in an orthography determines its system-wide phoneme-grapheme and grapheme-phoneme consistency. In turn, such language-specific inconsistency makes the full "commensurability of measures" across languages (cf. Share, 2008) very challenging. Fortunately, the sources of inconsistency tend to be similar in many alphabetic orthographies. Spelling tasks can thus be created by equating at least certain orthographic dimensions.

Guidelines for further research

By their nature, alphabetic orthographies reflect, to some extent, the phonological structure of words. For this reason, it is generally possible to control for structural features such as word length in terms of syllables and letters, as well as syllable structure. Furthermore, in real-word spelling tests, word frequency can be controlled across languages, if not perfectly equated, as long as appropriate lexical corpora are available. Brysbaert and New (2009) suggest that the optimal corpus size for representative frequency estimates is approximately 16 million words. However, estimates from smaller corpora can also be useful. For example, corpora such as CPWD (British English) (Masterson, Dixon, Stuart, & Lovejoy, 2010), Manulex (French) (Lété, Sprenger-Charolles & Colé, 2004), Weslalex (Czech, Slovak, Polish) (Authors, 2011) or WoC-

GR (Greek, Authors, 2007) contain words extracted from children's school texts and have a quite exhaustive cover of the printed words that children encounter in school. The representativeness of the source material is an important factor in determining corpus quality. Moreover, they provide various standardized frequency indices, such as the estimated usage per million (U), that allow for a comparison across languages of the relative frequency of specific words. Such indices were used in the ELDEL project for the stimuli selection for the silent reading measure of cognate words (Authors, 2013). When such corpora are not available, reasonable proxy measures may be Age of Acquisition (AoA) and/or familiarity norms as these tend to be highly correlated with lexical frequency and can be readily obtained from adult raters (e.g., Morrison, Chappell, & Ellis, 1997). AoA and familiarity estimates are also useful in selecting cognate items.

The creation of cross-linguistically comparable pseudoword measures is relatively more straightforward. The issue of matching on lexical frequency is eliminated, and if the aim is to match the items on phonographemic features such as length and phonological complexity, this is feasible across most alphabetic orthographies. Languages and orthographies vary with respect to the frequency distributions of sublexical phonological or orthographic units (e.g., at the level of syllable, body, rime, phoneme). Suppose the pseudoword spelling task aims to compare children's ability to represent specific phoneme-grapheme correspondences at any particular grain size, with or without regard for contextual constraints. In that case, factors such as frequency of occurrence of the corresponding phonological and/or spelling units of interest must be considered, either by selecting comparably frequent units or controlling for frequency variations. Fortunately, frequency statistics for sublexical units can quite reliably be obtained from corpora of relatively modest size because they represent relatively small inventories of units (e.g., Dockum & Bowern, 2019). Furthermore, direct comparisons on like-for-like

measures can be facilitated if carried out at the lowest common denominator level. For example, if one of the languages does not permit word-final consonant clusters, children's ability to spell these structures cannot be directly compared across all languages.

As a general rule, word-level measures of spelling tend to be adequately reliable across alphabetic orthographies. The greater challenge in creating comparable spelling measures of real words lies in selecting items that are comparable not only in terms of phonological structure and frequency but also in terms of their orthographic complexity. To the extent that acquiring orthographic knowledge about the morphophonological, morphological and etymological bases of phono-graphemically inconsistent spellings might progress universally, it is important to ensure that the spelling items contain similar types of inconsistencies across all orthographies in question. Moreover, the word choices should be educationally and developmentally appropriate. Such competing demands on the constraints of the stimuli need to be balanced, and the word list choices will ultimately be guided by the primary objectives of the test under consideration. For example, the approach followed in the ELDEL project led to selecting spelling items, which were matched across languages for phonological structure, orthographic inconsistency content, grade level, and relative frequency to a reasonably high degree[4].

In creating comparable triple foundation measures across languages, researchers need to decide whether they plan to make direct comparisons of attainment levels or whether they aim to assess the power and patterns of prediction across languages. In the latter case, the constraints on stimulus matching may be more relaxed than in the former. For example, if, on the one hand, the researchers are interested in whether English and Greek children can isolate phonemes to the same level of proficiency by the end of kindergarten, then the English and Greek task versions

[4] For more information, visit www.eldel-mabel.net

need to be quite closely matched at the item level. Although English and Greek may differ in the total number of phonemes in the language, the test should include those target phonemes that occur typically in both languages or share numerous articulatory features (e.g., English /g/ - Greek /γ/). If, on the other hand, the interest is in comparing the importance of the PA construct as a predictor of grade one spelling ability, then it is the reliability and validity of the construct that is important.

Items of tasks tapping PA skills should be comparable across languages for their frequency, familiarity, grammatical class, and so on if using real words (see also Authors, 2012). Using pseudoword items can facilitate the comparability of structures and phoneme classes across languages. Item-level matching should not ignore language-specific phonological constraints using legal constructions and phoneme sequences within a language while selecting sounds that are as closely approximated as the languages in question allow.

Letter knowledge is relatively much easier to measure across languages. In the simplest and most reliable task format, the child says aloud the name and the sound associated with each of the alphabet letters; tasks that involve naming aloud are preferable to forced-choice selection measures as they reduce the likelihood of guessing. Moreover, most languages' alphabets contain a sufficient number of items to ensure high internal consistency reliability.

Finally, RAN measures can also be fairly equated across languages. RAN tasks are typically very stable across versions (naming colours, objects, digits or letters) and across repeated trials in all languages (e.g., Authors, 2012). However, given that RAN tasks are timed, versions with longer words are likely to require more time to complete. In such a case, direct comparisons of naming speed in languages with longer words (e.g., Hungarian, Spanish) with languages with shorter words (e.g., English) require caution. For the alphanumeric versions of

RAN, it is generally possible to select digits and letters that are of comparable length across many (perhaps most) European languages.

Balancing the efforts to achieve cross-linguistic similarity is an equally important consideration in selecting phonological, morphological, and orthographic structures. For example, where languages have fundamentally different syllable structures, fully matching PA measures may not be possible. It must be ensured that the resulting PA measures estimate the same metalinguistic skill and are adequately valid and reliable. It is also important to weigh the importance of including language-specific items (e.g., the use of letters with diacritics in one language) that add to the measure's ecological validity in question.

In short, research on spelling development across languages promises to advance our understanding not of the basic skills of written communication but also to shed light on the universals and specifics of how children build complex lexical representations that encode various sources of inconsistency. While test batteries cannot be identical translations across different languages, especially when real word materials are included in the assessments, researchers can go a very long way to ensure that their measures are highly comparable either in terms of measuring the same constructs or in terms of measuring similar items (for example with the use of cognates), or when possible, both. Numerous research groups have worked on cross-linguistic projects of literacy development, funded by the European Community, over recent decades (e.g., ELDEL: Authors, 2012; NeuroDys: Landerl et al., 2013, Moll et al., 2014; ProRead: Authors, 2010; COST A8: Authors, 2003) and have met with and found solutions for various methodological challenges that are inherent in this type of research.

Reading Comprehension

Target skills and constructs

Text comprehension is the hallmark of literacy and a gateway to learning, problem-solving, and decision-making, both in and out of school. As a result, significant effort has been devoted to developing reading comprehension measures serving evidence-based educational decisions that affect millions of students across linguistic and sociocultural backgrounds. Reading comprehension is a multi-component process underlying the construction of the meaning of an extended, self-contained text that can serve a variety of functions. Although decoding is a prerequisite (Hoover & Gough, 1990; Joshi & Aaron, 2000; Vellutino, Tunmer, Jaccard, & Chen, 2007), comprehension further entails the coordination of linguistic knowledge, such as syntactic awareness and semantic skills, with cognitive and metacognitive processes, such as encoding, integration, and monitoring (Bates, Devescovi, & Wulfeck, 2001; Yeari, 2017). Therefore, comprehension involves applying and coordinating several automatic and strategic processes that transform language into thought.

As readers progress through the text, they access word meanings to form propositions, which, in turn, they connect with previously formed propositions to construct an integrated representation of the content of the text that can support whatever goal motivated the reading in the first place (Kintsch, 1988; Long & Lea, 2005; Perfetti & Stafura, 2014). The representational outcome of comprehension can vary in terms of coherence and elaboration. Mental representations can be fragmented, containing explicitly stated but largely unconnected ideas. At the other end, they can be highly coherent at the local and global level, containing a rich interconnected network of explicit and inferred ideas and elaborations that serve to integrate the text with prior knowledge, to extend its meaning, and to determine its implications (Kintsch 1988; McNamara & Magliano, 2009). Inferencing is crucial for coherence and elaboration. However, the nature, the amount, and the quality of the inferences generated during reading is a

function of the reader's prior knowledge (Kulesz, Francis, Barnes, & Fletcher, 2016; McNamara & Magliano, 2009). The availability and activation of language, discourse, domain, and general world knowledge allow readers to specify the relations that may hold between adjacent and more distant sentences and text parts, contributing, thereby, to local and global coherence, respectively (Best, Floyd, & McNamara, 2008; Deacon & Kiefer, 2018; Denton et al., 2015; Graesser, McNamara, & Louwerse, 2003; Perfetti & Stafura, 2014).

Comprehension assessment attempts to capture representational coherence and elaboration by targeting processes related to domain-general cognitive and metacognitive processes universally applied to make sense of the world (see PIRLS 2021; IEA[5], 2019). This contributes to the relatively seamless communication between studies examining comprehension in different languages (e.g., Blanc & Tapiero, 2001; Garcia, Bustos & Sanchez, 2015; Kaakinen & Hyönä, 2008; Mulder & Sanders, 2012), as well as to the paucity of cross-linguistic comprehension research beyond an international assessment context. Although it is unlikely that language characteristics (e.g., morphology, syntactic structure) alter the processes of encoding, integration and monitoring per se, they may introduce differences in the cognitive resources required for their smooth execution. This possibility remains largely unexplored in the field of comprehension research.

Assessment Issues

National and international standardized measures are typically employed to facilitate comparisons across readers in terms of overall comprehension performance (Martin, Mullis, & Hooper, 2017; McQueen & Mendelovits, 2003; IEA, 2019; RAND Reading Study Group, 2002; Tengberg, 2017). They are likely to involve sentences or short passages followed by multiple-

[5] International Association for the Evaluation of Educational Achievement

choice and/or short-answer questions or sentence verification, picture selection, or cloze tasks. Readers are asked to select or provide an answer, verify the validity of a statement, select the picture that best corresponds to the situation depicted in text or sentence, or choose from a selection the appropriate word that is missing (e.g., Leslie & Caldwell, 2011; MacGinitie, MacGinitie, Maria, & Dreyer, 2000; Wiederholt & Bryant, 1992; Woodcock, McGrew, & Mather, 2001). These formats are relatively time- and resource-efficient and lend themselves well to group administration. However, commonly used standardized tests have also been criticized for targeting lower-level aspects of comprehension and confounding text comprehension with prior knowledge and/or reasoning skill (Basaraba, Yovanoff, Alonzo, & Tindal, 2013; Campbell, 2005; Keenan & Betjemann, 2006).

Disentangling prior knowledge from meaning construction is a tall order and, one could argue, a futile attempt. The knowledge critical for comprehension is indeed also sensitive to demographic and language variations (Snyder, Caccamise, & Wise, 2005). Although passages and items are carefully calibrated for readability, age-appropriateness, and reliability, selections are often based on overarching and culturally driven assumptions regarding the familiarity of topics, vocabulary, and language structures. This is problematic considering that individual performance on a standardized test is interpreted according to normative data obtained at the district, national, or international levels. This potential problem is even more pronounced in the context of international assessment, where linguistic, educational, and cultural aspects can threaten measurement invariance (Asil & Brown, 2016; International Test Commission, 2018; Oliveri & von Davier, 2011). In general, mismatches between assessment and students' knowledge and experiences can significantly reduce the comparability of outcomes and the test's validity (Snyder et al., 2005).

For higher-level comprehension processes to kick in, passages must be long enough to make inferencing and integration possible and necessary. Written text, however, can serve different functions and pose different demands on readers (Authors, 2005; Eason, Goldberg, Young, Geist, & Cutting, 2012; Wu, Barquero, Pickren, Barber, & Cutting, 2020). Expository texts are primarily informational or learning texts, likely to have abstract and unfamiliar content and variable structure (Best et al., 2008). Lack of relevant background knowledge can limit inferencing and increase monitoring demands (Denton et al., 2015). In comparison to narratives, expository texts contribute significantly to item difficulty in comprehension assessment (Eason et al., 2012; Kulesz et al., 2016). Therefore, measures must include both narrative and expository selections.

Similarly, the questions or tasks that follow the reading of passages must be designed to assess the different levels of representational outcomes. Skilled readers identify and represent important text content along with information left implicit in the text, such as interconnections and implications. They may also form evaluations about the relevance and applicability of the content or author intentions (Cain, 2010; van den Broek & Helder, 2017; Wolfe & Goldman, 2005). Many standardized tests strive to assess different levels of comprehension outcomes by including various literal, inferential, and evaluative questions (e.g., IEA, 2019; Williams, Ari, & Santamaria, 2011). Inferential and evaluative questions contribute more to item difficulty than questions targeting vocabulary and memory for content, and that contribution is more pronounced with expository texts (Basaraba et al., 2013; Eason et al., 2012; Kulesz et al., 2016). Although practical and psychometric considerations limit the number and kinds of questions that can follow a test passage, the questions merit deeper examination regarding how they provide an index of at least the local and global coherence level that a reader has achieved.

Guidelines for further research

The complexity of reader-text interactions influencing comprehension and the variability of reading contexts increase the likelihood of assessment reflecting an incomplete picture regarding a reader's comprehension ability and may curtail interpretation of assessment outcomes. As a result, a substantial amount of research has focused on the consistency and comparability of measures and the extent to which they tap theoretically important component processes of comprehension.

Although a psychometrically sound test that includes different types of texts of sufficient length, followed by tasks that target representational coherence and elaboration, might be a good place to start, further probing is needed to determine potential extraneous sources of variability and provide a more fine-grained assessment of component processes (e.g., Cain, 2010). This is crucial in cross-language comprehension assessment, where performance differences can be attributable to additional language-specific and contextual factors (e.g., International Test Commission, 2018; Snyder et al., 2005). Further research should focus more explicitly on cross-linguistic issues to determine any language-specific influences on comprehension processes and outcomes.

In the context of international assessment, the focus has been on carefully developing and calibrating materials, items, and scoring rubrics for equivalence across languages and cultures. However, a lot more work is needed to minimize construct, method, or item bias (inequivalence) that may result from differences in language and in addition to those of culture, background knowledge, sample characteristics, test familiarity, administration conditions, test-taking strategies, and response styles (e.g., International Test Commission, 2018; Rios & Sireci, 2014). Therefore, one could argue that instead of striving for fine-grained equivalence in methods and

items, cross-language assessment should focus more on construct equivalence in conjunction with the item and reading goal comparability to assess comprehension in more culture-appropriate and authentic contexts.

Considering, however, the complexity of the construct and the focus of assessment on the outcomes of comprehension and not the processes per se, further research needs to validate existing assessment methods with online measures. Strategically selected comparisons between online and offline performance can provide important information regarding the accuracy and the degree to which a standardized measure captures the underlying processes. Moreover, cross-language research employing online and offline measures can illuminate the effects of language-specific characteristics on the cognitive and metacognitive processes. For example, research employing reading-time and eye-tracking methodologies can help illuminate if and how differences in syntactic structures influence establishing connections within and across text sentences, that is, integration. Recent international multi-lab collaborations (e.g., Kuperman et al., 2021, MECO[6] Study) make these comparisons possible.

Finally, given the multidimensionality of comprehension and the pervasive influence of background knowledge, assessment targeting primarily the representational coherence achieved by readers reading different texts, for different authentic reasons, would be a practical first-step approach. Assessing the level of coherence, as opposed to elaboration, would be optimal as it requires inferencing and integration of adjacent and more distant text parts while reducing the risk of general knowledge and sentence-level elaboration confounding. Specifying the assessment target at this level ensures that measures tap higher-level cognitive and metacognitive processes while permitting inferences regarding readers' skill in prerequisite lower-level

[6] Multilingual Eye-movement Corpus, https://meco-read.com/

processes. Suboptimal performance at this level would indicate the need for a follow-up, more targeted, and fine-grained assessment to pinpoint specific weaknesses and to rule out potential language-specific or knowledge confounds.

Writing Research Across Languages

Target skills and constructs

Writing sentences and texts is another fundamental component of literacy. It is the result of several interacting cognitive processes (namely, planning, translating and controlling; see, Kellogg, 1996; Authors, 2012) that operate on semantic, linguistic and motor representations in a capacity-limited system (Authors, 2004, 2012) in order to integrate constraints related to language use and linguistic systems[7]. In his Writer(s)-Within-Community model of writing, Graham (2018) emphasizes that writing is shaped by sociocultural influences and contexts that result from contexts related to writers' communities. These contexts contribute to the regulation and control of writing, moderated by writers' emotions, personality traits and physical states. Such interaction between communities of writers, cognitive resources and individual characteristics is, therefore, fundamental to the development and learning of writing. Thus, the text's features result from writers' efficiency in managing the cognitive processes involved in writing and the contextual and linguistic constraints related to the writer's culture and language, according to writers' discourse and vocabulary knowledge (Olinghouse & Graham, 2009)[8]. Therefore, writing assessment should target writing dimensions at both the product and process

[7] For example, rhetorical goals constrain vocabulary choice, and morphological properties of a language (analytic/ isolating or synthetic as agglutinative languages…) determine words characteristics (for a comparison of French and English, see Reilly et al., 2014).

[8] Other types of knowledge are required for composing a text (e.g., conceptual knowledge related to domain of the text). However, the linguistic dimensions of a text are mainly determined by writers' linguistic knowledge and skills (e.g., vocabulary or morphosyntactical knowledge).

levels to connect changes in processes to changes in text features. Two main approaches have been developed for that purpose: The process-centred perspective that assesses the characteristics of the cognitive operations required to plan, translate, review, and transcribe a text, and the product-centred perspective that assesses texts.

The process-centred approach aims at opening a window on the mental processes that create texts. Process-centred studies track the writing processes in real-time, for example, with verbalization techniques and dual tasks (Authors, 2004; Authors, 2002). Also, process-centred analyses rely on the recording of handwriting (e.g., with a tablet with Eye and Pen by Alamargot Chesnet, Dansac, & Ros, 2006; with an electronic pen with Handspy by Alves, Leal, & Limpo, 2018) or on keylogging in the case of typing (e.g., Inputlog by Leijten & van Waes, 2006 and Scriptlog by Wengelin et al., 2009), or of eye movements (Alamargot et al., 2006). Various temporal parameters of writing can then be analyzed, such as pause length and localization in the text, bursts of written language, revision operations. Although process-centred methods inform about the cognitive facet of writing strategies, it remains necessary to relate cognitive functioning to text characteristics to explain better how writing processes contribute to text quality. This relationship between writing processes and text quality has not been systematically studied, and it is therefore difficult to conclude about its nature. There are even fewer real-time studies comparing different languages, which again prevent knowing if and how different language systems create changes in cognitive operations and, thus, in products.

The product-centred approach aims to describe the various dimensions of texts to assess performance level and identify struggling writers (Dockrell & Connelly, in press). At least three text-related skills can be assessed: the pragmatic skills inform how writers adapt their communicative goals to the writing context; the conceptual or semantic skills (e.g., ideas,

structure, or coherence) address how meaning is communicated through texts. These skills can be assessed for different language units, such as a word, clause, sentence, paragraph, or text.

Text assessment involves holistic or analytical approaches and objective or subjective measures (see also Schriver, 1989, who proposed to distinguish between "text-focused" and "reader-focused" evaluations). Analytical and objective measures are crucial for demonstrating how different languages are processed since they focus on texts' specific linguistics characteristics (spelling errors, repetitions, abstract words, passive sentences, or cohesion devices). More generally, product measures at different levels (spelling, sentence structure, connectives) inform on how writers integrate the different facets of a language. However, writing is the result of several complex interactions between writers' skills, task characteristics, and writing contexts. Hence, a considerable challenge is developing indexes or criteria that measure the above skills across different languages. Therefore, text composition measures have to be particularly well-conceived and operationalized to be sensitive to variations between languages.

Assessment Issues

Globally, three main dimensions can be assessed: productivity (or fluency), accuracy and complexity at the word-, sentence-, and discourse- levels (Troia, Shen, & Brandon, 2019). Productivity refers to parameters about the quantity and rhythm of the composition. For example, higher writing rates (such as the number of words per minute or T-units) reflect high cognitive fluency. However, writing rate may be affected by typological features: agglutinative languages presumably result in lower compositional fluency than analytic languages because of longer words (for a comparison between English, Hebrew, Icelandic and Swedish, see Stromquist et al., 2002). Accuracy refers to how writers use the language system (assessed via spelling, lexical and grammar use) by adjusting the clarity of the message delivered to the target language's norms

(Skehan, 1998). Accuracy may be affected by features of the language used. For example, fewer spelling errors are observed in transparent than in opaque languages (for a comparison between Spanish and English, see Llaurado & Dockrell, 2020). Complexity measures assess lexical usage and syntactic structures, generally at the sentence, clause or word levels (or in T-unit such as the number of clauses per T-unit). Research has highlighted the difficulty to find standard complexity measures that can be applied to different languages (e.g., Newmeyer & Preston, 2014).

Natural language processing (NLP) tools, which automatically analyze the features of texts, facilitate text assessment (Paroubek, Chaudiron, & Hirschman, 2007). NLP is a subfield of linguistic, computer science, and artificial intelligence whose aim is to process, understand and extract meanings from natural languages by calculating language measures that index its features (other NLP branches are automatic language generation and comprehension, and translation). Applied to writing, NLP or automated essay scoring tools can provide researchers with standardized scoring rubrics such as cohesion (Coh-Metrix, McNamara et al., 2010; Crossley et al., 2016), lexical sophistication (Kim, Crossley, & Kyle, 2018) or semantic dimensions (e.g., Latent Semantic Analyses; Dumais, 2015), and sentiment analysis (Linguistic Inquiry and Word Count, LIWC; Pennebaker et al., 2001). Objectivity, replicability, and ease of implementation are the main qualities of NLP tools (McNamara, Louwerse, McCarthy, & Graesser, 2010), which can also be used to assess texts written in several languages as soon as libraries are available in the targeted language. For example, LIWC provides about 90 variables about vocabulary, syntax, emotions, punctuation, topics and the like in different languages (e.g., English, German, and French). The Natural Language Toolkit (NLTK), an open-source platform in Python, also provides a suite of text processing libraries for classification, tokenization, stemming, parsing, or

semantic analyses that supports several languages. Although NLP measures are generally not norm-referenced and defining a gold standard for language measures is difficult, NLP tools could contribute to a better understanding of the texts' quality and their relationships with the writing process.

In sum, among the multiple proficiencies required to compose a text, some are more or less prone than others to be directly influenced by the typological properties of a language. For instance, higher-order processes related to decision making, reasoning, or problem-solving are not language-dependent and may be easier to compare across languages. However, because texts and discourses are shaped by the discourse communities in which they occur (Graham, 2018), higher-order cognitive and metalinguistic skills, as well as text features, can be affected by language differences. This is the case of argument structure that differs across cultures (Uysal, 2012): how arguments are linked in argumentative texts differs and involves different sentence structures or connectors. Similarly, regarding text genres, Ragnarsdóttir et al. (2002) showed that cross-linguistic similarities and differences about typological structures of verbs used in five languages allowed for differentiating text genres. Therefore, higher-order writing skills are also dependent, presumably to a lesser extent than lexical and syntactic proficiencies, on the linguistic contexts, such as information structure and the inherent lexical properties of arguments (Duguine, Huidobro, & Madariaga, 2010; Witzlack-Makarevich & Seržant, 2018). In addition, some methods appear more suitable than others for examining variations between languages in writing. Objective and analytical methods, whether carried out by humans or with computers, seem to be better suited for these inter-language comparisons.

Discussion

Research on literacy has become universal. The need to study the acquisition or development of reading, spelling, writing and comprehension skills across alphabetic languages to better understand these phenomena is apparent. This paper has addressed some of the most important methodological challenges that concern the relevant research, suitable for cross-linguistic comparisons. We have shown that several unique challenges for literacy research across alphabetic languages arise for many reasons. This also means that any limitations on the reliability and reproducibility of the data produced in cross-linguistic literacy research may affect the validity of the models and theories and the theories' application in practice.

We have shown that while literacy studies across European languages frequently acknowledge similar methodological issues, the limitations addressed or the standards demanded by those studies are generally overlooked or have not been met yet. Despite the considerable progress in understanding the centrality of cognitive and linguistic skills, only a handful of studies reveal how these skills develop in different readers at various levels of development and across different orthographies (e.g., Authors, 2017; Georgiou et al., 2008, 2020; Authors, 2010). Our review shows that establishing comparability or equivalence at each stage of the research process is a fundamental requirement. Also, the lack of a standardized approach in designing relevant experiments further deepens the replicability crisis in this research area.

To tackle these challenges, we need to determine the degree of cross-linguistic diversity and select analysis units to assess. We also need to explore new methodologies based on constructs that will be carefully defined across languages through preliminary research. We need to aim for the data's maximum comparability by using similar testing conditions and matched samples. Last, at the level of analysis, we need to determine measures that are reflective and formative indicators of the constructs of interest and make proper decisions about needing

conversions of the measurement of units. We have provided some possible solutions in the present paper. To the extent that such a perspective is possible, we have also presented some guidelines for each interest area. With interest shown in neurosciences advances and how the neuroscientific findings are translated into practice (e.g., Papadatou-Pastou, Haliou, & Vlachos, 2017), the article concludes with an appraisal about the significant contribution of neuroscience in managing the challenges of cross-linguistic research.

Looking into the future: Brain research issues and the contribution of neuroscience

Word level reading has been the target of brain research for the last two decades. However, one of the main questions is how many relevant publications have focused on reading development and used a cross-linguistic design. This research field's general approach is to investigate the full-blown reading networks of adults and look for functional patterns and structural characteristics associated with reading. The majority of the studies investigated highly educated college students and found that the functional and structural networks of reading are fairly the same in alphabetic languages and resulted from integrating speech and script as a general process under the specific impact of culture and education. While differences in the integration-related changes in brain correlates are easy to compare even without using a cross-linguistic design, proper neuroscientific answers to the impact of culture and education are impossible without cross-linguistic investigations. However, our understanding of the impact of multifactorial changes in the literate brain remains opaque as long as no results of large-scale longitudinal studies on the same or reliably similar factors are available. Moreover, sensitive measures like the neuroscientific ones do not give comparable results without having a clear view of literacy development's crucial factors. The prime prerequisite of neuroscientific investigations is the high replicability of data in literacy development across languages, as in

behavioral studies. Likewise, the reliability of conclusions and probability of models addressing both language-universal and specific aspects are important.

We should be aware that neuroscientific methods provide correlative data whose predictive value is limited. Therefore, before we ask how useful brain data could be used to answer whether a process is the same or different in two or more languages, we should be aware that neuroscience is not just about looking for brain areas. Therefore, an important step toward understanding literacy development would be if systematic studies were used and a cross-linguistic design was applied to answer simple and more development-related questions. Developmental cross-linguistic investigations using brain methods for comparing the core factors' expression during reading development are scarce. This is valid for investigating the phonological awareness changes on the syllabic and phonemic levels. Moreover, there are no cross-linguistic neuroscientific studies on several issues, even not on simple matters. For example, phoneme level awareness is seen as the consequence and not the prerequisite of reading (for review, see Deheane, Cohen, Morais, & Kolinsky, 2015). Very few neuroscientific studies aimed at looking for how the developing brain changes in this respect. It is well known that the illiterates' spelling is underdeveloped, and their brain is different from those who are literate (Petersson, Silva, Castro-Caldas, Ingvar, & Reis, 2007), but further systematic studies are not available. It is a sound though not obvious question to ask how the developing brain performs letter-speech sound (LSS) and script-speech integrations and what the consequences are on reading performance. As Blomert (2011) suggested, learning to read in alphabetic orthographies started with learning a script code consisting of LSS pairs. Typically developing children learn the LSS associations within months, though it takes considerably longer to process them as new audio-visual objects, constructed via integration automatically. We have limited answers to how

the first grade's reading requirements shape the brain and to what extent this newly emerged system is used for reading words of another language. A recent cross-linguistic study by Leppanen et al. (2019), involving Finnish, French, Hungarian and German cohorts, showed that an electrophysiological response (MMN) indexing speech and non-speech sound discrimination was extremely reproducible and supported the view that auditory change detection is the same in typically developing schoolchildren irrespectively of their native language. However, the limited generalizability of the atypical brain responses found in dyslexia across language environments does not question the usefulness of neuroscientific methods per se. It rather questions the common neurobiological factors of dyslexia.

Recently, the fast development of script expertise is studied in several laboratories. Authors (2020) show that tuning for print is a fast process and emerges as early as the first school year. Moreover, tuning for familiar letter strings can be enhanced by the parallel presentation of print and sound, suggesting an important role of orthographic-phonological mapping in print awareness development. However, there are no neuroscientific cross-linguistic studies exploring the development of orthography and its relationship to phonology development. Therefore, the question that arises is what cross-linguistic neuroscience studies should look for? What should they focus on if we wished to contribute to developing new theoretical models, instruction, or intervention? The topics are different, but the methodological challenges are the same. Therefore, there is a need for change in methods, design, size and scale of systematic investigations to advance, in turn, researchers' methodological awareness. Here are some suggestions to enhance the cross-linguistic research through neuroscience: (1) Pay attention to one dimension as a start and decompose it according to the languages studied, focusing on procedural equivalence; (2) Combine behavioural and neuroscientific methods to

investigate processes assumed to show subtle differences in various languages; (3) Design studies with careful stimulus selection, strict, broadly accepted protocol for comparing reading-related processes; (4) Perform large-scale, longitudinal cross-linguistic studies with a focus on reading development; (5) Move from word-level investigations to sentences and texts; (6) Develop studies for investigating natural reading by using parallel measures; and (7) Develop ecologically valid designs and move from laboratory to classroom where the technical conditions allow group measurements in a school setup.

Conclusions

We have attempted to record various methodological challenges in literacy research across languages. Until we improve our methodology, the challenges will always be present. However, cross-linguistic research will continue to move on, requiring greater methodological rigour from researchers in the field. The methodological issues we have presented here are just some of the challenges of current or future research, and we do recognize that we have not examined them all in this paper. For instance, factors such as the growth and degree of longitudinal stability across language skills (e.g., Georgiou et al., 2020), the potential importance of individual differences (e.g., Pugh & Verhoeven, 2018), or the appropriateness of different methods (e.g., see the Reading-Level match design for research in dyslexia, Authors, 2020) could make the list of methodological challenges grow long. As we discussed, future research should explore several other issues, perhaps with the contribution of neuroscience. This new direction needs to be systematic, examining methodological developments in literacy research to provide additional insights and a comprehensive and informed review. Even more, it needs to investigate the growth and the degree of longitudinal stability of various factors determining the development of literacy across languages.

References

Authors (1993).

Authors (2002).

Authors (2003).

Authors (2003).

Authors (2004).

Authors (2005).

Authors (2005).

Authors (2007).

Authors (2008).

Authors (2009).

Authors (2010).

Authors (2010).

Authors (2010).

Authors (2012).

Authors (2012).

Authors (2012).

Authors (2013).

Authors (2013).

Authors (2015).

Authors (2016).

Authors (2017).

Authors (2018).

Authors (2019).

Authors (2019).

Authors (2019).

Authors (2019).

Authors (2020).

Authors (2020).

Authors (in press).

Authors (in press).

Alamargot, D., Chesnet, D., Dansac, C., & Ros, C. (2006). Eye and pen: A new device for studying reading during writing. *Behavior Research Methods, Instruments, & Computers, 38*, 287-299.

Alves, R., Limpo T., & Joshi R. (Eds.) (2020). *Reading-writing connections: Towards integrative literacy science*. Cham, Switzerland: Springer.

Alves, R. A., Leal, J., & Limpo, T. (2019). Using Handspy to study writing in real-time: A comparison between low- and high-quality texts in Grade 2. In R. Fidalgo & T. Olive (Series Eds.), *Studies in Writing* (pp. 50-70). Leiden, NL: Brill.

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders* (DSM-5®). Washington, DC: American Psychiatric Association.

Anthony, J. L., & Lonigan, C. J. (2004). The nature of phonological awareness: Converging evidence from four studies of preschool and early grade school children. *Journal of Educational Psychology, 96*, 43-55.

Asil, M., & Brown, G. T. (2016). Comparing OECD PISA reading in English to other languages: Identifying potential sources of non-invariance. *International Journal of Testing, 16*, 71-93.

Basaraba, D., Yovanoff, P., Alonzo, J., & Tindal, G. (2013). Examining the structure of reading comprehension: Do literal, inferential, and evaluative comprehension truly exist? *Reading & Writing, 26,* 349-379.

Bates, E., Devescovi, A., & Wulfeck, B. (2001). Psycholinguistics: A cross-language perspective. *Annual Review of Psychology, 52*, 369-396.

Berninger, V. W., Swanson, H. L., & Griffin, W. (2015). Understanding developmental and learning disabilities within functional-systems frameworks: Building on the contributions of J. P. Das. In T. C. Papadopoulos, R. K. Parrila, & J. R. Kirby (Eds.), *Cognition, intelligence, and achievement* (pp. 397-418). San Diego, CA: Academic Press.

Berman, R., & Verhoeven, L. (2002). Cross-linguistic perspectives on the development of text-production abilities: Speech and writing. *Written Language & Literacy, 5*, 1-43.

Best, R. M., Floyd, R. G., & McNamara, D. S. (2008). Differential competencies contributing to children's comprehension of narrative and expository texts. *Reading Psychology*, *29*, 137-164.

Blomert, L. (2011). The neural signature of orthographic-phonological binding in successful and failing reading development. *Neuroimage,* 57, 695-703.

Borleffs, E., Maassen, B. A., Lyytinen, H., & Zwarts, F. (2017). Measuring orthographic transparency and morphological-syllabic complexity in alphabetic orthographies: a narrative review. *Reading and Writing, 30*, 1617-1638.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*, 977-990.

Cain, K. (2010). Reading for meaning: The skills that drive comprehension development. In K. Hall, U. Goswami, C. Harrison, S. Ellis, & J. Soler (Eds.). *Interdisciplinary perspectives on learning to read: Culture, cognition and pedagogy* (pp. 74-86). Routledge.

Campbell, J. R. (2005). Single instrument, multiple measures: Considering the use of multiple item formats to assess reading comprehension. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 347-368). Mahwah, NJ: Lawrence Erlbaum Associates.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review, 108*, 204-256.

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. *The Journal of Second Language Writing, 32,* 1-16.

Deacon, S. S., & Kiefer, M. (2018). Understanding how syntactic awareness contributes to reading comprehension: Evidence from mediation and longitudinal models. *Journal of Educational Psychology, 110,* 72-86.

Deheane, S., Cohen, L., Morais, J. & Kolinsky, R. (2015) Illiterate to literate: Behavioural and cerebral changes induced by reading acquisition. *Nature Reviews*, *16,* 234-244.

Denton, C. A., Enos, M., York, M. J., Francis, D. J., Barnes, M. A., Kulesz, P. A., Fletcher, J. M., & Carter, S. (2015). Text-processing differences in adolescent adequate and poor

comprehenders reading accessible and challenging narrative and informational text. *Reading Research Quarterly, 50,* 393-416.

Desrochers, A., Manolitsis, G., Gaudreau, P., & Georgiou, G. (2018). Early contribution of morphological awareness to literacy skills across languages varying in orthographic consistency. *Reading and Writing, 31*, 1695-1719.

Dockrell, J. E., & Connelly, V. (in press). Capturing the challenges in assessing writing: development and writing dimensions. In T. Limpo & T. Olive (Eds.), *Executive functions and writing*. Oxford University Press.

Dockrell, J. E., Papadopoulos, T. C., Mifsud, C. L., Bourke, L., Vilageliu, O.,…Gerdzhikova, N. (in press). Teaching and learning in a multilingual Europe: Findings from a cross-European study. *European Journal of Psychology of Education*.

Duguine, M., Huidobro, S., & Madariaga, N. (Eds) (2010). *Argument structure and syntactic relations: A cross-linguistic perspective*. Amsterdam, The Netherlands: John Benjamins Publishers.

Dumais, S. T. (2005). Latent semantic analysis. *Annual Review of Information Science and Technology, 38*, 188-230.

Durgunoğlu, A.Y., & Öney, B. A (1999). Cross-linguistic comparison of phonological awareness and word recognition. *Reading and Writing 11*, 281-299.

Eason, S. H., Goldberg, L. F., Young, K. M., Geist, M. C., & Cutting, L. E. (2012). Reader-text interactions: How differential text and question types influence cognitive skills needed for reading comprehension. *Journal of Educational Psychology, 104*, 515-528.

Fitzgerald, J., & Shanahan, T. (2000). Reading and writing relations and their development. *Educational Psychologist, 35*, 39-50.

Furnes, B., & Samuelsson, S. (2011). Phonological awareness and rapid automatized naming predicting early development in reading and spelling: Results from a cross-linguistic longitudinal study. *Learning and Individual Differences, 21*, 85-95.

Garcia, J. R., Bustos, A., & Sanchez, E. (2015). The contribution of knowledge about anaphors, organizational signals and refutations to reading comprehension. *Journal of Research in Reading, 38,* 405-427.

Georgiou, G. K., Das, J. P., & Hayward, D. (2009). Revisiting the 'simple view of reading' in a group of children with poor reading comprehension. *Journal of Learning Disabilities, 42*, 76-84.

Georgiou, G. K., Parrila, R. K., & Papadopoulos, T. C. (2008). Predictors of word decoding and reading fluency across languages varying in orthographic consistency. *Journal of Educational Psychology, 100*, 566-580.

Georgiou, G. K., Torppa, M., Landerl, K., Desrochers, A., Manolitsis, G., de Jong, P. F., & Parrila, R. (2020). Reading and spelling development across languages varying in orthographic consistency: Do their paths cross? *Child Development, 91*, e266-e279.

Georgiou, G. K., Torppa, M., Manolitsis, G., Lyytinen, H., & Parrila, R. (2012). Longitudinal predictors of reading and spelling across languages varying in orthographic consistency. *Reading and Writing, 25*, 321-346.

Graesser, A. C., McNamara, D. S., & Louwerse, M. M. (2003). What do readers need to learn in order to process coherence relations in narrative and expository text? In A. P. Sweet & C. E. Snow (Eds.), *Rethinking reading comprehension* (pp. 82-98). New York: Guilford Press.

Graham, S. (2018). A revised writer(s)-within-community model of writing. *Educational Psychologist, 53*, 258-279.

Harm, M. W., & Seidenberg, M. S. (1999). Phonology, reading acquisition, and dyslexia: Insights from connectionist models. *Psychological Review, 106*, 491-528.

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing, 2,* 127-160.

IEA (2019). *PIRLS 2021 Assessment frameworks*. Chestnut Hill, MA: Boston College.

International Literacy Association. (2018). *Reading fluently does not mean reading fast* [Literacy leadership brief]. Newark, DE: Author.

International Test Commission (2018). *ITC Guidelines for Large-Scale Assessment of Linguistically and Culturally Diverse Populations* [www.InTestCom.org]

Joshi, R. M., & Aaron, P. G. (2000). The component model of reading: A simple view of reading made a little more complex. *Reading Psychology, 21,* 85-97.

Katz, L., & Frost, R. (1992). The reading process is different for different orthographies: The orthographic depth hypothesis. In R. Frost & L. Katz (Eds.), *Advances in psychology*: *Orthography, phonology, morphology, and meaning* (pp. 67-84). North-Holland: Elsevier.

Katzir, T., Shaul, S., Breznitz, Z., & Wolf, M. (2004). The universal and the unique in dyslexia: A cross-linguistic investigation of reading and reading fluency in Hebrew-and English-speaking children with reading disorders. *Reading and Writing, 17*, 739-768.

Keenan, J. M., & Betjemann, R. S. (2006). Comprehending the gray oral reading test without reading it: Why comprehension tests should not include passage-independent questions. *Scientific Studies of Reading, 10,* 363-380.

Kellogg, R. T. (1996). A model of working memory in writing. In M. Levy & S. E. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 57-71). Mahwah, NJ: Lawrence Erlbaum Associates.

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*, 120-141.

Kim, Y. G. (2020). Interactive dynamic literacy model: An interactive theoretical framework for reading-writing relations. In R. Alves, T. Limpo, & M. Joshi (Eds), *Reading-writing connections* (pp. 11-34). Cham, Switzerland: Springer.

Kintsch, W. (1988). The use of knowledge in discourse processing: A construction-integration model. *Psychological Review, 95*, 163-182.

Kirby, J. R., Georgiou, G. K., Martinussen, R., & Parrila, R. (2010). Naming speed and reading: From prediction to instruction. *Reading Research Quarterly, 45*, 341-362.

Kulesz, P. A., Francis, D. J., Barnes, M. A., & Fletcher, J. M. (2016). The influence of properties of the test and their interaction with reader characteristics on reading comprehension: An explanatory item response study. *Journal of Educational Psychology, 108*, 1078-1097.

Kuo, L. J., & Anderson, R. C. (2006). Morphological awareness and learning to read: A cross-language perspective. *Educational Psychologist, 41*, 161-180.

Landerl, K. (2019). Behavioral precursors of developmental dyslexia. In L. Verhoeven, & C. Perfetti (Eds.), *Developmental dyslexia across languages and writing systems* (pp. 229-252). Cambridge, UK: Cambridge University Press.

Landerl, K., Freudenthaler, H. H., Heene, M., de Jong, P. F., Desrochers, A., Manolitsis, G., . . . Georgiou, G. K. (2019). Phonological awareness and rapid automatized naming as

longitudinal predictors of reading in five alphabetic orthographies with varying degrees of consistency. *Scientific Studies of Reading, 23*, 220-234.

Landerl, K., Ramus, F., Moll, K., Lyytinen, H., Leppänen, P. H., Lohvansuu, K., ... & Schulte-Körne, G. (2013). Predictors of developmental dyslexia in European orthographies with varying complexity. *Journal of Child Psychology and Psychiatry, 54*, 686-694.

Leijten, M., & Van Waes, L. (2006). Inputlog: New perspectives on the logging of on-line writing. In K. P. H. Sullivan & E. Lindgren (Eds.), *Computer keystroke logging and writing: Methods and applications* (pp. 73-94). Oxford, UK: Elsevier.

Leslie, L., & Caldwell, J. S. (2011). *Qualitative Reading Inventory-5*. Boston, MA: Pearson Education.

Lété, B., Peereman, R., & Fayol, M. (2008). Consistency and word-frequency effects on spelling among first-to fifth-grade French children: A regression-based study. *Journal of Memory and Language, 58*, 952-977.

Llaurado, A., & Dockrell, J. (2020). The impact of orthography on text production in three languages: Catalan, English, and Spanish. *Frontiers in Psychology, 11*, 878.

Long, D. L., & Lea, R. B. (2005). Have we been searching for meaning in all the wrong places: Defining the "search after meaning" principle in comprehension. *Discourse Processes,* 39, 279-298.

MacGinitie, W. H., MacGinitie, R. K., Maria, K., & Dreyer, L. (2000). *Gates-MacGinitie Reading Tests* (4th ed.). Chicago, IL: Riverside Publishing.

Manolitsis, G., Georgiou, G., Stephenson, K., & Parrila, R. (2009). Beginning to read across languages varying in orthographic consistency: Comparing the effects of non-cognitive and cognitive predictors. *Learning and Instruction, 19*, 466-480.

Martin, M. O., Mullis, I. V. S., & Hooper, M. (Eds.). (2017). *Methods and Procedures in PIRLS 2016.* Retrieved from Boston College, TIMSS & PIRLS International Study Center: https://timssandpirls.bc.edu/publications/pirls/ 2016-methods.html

McBride, C. (2016). *Children's literacy development: A cross-cultural perspective on learning to read and write*. New York, NY: Routledge.

McClung, N. A., & Pearson, P. D. (2019). Reading comprehension across languages seven European orthographies and two international literacy assessments. *Written Language and Literacy, 22*, 33-66.

McNamara, D. S., & Magliano, J. (2009). Toward a comprehensive model of comprehension. In B. Ross (Ed.), *The psychology of learning and motivation* (Vol. 51, pp. 297-384). Burlington: Academic Press.

McNamara, D. S., Louwerse, M. M., McCarthy, P. M., & Graesser, A. C. (2010). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes*, *47*, 292-330.

McQueen, J., & Mendelovits, J. (2003). PISA reading: Cultural equivalence in a cross-cultural study. *Language Testing, 20,* 208-224.

Milfont, T. L., & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of Psychological Research, 3*, 111-130.

Moll, K., Ramus, F., Bartling, J., Bruder, J., Kunze, S., Neuhoff, N., . . . Landerl, K. (2014). Cognitive mechanisms underlying reading and spelling development in five European orthographies. *Learning and Instruction, 29*, 65-77.

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*. Washington, DC: National Institute of Child Health and Human Development.

Newmeyer, F. J., & Preston, L. B. (Eds.) (2014). *Measuring grammatical complexity*. Oxford, UK: Oxford University Press.

Olinghouse, N. G., & Graham, S. (2009). The relationship between the discourse knowledge and the writing performance of elementary-grade students. *Journal of Educational Psychology, 101*, 37-50.

Oliveri, M. E., & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Journal of Psychological Testing and Assessment Modeling, 53,* 315-333.

Papadatou-Pastou, M., Haliou, E., & Vlachos, F. (2017). Brain knowledge and the prevalence of neuromyths among prospective teachers in Greece. *Frontiers in Psychology, 8*, 804.

Paroubek, P., Chaudiron, S., & Hirschman, L. (2007). Principles of evaluation in natural language processing. *Traitement Automatique Du Langage, 48*, 7-31.

Parrila, R., Aunola, K., Leskinen, E., Nurmi, J. E., & Kirby, J. R. (2005). Development of individual differences in reading: Results from longitudinal studies in English and Finnish. *Journal of Educational Psychology, 97*, 299.

Patel, T. K., Snowling, M. J., & de Jong, P. F. (2004). A cross-linguistic comparison of children learning to read in English and Dutch. *Journal of Educational Psychology, 96*, 785.

Perfetti, C. & Harris, L. (2017). Learning to read English. In L. Verhoeven, & C. Perfetti (Eds.), *Learning to read across languages and writing systems* (pp. 347-370). Cambridge, UK: Cambridge University Press.

Perfetti, C., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading, 18,* 22-37.

Peterson, R. L., & Pennington, B. F. (2012). Developmental dyslexia. *The Lancet, 379*, 1997-2007.

Petersson, K. M., Silva, C., Castro-Caldas, A., Ingvar, M., & Reis, A. (2007). Literacy: A cultural influence on functional left-right differences in the inferior parietal cortex. *European Journal of Neuroscience, 26*, 791-799.

Pugh, K., & Verhoeven, L. (2018). Introduction to this special issue: Dyslexia across languages and writing systems. *Scientific Studies of Reading, 22*, 1-6.

Ragnarsdóttir, H., Aparici, M., Cahana-Amitay, D., van Hell, J. & Viguié, A. (2002). Verbal structure and content in written discourse: Expository and narrative texts. *Written Language and Literacy, 5*, 95-126.

RAND Reading Study Group (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Washington, DC: RAND Education.

Rau, A. K., Moll, K., Moeller, K., Huber, S., Snowling, M. J., & Landerl, K. (2016). Same same, but different: Word and sentence reading in German and English. *Scientific Studies of Reading, 20*, 203-219.

Reilly, J., Bernicot, J., Olive, T., Uzé, J., Wulfeck, B., Favart, M., & Appelbaum, M. (2014). Written narratives from French and English-speaking children with language impairment. In B. Arfé, J. Dockrell, & V. W. Berninger, (Eds.). *Writing development and instruction*

in children with hearing, speech and oral language difficulties (pp. 176-187). Oxford, UK: Oxford University Press.

Rios, J. A., & Sireci, S. G. (2014). Guidelines versus practices in cross-lingual assessment: A disconcerting disconnect. *International Journal of Testing, 14,* 289-312.

Schmalz, X., Beyersmann, E., Cavalli, E., & Marinus, E. (2016). Unpredictability and complexity of print-to-speech correspondences increase reliance on lexical processes: More evidence for the orthographic depth hypothesis. *Journal of Cognitive Psychology, 28*, 658-672.

Schmalz, X., Marinus, E., Coltheart, M., & Castles, A. (2015). Getting to the bottom of orthographic depth. *Psychonomic Bulletin and Review, 22*, 1614-1629.

Seidenberg, M. S. (2011). Reading in different writing systems: One architecture, multiple solutions. In P. McCardle, B. Miller, J. R. Lee, & O. J. L. Tzeng (Eds.), *Dyslexia across languages: Orthography and the brain-gene-behavior link* (p. 146–168). Baltimore, MD: Paul H Brookes Publishing.

Share, D. L. (2008). On the anglocentricities of current reading research and practice: The perils of overreliance on an "outlier" orthography. *Psychological Bulletin, 134*, 584-615.

Smith, A. C., Monaghan, P., & Huettig, F. (2021). The effect of orthographic systems on the developing reading system: Typological and computational analyses. *Psychological Review, 128*, 125.

Snyder, L., Caccamise, D., & Wise, B. (2005). The assessment of reading comprehension: Considerations and cautions. *Topics in Language Disorders, 25,* 33-50.

Soodla, P., Torppa, M., Kikas, E., Lerkkanen, M. K., & Nurmi, J. E. (2019). Reading comprehension from grade 1 to 6 in two shallow orthographies: Comparison of Estonian

and Finnish students. *Compare: A Journal of Comparative and International Education, 49*, 681-699.

Spencer, L. H., & Hanley, J. R. (2003). Effects of orthographic transparency on reading and phoneme awareness in children learning to read in Wales. *British Journal of Psychology, 94*, 1-28.

Stromquist, S., Johansson, V., Kriz, S., Ragnardóttir, R., Aiseman, R., & Ravid, D. (2002). Toward a cross-linguistic comparison of lexical quanta in speech and writing. *Written Language & Literacy 5*, 45-67.

Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement, 27*, 159-203.

Tengberg, M. (2017). National reading tests in Denmark, Norway, and Sweden: A comparison of construct definitions, cognitive targets, and response formats. *Language Testing, 34,* 83-100.

Treiman, R., & Kessler, B. (2014). *How children learn to write words*. Oxford, UK: Oxford University Press.

Troia, G., Shen, M., & Brandon, D. (2019). Multidimensional levels of language writing measures in grades four to six. *Written Communication, 36*, 231-266.

Uysal, H. H. (2012). Argumentation across L1 and L2 writing: Exploring cultural influences and transfer issues. *Vigo International Journal of Applied Linguistics, 9*, 133-159.

van Daal, V. H., & Wass, M. (2017). First-and second-language learnability explained by orthographic depth and orthographic learning: A 'natural' Scandinavian experiment. *Scientific Studies of Reading, 21*, 46-59.

van den Broek, P., & Helder, A. (2017). Cognitive processes in discourse comprehension: Passive processes, reader-initiated processes, and evolving mental representations. *Discourse Processes, 54,* 360-372.

Vellutino, F. R., Tunmer, W. E., Jaccard, J. J., & Chen, R. (2007). Components of reading ability: Multivariate evidence for a convergent skills model of reading development. *Scientific Studies of Reading, 11,* 3-32.

Verhoeven, L. & Perfetti, C. (2017). Introduction: Operating principles in learning to read. In L. Verhoeven, & C. Perfetti (Eds.), *Learning to read across languages and writing systems* (pp. 1-30). Cambridge, UK: Cambridge University Press.

Verhoeven, L., & Perfetti, C. (Eds.) (2017). *Learning to read across languages and writing systems*. Cambridge, UK: Cambridge University Press.

Wengelin, A., Torrance, M., Holmqvist, K., Simpson, S., Galbraith, D., Johansson, V., & Johansson, R. (2009). Combined eye tracking and keystroke-logging methods for studying cognitive processes in text production. *Behaviour Research Methods*, *41*, 337-351.

Wiederholt, L., & Bryant, B. (1992). *Examiner's manual: Gray Oral Reading Test-3*. Austin, TX: Pro-Ed.

Williams, R. S., Ari, O., & Santamaria, C. N. (2011). Measuring college students' reading comprehension ability using cloze tests. *Journal of Research in Reading, 34,* 215-231.

Witzlack-Makarevich, A. & Seržant, I. A. (2018). Differential argument marking: Patterns of variation. In A. Witzlack-Makarevich & I. A. Seržant (Eds.), *Diachrony of differential argument marking* (pp. 1-40). Berlin, Germany: Language Science Press.

Wolfe, M. B., & Goldman, S. R. (2005). Relations between adolescents' text processing and reasoning. *Cognition and Instruction, 23,* 467-502.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III tests of achievement*. Itaska, IL: Riverside.

Wu, Y., Barquero, L. A., Pickren, S. E., Barber, A. T., & Cutting, L. E. (2020). The relationship between cognitive skills and reading comprehension of narrative and expository texts: A longitudinal study from Grade 1 to Grade 4. *Learning and Individual Differences, 80*.

Yeari, M. (2017). The role of working memory in inference generation during reading comprehension: Retention, (re)activation, or suppression of verbal information? *Learning and Individual Differences, 56,* 1-12.

Ziegler, J. C., Bertrand, D., Tóth, D., Csépe, V., Reis, A., Faísca, L., ... & Blomert, L. (2010). Orthographic depth and its impact on universal predictors of reading: A cross-language investigation. *Psychological Science, 21*, 551-559.

Ziegler, J. C., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: A psycholinguistic grain size theory. *Psychological Bulletin, 131*, 30-29.

Ziegler, J. C., Perry, C., Ma-Wyatt, A., Ladner, D., & Schulte-Körne, G. (2003). Developmental dyslexia in different languages: Language-specific or universal? *Journal of Experimental Child Psychology*, *86*, 169-193.

Zorzi, M. (2010). The connectionist dual process (CDP) approach to modelling reading aloud. *European Journal of Cognitive Psychology, 22*, 836-860.