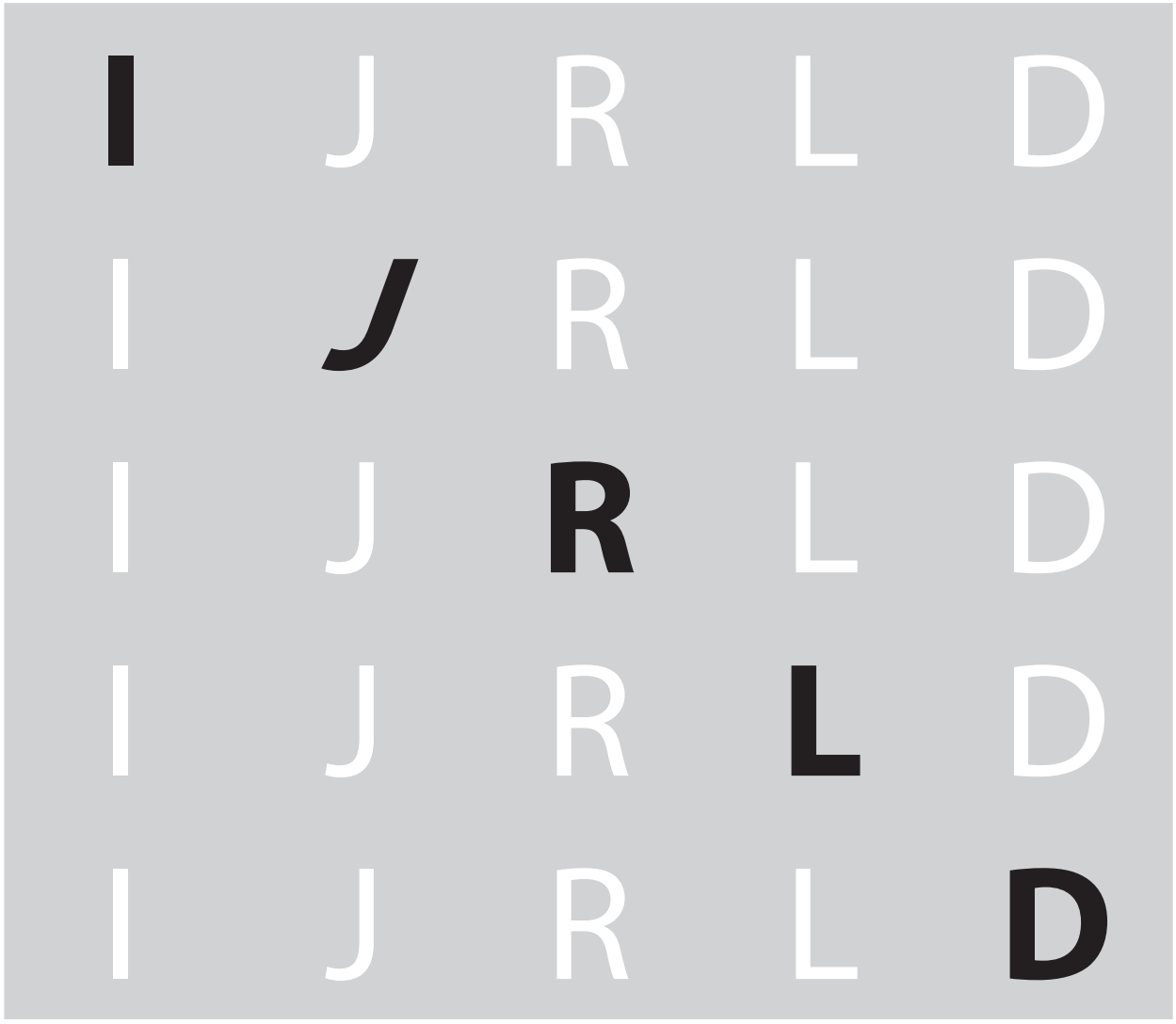
CBM language measures as indicators of foreign-language learning: technical adequacy of scores for secondary-school students

Hoefnagel, L.H.; Espin, C. A.; Rippe, R.C.A.

International Journal for Research in Learning Disabilities

volume 5, issue 1, 2021

formerly *Thalamus*



Journal of the
International Academy for Research in Learning Disabilities

Editor: David Scanlon, Boston College, United States

International Academy for Research in Learning Disabilities

International **J**ournal for **R**esearch in **L**earning **D**isabilities
formerly *Thalamus*
journal of the *International Academy for Research in Learning Disabilities*

David Scanlon, Editor, Boston College, USA
Kirsten McBride, Copy Editor, Kansas City, Missouri, USA

IJRLD
International Journal for Research in Learning Disabilities

Volume 5, No. 1, 2021

Table of Contents

Guest Reviewers

Faye Antoniou, *National and Kapodistrian University of Athens*
Kevin Chung, *The Education University of Hong Kong*
Christine Espin, *Leiden University*
Jennifer Krawec, *University of Miami*
Rebecca Louick, *St. John's University*
Michael Paal, *Carl von Ossietzky Universität Oldenburg*

The "Write Stuff": What Do We Know About Developmental Dysgraphia?

Catherine McBride and Zebedee Rui En Cheah
The Chinese University of Hong Kong

Editor's Note: Catherine McBride had been selected by the Academy to deliver the 2020 Cruickshank Memorial Lecture. That lecture was canceled along with the annual conference.

Abstract

As researchers come to recognize the origins of dysgraphia, we can better suggest optimal approaches to remediation. In defining dysgraphia, we review the writing process, research on the development of writing, and various factors related to either spelling difficulties, visual-motor difficulties, or both, that might interfere in the process of writing. We conclude by exploring some potentially helpful remediation techniques that should be considered as educators, clinicians, researchers, teachers, and parents work together to ameliorate the potentially devastating consequences of dysgraphia.

Keywords: Writing, dyslexia, multiscriptalism, motor skills, visual-spatial skills, spelling, dictation, copying, handwriting, dysgraphia remediation

My (e.g., McBride, 2019) interest in dysgraphia developed gradually. I have taken lessons in Chinese on and off over the years. Each time I do, I try to write the Chinese characters assigned to me clearly and neatly, and every time I fail. My writing of Chinese looks out of proportion and often unclear, writing that might have been done by a child in kindergarten or first grade. I am further compelled to mention that I received the grade of C in handwriting in second grade (age 7). This was lower than most of the other grades I ever received, and I remember feeling shame at this evaluation. At the same time, however, I was quite a good speller, so handwriting and spelling for me were not conflated.

Given my interest in cross-cultural literacy, I came to the research topic of dysgraphia relatively late. What particularly piqued my interest was a former student who did a project on Chinese dysgraphia (see McBride, 2019). The project was well researched, and the student's personal story was even more compelling: He had always had great difficulties in writing but never had problems in reading, in either Chinese or English. Before I got to know him and his story, I had always assumed that dysgraphia was primarily a by-product of the much better understood phenomenon of dyslexia. Now I realize that the phenomenon of dysgraphia is more complicated.

A focus on handwriting per se can help us to identify children with specific learning disabilities, such as dyslexia and dysgraphia. Although those with dyslexia are typically characterized as manifesting pronounced difficulties in spelling and word reading (Lyon et al., 2003), they also have distinctive handwriting characteristics. That is, children with dyslexia often manifest slow and poor-quality handwriting (Gosse & Van Reybroeck, 2020). Compared to those without dyslexia, Chinese children with dyslexia write significantly more slowly, with lower accuracy, greater character size, and more size variability (Lam et al., 2011), whereas children with dyslexia in alphabetic scripts tend to have greater spelling error rates attributable mainly to their impairments in phonology (Sterling et al., 1998). Given these writing-related correlates of dyslexia, it is important to consider more precisely the nature of dysgraphia. What is it, and how can we separate dysgraphia from dyslexia?

McBride, C., & Rui En Cheah, Z. (2021). The "Write Stuff": What Do We Know About Developmental Dysgraphia? *International Journal for Research in Learning Disabilities, 5*(1), 3-12. https://doi.org/10.28987/ijrld.5.1.3

Editor's Note: Due to an editing error, the citation associate with this article should correctly read:
McBride, C., & Cheah, Z. R. E. (2021). The "Write Stuff": What Do We Know About Developmental Dysgraphia? International Journal for Research in Learning Disabilities, 5(1), 3-12. https://doi.org/10.28987/ijrld.5.1.3

In this article, we begin by defining dysgraphia. Conceptualizations of dysgraphia are perhaps even more confusing and variable than are concepts of dyslexia. Nevertheless, it is crucial to settle on a consistent definition as well as to understand how dysgraphia is diagnosed across cultures. We then consider the writing process more generally. In order to understand writing difficulties, we must first grasp how writing progresses in a typically developing child or, indeed, in an adult learning to write in a new script. In reviewing the writing process, we particularly highlight the motoric and visual-orthographic skills required for writing across scripts. Our discussion concludes with a review of some approaches to remediation techniques for helping those with dysgraphia.

Defining Dysgraphia

According to the 5th edition of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-5), dysgraphia is categorized as a specific learning disorder (American Psychiatric Association, 2013), with an estimated prevalence of around 7-15% (Katusic et al., 2009). Writing difficulties may be associated not only with dysgraphia but also with other disorders such as developmental coordination disorder (Biotteau et al., 2019), dyslexia (Gosse & Van Reybroeck, 2020), autism (Mayes et al., 2019), and attention deficit hyperactive disorder (Rosenblum et al., 2008). Dysgraphia can appear alone, but it can also co-occur with other developmental disorders (Chung & Patel, 2015). Therefore, to ensure accurate identification, it is imperative for a test used to diagnose dysgraphia to be valid and reliable.

What is the definition of dysgraphia? These are two of the definitions that have been offered in an attempt to describe this difficulty precisely: Hamstra-Bletz and Blöte (1993) defined dysgraphia as a disturbance in the production of written language in relation to the mechanics of writing. In contrast, Chung et al. (2020) defined dysgraphia in a more comprehensive way, describing dysgraphia as a "disorder of writing ability at any stage, including problems with letter formation/legibility, letter spacing, spelling, fine motor coordination, rate of writing, grammar, and composition" (p. S46).

Children have been estimated to spend around 31-60% of their school day performing tasks involving handwriting and fine-motor tasks (Feder & Majnemer, 2007). Although attitudes about handwriting may be changing given the prevalence of computers and cell phones, good handwriting skills remain important (e.g., Askvik et al., 2020; Kiefer et al., 2015). Poor writing and speaking development are related to various negative outcomes, including academic difficulties, as well as social-emotional and behavioral problems (Grigorenko, 2007). As a result, early identification is crucial for children with dysgraphia.

Deuel's (1995) classification of dysgraphia into three categories is potentially useful for a precise diagnosis of dysgraphia (McBride, 2019). This classification focuses on various abilities that contribute to the writing process as a means of identifying the specific difficulties of each individual who exhibits symptoms of dysgraphia. For example, Deuel (1995) focused on four tasks that are used to test for dysgraphia in children; these are oral spelling skills, copying skills, drawing skills, and finger-tapping speed. By focusing on strengths and weaknesses across these skills, one can potentially distinguish across three types of dysgraphia, namely, dyslexic dysgraphia, spatial dysgraphia, and motor dysgraphia.

The three subtypes of dysgraphia distinguished by Deuel (1995) are defined based on strengths and weaknesses of subskills as follows. *Dyslexic dysgraphia* implies struggles related to writing that are caused by corresponding difficulties with spelling. Insecure spellers may write sub-optimally because of confusion around how to represent symbols. However, these children do not have spatial or motor difficulties per se. Rather, their main difficulties have to do with spelling or dictation; such children have difficulties in spelling orally, but their spatial and motor skills are largely intact. In contrast, those with *spatial dysgraphia* have difficulties in the production of writing in both spontaneous and copied written text (Deuel, 1995). Although these children have problems in representing text as well as two-dimensional drawings or other symbols, they do not struggle with motor movements; importantly, they do not have difficulties in spelling when asked to do so orally. Finally, those with *motor dysgraphia*, in addition to manifesting illegible writing in both spontaneous and copied contexts, demonstrate abnormal drawing and handwriting velocity (Deuel, 1995).

A group of researchers has distinguished movements involved in handwriting apart from cognitive-linguistic skills related to spelling or dictation itself. For example, analyzing variables from 42 studies on handwriting movements in relation to dysgraphia, Danna et al. (2013) described the variables as falling into the three main categories of temporal, kinematic, and dynamical.

There is some support for each of the three subtypes of dysgraphia considered by Deuel (1995). For example, some researchers have focused on the overlap between dysgraphia and developmental dyslexia (Döh-

la et al., 2018). Döhla and Heim (2016) presented a review showcasing the similarity in underlying cognitive skills across dyslexia and dysgraphia; these include, for example, a phonological awareness deficit and an automatization deficit (Nicolson & Fawcett, 2011). However, some research has found that a central distinction between those with dyslexic dysgraphia and those with dyslexia only is that, typically, students with dyslexic dysgraphia tend to read at grade level, whereas those with dyslexia do not (Brown, 2019). Another distinction between dyslexia and dysgraphia involves differences in language mapping processes. Specifically, Berninger (2008) suggested that dysgraphia is a consequence of an inefficient mapping process involving verbal memory only in the direction of phonological to orthographic, whereas dyslexia results from inefficiency in mapping in both directions, namely, from orthographic to phonological and from phonological to orthographic.

In contrast, spatial dysgraphia is presumably caused primarily by visual-spatial difficulties, which, in turn, contribute to the prevalence of handwriting difficulties (Tal-Saban & Weintraub, 2019). Hécaen and Albert (1978, as cited in Rode et al., 2006) defined spatial dysgraphia according to four main features: (a) right-page preference, with writing often crowded onto the right side of the page; (b) inclination, particularly a failure to produce oblique lines and to write horizontally; (c) broken lines (i.e., leaving unusually large spaces between words leading to fragmentation of lines into small segments); and (d) graphic errors, including incorrect productions of strokes of given letters or characters. Presumably, these features could be generalized to various writing systems, such as Arabic or Persian, in which writing takes place from right to left (see McBride & Mohseni, 2020). Individuals with visual-spatial dysgraphia likely do not manifest oral spelling difficulties, but show distortions in how they copy or draw symbols or pictures.

Finally, dysgraphia is also sometimes associated with poor motor control (Smits-Engelsman & Van Galen, 1997). Smits-Engelsman and Van Galen (1997) found that handwriting movement with dysgraphia produces greater "noise" and that often poor writing is related to failure to obey spatial constraints, resulting in a lack of consistency in handwriting. On the other hand, Nicolson and Fawcett (2011) suggested that dysgraphia may reflect a lack of automaticity at the cognitive level, including an impairment in the cerebellar-motor circuit. Further, adopting a test of tapping ability, Ben-Pazi et al. (2007) demonstrated that some children with poor handwriting quality manifested dysrhythmia.

Other tests of finger tapping or finger succession (touching every finger to the thumb) can distinguish some children with writing difficulties, presumably because these children have specific difficulties with motor control (Berninger et al., 2006; Deuel, 1995). There is still a lot to be learned about dysgraphia across scripts. Deuel's (1995) classification of forms of dysgraphia is helpful, but in practice more research studies on different aspects of dysgraphia are needed.

Some aspects of dysgraphia must also be considered in relation to the individual script. In particular, writing in Chinese is deemed more difficult than writing in English, given that Chinese requires greater visual discrimination of the fine differences in the forms and positions of strokes (Lam et al., 2011). McBride (2016) concluded that learning to both read and write in Chinese demands greater visual skills than in alphabetic orthographies. In addition, the writing styles in alphabetic and non-alphabetic writing systems differ. Alphabetic languages stress the importance of smoothness and continuity in writing (Rosenblum et al., 2003), whereas writing in Chinese often involves sharp turns of strokes and more pen lifts (Tseng, 1998). The differences in both the nature of the scripts and the writing styles in Chinese highlight the importance of research on different scripts in order to shed light on additional issues related to dysgraphia.

Testing for Dysgraphia

Despite a relative lack of consensus on the nature of dysgraphia worldwide, some fairly popular tests of handwriting have been used to diagnose dysgraphia. For example, the Concise Assessment Method for Children's Handwriting (BHK; Hamstra-Bletz et al., 1987), a test of dysgraphia in Latin-alphabet-based writing (Asselborn et al., 2018), can be used to evaluate both quality and speed of writing (Van Waelvelder et al., 2012). The BHK consists of 13 criteria that are believed to provide a detailed analysis of the handwriting profiles of children who are either at risk for reading difficulties or who actually have dysgraphia (Overvelde & Hulstijn, 2011). Another test for dysgraphia in alphabetic writing is the Children's Handwriting Evaluation Scale-Manuscript (CHES-M; Phelps & Stempel, 1988). Handwriting characteristics scored on this test include letter form, spacing, rhythm, and general appearance (Feder & Majnemer, 2003). However, the validity and reliability of the CHES-M have been questioned (Van Waelvelder et al., 2012).

In Chinese, a few diagnostic tests of Chinese handwriting have also been developed, including the Chinese Handwriting Analysis System (CHAS; Li-

Tsang et al., 2013) and Tseng's Handwriting Speed Test (Tseng & Hsueh, 1997). Tests of visual-motor integration (e.g., Beery-Buktenica Developmental Test of Visual-Motor Integration; Beery et al., 1997) may also be used for the diagnosis of dysgraphia (Chung & Patel, 2015). Visual-motor integration skills are related to writing across the different scripts of English (Chung & Patel, 2015), Korean (Lee et al., 2019), and Chinese (Chung et al., 2018).

What Does the Writing Process Entail?

Having reviewed some important diagnostic tools that are sometimes used to evaluate writing difficulties, we must take a step back and consider the skills that are required for writing. Writing is a fundamental component of literacy. Children and adults who struggle to acquire writing skills face multiple impediments to their daily lives in activities such as note taking or providing signatures on documents (McCloskey & Rapp, 2017). In the context of dysgraphia, writing involves both visual-motor skills (i.e., the physical capacity to write) and sufficient orthographic/spelling/dictation knowledge to produce a given word in a given script. These two broad elements of the motoric and visual-orthographic aspects of word writing development and impairment can be considered somewhat separately.

In the research of reading and writing, different models of word writing have been proposed; some of these even extend to writing composition (Bereiter & Scardamalia, 1987; Hayes & Flower, 1980; Kellogg, 1996). One mode of the word writing process comes from McCloskey and Rapp (2017).

In this spelling-to-dictation model (McCloskey & Rapp, 2017), both long-term phonological and orthographic memory are strongly emphasized. For example, when we hear or think of a word (e.g., *buy*), this activates a phonological-grapheme representation in our phonological long-term memory (e.g., /baɪ/), in turn activating our lexical-semantic representation to help us to understand the context of the word (e.g., *buy a snack*). Finally, we retrieve the spelling of the selected word (e.g., *buy*, but not *by* or *bye*) from our orthographic long-term memory (McCloskey & Rapp, 2017). After the retrieval of the abstract letter/grapheme representation in our orthographic memory, both motor planning and production processes are required to produce writing (McCloskey & Rapp, 2017). For example, the abstract letter representation is first converted into an appropriate form of allographs (defined here as variants of a grapheme, such as "A" vs. "a"). Next, the allographic representation activates our graphic motor plans to enact the writing process, including where to begin on the page as well as the direction and movements of the pen (McCloskey & Rapp, 2017). Finally, our motor system executes the graphic motor plans we made and writes the word that we had in mind.

Thus, handwriting is a multi-componential task. It includes perceptual, attentional, linguistic, and motor skills (Asselborn et al., 2018). The production of written words involves both the central and peripheral processes. The central process is responsible for the cognitive processes of retrieving, assembling, and selecting the orthographic representation from the orthographic memory whereas the peripheral process is responsible for the generation of motor actions to produce writings (Delattre et al., 2006; Purcell et al., 2011). The process of motor memory is unique to the entire writing operation (McCloskey & Rapp, 2017).

Poor handwriting sometimes reflects deficits in the central process. Kandel et al. (2017) argued that there is an interaction between the central and peripheral processes such that spelling modulates motor production in children's writing. These researchers suggested that the retrieval of the lexical orthographic representation during spelling continues during the production of handwriting, thus affecting the handwriting process (Kandel et al., 2017). A poor memory for either orthographic representation or motor movements can lead to handwriting deficiencies (Kandel et al., 2017; McCloskey & Rapp, 2017). Importantly, in their study of Chinese handwriting, Zhang and Feng (2017) also found that the central processes of handwriting affect the actual execution of handwriting.

Complexity in Writing

What about the visual- orthographic characteristics involved in writing? Orthographic complexity in an alphabetic script refers to the complications of spelling words when their written representations deviate from the basic one-to-one phoneme-grapheme correspondence (Arfé et al., 2020). The orthography-phonology mapping varies across languages. For example, a transparent one-to-one mapping occurs in Finnish, but a more abstract mapping is found in English (Wang et al., 2009). To elaborate, the less transparent mapping of the phoneme /k/ in the English language might include different spellings such as c in *cat*, ch in *character*, ck in *check*, and k in *kick* (Wang et al., 2009). The inconsistency across orthographies causes differences not only in reading development across languages (Ziegler & Goswami, 2005) but also in writing development. For example, in Chinese, orthographic complexity relates to the number of strokes, number of radicals, and the

spatial composition of the radicals (Wang et al., 2020). When the number of strokes in Chinese characters increases, a person's handwriting is more error-prone and slower at accessing orthographic codes (Wang et al., 2020). Furthermore, individuals require less time to access orthographic codes and hand-write left-right characters than characters with other compositions, such as those that are top-down (Wang et al., 2020) due to the familiarity effect (e.g., 75% of Chinese semantic radicals appear on the left side of the character; Feldman & Siok, 1997).

In conceptualizing dysgraphia in the context of the writing process, we must keep in mind that there are differences in visual complexity and discriminability of visual forms of graphs across distinctive writing systems and that such differences can affect the perceptual learning of grapheme forms (Chang et al., 2018). Grapheme complexity is strongly associated with both learning time and learning difficulty – more time is needed to learn an orthography with higher grapheme complexity (Chang et al., 2016).

Chang et al. (2018) devised a grapheme complexity measure to capture the differences in visual complexity across writing systems. This measure includes four components, namely, perimetric complexity, number of disconnected components, number of connected points, and number of simple features. The authors (2018) compiled an ordering of grapheme complexity across 131 languages, with traditional Chinese script highlighted as the most complex written language, and abjads and alphabets showing equally lower complexity levels. Furthermore, the visual complexity of scripts has an effect on the perceptual load, suggesting that the increasing visual complexity of scripts may increase processing difficulty (alphasyllabry: Rao et al., 2011; abjad: Abdelhadi et al., 2011). Hence, the acquisition of writing skills may be affected by the visual complexity of a given writing system.

The importance of visual-motor skills is particularly highlighted in the acquisition of Chinese. Chinese writing acquisition usually relies heavily on drill-and-practice of writing or copying of each Chinese character over and over (Wu et al., 1999). Chinese learners are required to learn to copy Chinese characters in the correct stroke orders (Wang & McBride, 2017). Moreover, they must rely on fine-grained visual discrimination of the forms and positions of strokes in learning to write in Chinese (Lam et al., 2011). Copying, a primarily visual-motor integrated skill, has been shown to explain unique variance in Chinese word writing skills (Wang et al., 2014), whereas poor visual-motor integration is one of the prominent issues faced by Chinese children with slow handwriting (Tseng & Chow, 2000).

Handwriting consolidates memorization of graphemes (alphabetic: Longcamp et al., 2005; morphosyllabic: Guan et al., 2011; alphasyllabary: Bhide, 2018). Indeed, even novel copying skill is sometimes associated with reading (e.g., McBride-Chang et al., 2011) and word writing (e.g., McBride-Chang et al., 2011; Wang et al., 2014). However, the extent to which copying skills facilitate orthographic learning may be restricted to novice learners (Naka & Naoi, 1995; Vaughn et al., 1992). Given all that we understand about the writing process in the context of dysgraphia and more generally, how can we help children with dysgraphia?

Dysgraphia Remediation

As mentioned in relation to learning difficulties more generally (McBride, 2019), two overarching remediation strategies should be considered for those with dysgraphia; namely, work around and work through. "Work around" strategies are ways to deal with the problem and accomplish tasks and assignments despite it. Such strategies focus on how an individual with dysgraphia can produce good work in a given domain (e.g., handwriting an essay during an exam within a specified time period) using alternative techniques. The other type of remediation, sometimes referred to as "work through" strategies, consists of techniques by which individuals with dysgraphia can work on their difficulties by focusing on skills related to them. Here, the focus is on skills development, including motor and visual skills, that directly contribute to the process of handwriting.

"Work Around" Strategies

To start with, problems with handwriting can be remediated with the use of assistive technology. Allowing individuals with dysgraphia to present their work in an alternative medium to handwriting can help to free up their cognitive resources to better focus on higher-order skills in writing assignments (McBride, 2019). The speech-to-text function helps to convert the user's speech, produced orally, into text outputs with the usage of voice recognition software (Thiel et al., 2015). The use of speech-to-text technology is particularly beneficial for improving the quality of content, vocabulary, and syntax of the text, as well as contributing to longer and more complex texts overall (Thiel et al., 2015).

Typing using a keyboard is also generally accepted as an alternative form of written communication for children with dysgraphia in the eyes of occupational therapists (Penso, 1990). In survey research, Freeman and colleagues (2004) found that 93% of the

443 occupational therapists in the Canadian Association of Occupational Therapists who responded reported that they frequently recommended typing on a keyboard as a "work around" alternative for clients with dysgraphia. Similarly, Cochran-Smith (1991) noted several advantages of word processing compared to handwriting, including an increase in content quality and quantity, an increase in legibility, and more error-free texts as compared to written work.

Typing and handwriting essentially require distinctly different skills; indeed, there is typically a low-to-moderate correlation between these two writing forms (Rogers & Case-Smith, 2002). Rogers and Case-Smith (2002) found that students who write slowly or with poor legibility demonstrate an increase in both quantity and legibility of text when they adopt typing as their written form of communication. The impact of differing Chinese inputs in relation to typing remediation of Chinese character is relatively unclear, but typing is possible for Chinese as well, particularly using a Pinyin (phonological coding, primarily using Roman alphabet letters) system. Thus, both text-to-speech and typing are simple alternatives to handwriting, and are the clearest "work around" strategies for managing dysgraphia at school or at work.

"Work Through" Strategies

One potentially exciting approach to remediating handwriting difficulties incorporates neurofeedback. Neurofeedback, or EEG biofeedback, is viewed as a potentially useful, though relatively underdeveloped, treatment for several conditions, ranging from developmental disorders to mental illness to problems with physical balance (e.g., Hammond, 2007). The idea behind neurofeedback it to provide real-time audio and visual feedback about brain waves in order to retrain abnormal brainwave patterns to produce healthier patterns through operant conditioning (Hammond, 2007).

This technique has been used in an attempt to alleviate handwriting difficulties (Harandi & Moghadam, 2017; Walker, 2012). With only 5-10 neurofeedback training sessions, Walker (2012) succeeded in normalizing some abnormalities in cortical areas that are significant for handwriting in individuals with dysgraphia. In addition, these individuals' handwriting was also judged to have improved. To date, relatively few studies have been conducted on the utility of the neurofeedback for ameliorating dysgraphia. However, this technique may be worth integrating into future training studies for those with writing difficulties.

Further, a multisensory approach linking aspects of visual, auditory, and kinesthetic-tactile skills (Abdulkarim et al., 2017) has been recommended as a general remedy to handwriting problems in children (Amundson, 1992). The interaction of different modalities is believed to help students with dysgraphia to recognize cues that are provided by many different sensory channels in order to facilitate learning (Tafti & Abdolrahmani, 2014). Examples of multisensory modalities and activities used in remediation of handwriting problems include "sky writing" letters in the air, finger writing using finger painting, and finger writing in sand or rice (Woodward & Swinth, 2002). A multisensory approach has proved to be beneficial for students with dysgraphia; for example, researchers have documented some improvement in writing performances, writing expression, and spelling, as well as reduced social-emotional problems in students with dysgraphia (Abdulkarim et al., 2017).

At a general level, children with dysgraphia may also require early therapy in basic processes related to writing (for a review, see McBride, 2019). For example, some children benefit from exercises intended to strengthen their hands and fingers or focused on improving fine-motor movements. In addition, it is important that children establish a handedness preference. Children should be encouraged to favor one hand over the other for holding a pen to write; such dominance indicates a specialization of one hemisphere of the brain over the other in writing activities. Forcing children who are naturally left-handed to write with their right hand can cause difficulties (for a review, see McBride, 2019).

Another potentially useful focus for children with specific writing difficulties involves coordination between the hands to ensure optimal bilateral integration. For example, if one hand easily writes and the other easily holds the paper in place to facilitate writing, that collective, coordinated process makes the handwriting process easier. Finally, simply getting children interested in the writing implement, whether it is a pencil, pen, or marker, can be helpful for those with dysgraphia. Thus, one boy's interest in a beautiful pen his mother gave him paved the way for his renewed practice and ultimate mastery of writing (McBride, 2019).

Perhaps the act of copying graphemes itself can additionally help in ameliorating handwriting problems. As children with dysgraphia may purposely avoid writing tasks given the frustration such tasks can cause (Rahim & Jamaludin, 2019), they may lack general practice in handwriting. The act of handwriting practice through repeated direct and delayed copying of letters and words can lead to a more automatic graphomotor control in handwriting (Beeson, 2004). Furthermore, copying tasks can be used to improve both spelling production and handwriting. For example, the copy and recall treatment (CART) paradigm requires individuals to copy a target

word repeatedly and then try to recall the spelling in a written picture-naming format (Beeson et al., 2003). CART has been found to facilitate spelling, which, in turn, helped to ameliorate dysgraphia (Beeson et al., 2003). Interestingly, individuals with dysgraphia may show differences in impairment between writing styles. For example, some are impaired in print writing but not in their ability to write in cursive (Hanley & Peters, 1996; Ingles et al., 2014). Research has also focused on the idea of using cursive writing as an intervention strategy for dysgraphia (Indira & Vijayan, 2015; Nalpon & Chia, 2009). While teaching cursive writing to children appears to improve handwriting skills (Indira & Vijayan, 2015), it does not improve reading and spelling performance in children with dysgraphia (Nalpon & Chia, 2009).

Conclusion

This has been an overview of an important but under-studied learning difficulty. It is critical to understand the writing process in its entirety in order to establish what specific difficulties might interfere with that process, causing dysgraphia. Dysgraphia is, after all, extreme difficulty in the normal but very complicated process of writing. As technology progresses and children engage in more typing and less handwriting, dysgraphia may become less consequential and devastating for those who have it. However, handwriting remains important at least in some domains, and understanding this difficulty is helpful for teachers and parents worldwide.

Given the different scripts that are used globally and the varying demands of each, dysgraphia is a particular learning problem that may glean critical understanding from cross-scriptal, cross-cultural comparisons. In addition, from our observations and experience, we view the topic of dysgraphia as incorporating many disciplines of study. Those in computer science, neuroscience, education, psychology, and occupational therapy, among others, all contribute important understanding of dysgraphia We look forward to critical research globally on this topic in the years to come.

Acknowledgments

This research was supported by a General Research Fund of the Hong Kong Special Administrative Region Research Grants Council (14600818) to Catherine McBride.

References

Abdelhadi, S., Ibrahim, R., & Eviatar, Z. (2011). Perceptual load in the reading of Arabic: Effects of orthographic visual complexity on detection. *Writing Systems Research*, *3*(2), 117-127.

Abdulkarim, W. F., Abdulrauf, M. S., & Elgendy, A. A. (2017). The effect of a multi-sensory program on reducing dyspraxia and dysgraphia among learning disabled students in Rafha. *Journal of Educational Sciences and Psychology, Ii*(lxix), 1.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Author. https://doi.org/10.1176/appi.books.9780890425596

Amundson, S. J. (1992). Handwriting: Evaluation and intervention in school settings. In J. Case-Smith & C. Pehoski (Eds.). *Development of hand skills in the child* (pp. 63-78). American Occupational Therapy Association.

Arfé, B., Corato, F., Pizzocaro, E., & Merella, A. (2020). The effects of script and orthographic complexity on the handwriting and spelling performance of children with dyslexia. *Journal of Learning Disabilities*, *53*(2), 96-108. https://doi.org/10.1177/0022219419892845

Askvik, E. O., Van der Weel, F. R., & Van der Meer, A. L. (2020). The importance of cursive handwriting over typewriting for learning in the classroom: A high-density EEG study of 12-year-old children and young adults. *Frontiers in Psychology*, *11*, 1810.

Asselborn, T., Gargot, T., Kidziński, L., Johal, W., Cohen, D., Jolly, C., & Dillenbourg, P. (2018). Automated human-level diagnosis of dysgraphia using a consumer tablet. *NPJ Digital Medicine*, *1*(1), 1-9.

Beery, K. E., Buktenica, N. A., & Beery, N. A. (1997). *The Beery-Buktenica developmental test of visual-motor integration: VMI, with supplemental developmental tests of visual perception and motor coordination: administration, scoring and teaching manual*. Modern Curriculum Press.

Beeson, P. M. (2004). Remediation of written language. *Topics in Stroke Rehabilitation*, *11*(1), 37-48. https://doi.org/10.1310/D4AM-XY9Y-QDFT-YUR0

Beeson, P. M., Rising, K., & Volk, J. (2003). Writing treatment for severe aphasia. *Journal of Speech, Language, and Hearing Research*, *46*, 1038-1060. https://doi.org/10.1044/1092-4388(2003/083)

Ben-Pazi, H., Kukke, S., & Sanger, T. D. (2007). Poor penmanship in children correlates with abnormal rhythmic tapping: A broad functional temporal impairment. *Journal of Child Neurology*, *22*(5), 543-549.

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Erlbaum. https://doi.org/10.1016/B978-0-08-088583-4.50006-6

Berninger, V. W. (2008). Defining and differentiating dysgraphia, dyslexia, and language learning disability within a working memory model. In M. Mody & E. R. Silliman

(Eds.), *Brain, behavior, and learning in language and reading disorders* (pp. 103-134). Guilford Press.

Berninger, V. W., Abbott, R. D., Jones, J., Wolf, B. J., Gould, L., Anderson-Youngstrom, M., ... & Apel, K. (2006). Early development of language by hand: Composing, reading, listening, and speaking connections; three letter-writing modes; and fast mapping in spelling. *Developmental Neuropsychology*, *29*(1), 61-92. https://doi.org/10.1207/s15326942dn2901_5

Bhide, A. (2018). Copying helps novice learners build orthographic knowledge: Methods for teaching Devanagari Akshara. *Reading and Writing*, *31*(1), 1-33. Inhttps://doi.org/10.1007/s11145-017-9767-8

Biotteau, M., Danna, J., Baudou, É., Puyjarinet, F., Velay, J. L., Albaret, J. M., & Chaix, Y. (2019). Developmental coordination disorder and dysgraphia: signs and symptoms, diagnosis, and rehabilitation. *Neuropsychiatric Disease and Treatment*, *15*, 1873-1885. https://doi.org/10.2147/NDT.S120514

Brown, M. (2019). *Dysgraphia*. Southeastern University FireScholars.https://firescholars.seu.edu/cgi/viewcontent.cgi?article=1001&context=ccplus

Chang, L.Y., Chen, Y. C., & Perfetti, C. A. (2018). GraphCom: A multidimensional measure of graphic complexity applied to 131 written languages. *Behavior Research Methods*, *50*(1), 427-449.

Chang, L. Y., Plaut, D. C., & Perfetti, C. A. (2016). Visual complexity in orthographic learning: Modeling learning across writing system variations. *Scientific Studies of Reading*, *20*(1), 64-85.

Chung, K.K.H., Lam, C. B., & Cheung, K. C. (2018). Visuomotor integration and executive functioning are uniquely linked to Chinese word reading and writing in kindergarten children. *Reading and Writing*, *31*(1), 155-171. https://doi.org/10.1007/s11145-017-9779-4

Chung, P., & Patel, D. R. (2015). Dysgraphia. *International Journal of Child and Adolescent Health*, *8*(1), 27-36.

Chung, P. J., Patel, D. R., & Nizami, I. (2020). Disorder of written expression and dysgraphia: Definition, diagnosis, and management. *Translational Pediatrics*, *9*(Suppl 1), S46-S54.

Cochran-Smith, M. (1991). Word processing and writing in elementary classrooms: A critical review of related literature. *Review of Educational Research*, *61*(1), 107-155. https://doi.org/10.3102/00346543061001107

Danna, J., Paz-Villagrán, V., & Velay, J. L. (2013). Signal-to-noise velocity peaks difference:A new method for evaluating the handwriting movement fluency in children with dysgraphia. *Research in Developmental Disabilities*, *34*(12), 4375-4384.

Delattre, M., Bonin, P., & Barry, C. (2006). Written spelling to dictation: Sound-to-spelling regularity affects both writing latencies and durations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(6), 1330-1340.

Deuel, R. K. (1995). Developmental dysgraphia and motor skills disorders. *Journal of Child Neurology*, *10*(suppl 1), S6-S8. https://doi.org/10.1177/08830738950100S103

Döhla, D., & Heim, S. (2016). Developmental dyslexia and dysgraphia: What can we learn from the one about the other? *Frontiers in Psychology*, *6*, 2045.

Döhla, D., Willmes, K., & Heim, S. (2018). Cognitive profiles of developmental dysgraphia. *Frontiers in Psychology*, *9*, 2006. https://doi.org/10.3389/fpsyg.2018.02006

Feder, K. P., & Majnemer, A. (2003). Children's handwriting evaluation tools and their psychometric properties. *Physical & Occupational Therapy in Pediatrics*, *23*(3), 65-84. https://doi.org/10.1080/J006v23n03_05

Feder, K. P., & Majnemer, A. (2007). Handwriting development, competency, and intervention. *Developmental Medicine & Child Neurology*, *49*(4), 312-317. https://doi.org/10.1111/j.1469-8749.2007.00312.x

Feldman, L. B., & Siok, W. W. (1997). The role of component function in visual recognition of Chinese characters. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(3), 776-781. https://doi.org/10.1037/0278-7393.23.3.776

Freeman, A. R., MacKinnon, J. R., & Miller, L. T. (2004). Assistive technology and handwriting problems: What do occupational therapists recommend? *Canadian Journal of Occupational Therapy*, *71*(3), 150-160. https://doi.org/10.1177/000841740407100305

Gosse, C., & Van Reybroeck, M. (2020). Do children with dyslexia present a handwriting deficit? Impact of word orthographic and graphic complexity on handwriting and spelling performance. *Research in Developmental Disabilities*, *97*, 103553.

Grigorenko, E. L. (2007). Rethinking disorders of spoken and written language: Generating workable hypotheses. *Journal of Developmental & Behavioral Pediatrics*, *28*(6), 478-486. https://doi.org/10.1097/DBP.0b013e31811ff895

Guan, C. Q., Liu, Y., Chan, D. H. L., Ye, F., & Perfetti, C. A. (2011). Writing strengthens orthography and alphabetic-coding strengthens phonology in learning to read Chinese. *Journal of Educational Psychology*, *103*(3), 509-522. https://doi.org/10.1037/a0023730

Hammond, D. C. (2007). What is neurofeedback?. *Journal of Neurotherapy*, *10*(4), 25-36. https://doi.org/10.1300/J184v10n04_04

Hamstra-Bletz, L., & Blöte, A. W. (1993). A longitudinal study on dysgraphic handwriting in primary school. *Journal of Learning Disabilities*, *26*(10), 689-699. https://doi.org/10.1177/002221949302601007

Hamstra-Bletz, L., DeBie, J., & Den Brinker, B. P. L. M. (1987). *Concise evaluation scale for children's handwriting*. Swets & Weitlinger

Hanley, J. R., & Peters, S. (1996). A dissociation between the ability to print and write cursively in lower-case let-

ters. *Cortex*, *32*(4), 737-745. https://doi.org/10.1016/S0010-9452(96)80043-8

Harandi, V., & Moghadam, N. K. (2017). A comparison of the effectiveness of neurofeedback (NFB) training method and Fernald's multisensory approach on dictation performance among students suffering from dictation disorder (dysgraphia). *Focus on Medical Sciences Journal*, *3*(2), 1-8. https://doi.org/10.21859/focsci-03021421

Hayes, J. R., & Flower, L. S. (1980). Identifying the organization of writing processes. In L. W. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing* (pp. 3-30). Erlbaum.

Hécaen, H., & Albert, M. L. (1978). *Human neuropsychology*. John Wiley & Sons Inc.

Indira, A., & Vijayan, P. (2015). Teaching cursive hand writing as an intervention strategy for high school children with dysgraphia. *International Academic Journal of Social Sciences, 2*(12), 1-10.

Ingles, J. L., Fisk, J. D., Fleetwood, I., Burrell, S., & Darvesh, S. (2014). Peripheral dysgraphia: Dissociations of lowercase from uppercase letters and of print from cursive writing. *Cognitive and Behavioral Neurology*, *27*(1), 31-47.

Kandel, S., Lassus-Sangosse, D., Grosjacques, G., & Perret, C. (2017). The impact of developmental dyslexia and dysgraphia on movement production during word writing. *Cognitive Neuropsychology*, *34*(3-4), 219-251. https://doi.org/10.1080/02643294.2017.1389706

Katusic, S. K., Colligan, R. C., Weaver, A. L., & Barbaresi, W. J. (2009). The forgotten learning disability: Epidemiology of written-language disorder in a population-based birth cohort (1976–1982), Rochester, Minnesota. *Pediatrics*, *123*(5), 1306-1313. https://doi.org/10.1542/peds.2008-2098

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

Kiefer, M., Schuler, S., Mayer, C., Trumpp, N. M., Hille, K., & Sachse, S. (2015). Handwriting or typewriting? The influence of pen- or keyboard-based writing training on reading and writing performance in preschool children. *Advances in Cognitive Psychology*, *11*(4), 136-146. https://doi.org/10.5709/acp-0178-7

Lam, S. S., Au, R. K., Leung, H. W., & Li-Tsang, C. W. (2011). Chinese handwriting performance of primary school children with dyslexia. *Research in Developmental Disabilities*, *32*(5), 1745-1756. https://doi.org/10.1016/j.ridd.2011.03.001

Lee, C. H., Kim, E. B., Lee, O., & Kim, E.Y. (2019). Development of the Korean Handwriting Assessment for Children Using Digital Image Processing. *TIIS*, *13*(8), 4241-4254.

Li-Tsang, C. W., Wong, A. S., Leung, H. W., Cheng, J. S., Chiu, B. H., Linda, F. L., & Chung, R. C. (2013). Validation of the Chinese Handwriting Analysis System (CHAS) for primary school students in Hong Kong. *Research in Developmental Disabilities*, *34*(9), 2872-2883. https://doi.org/10.1016/j.ridd.2013.05.048

Longcamp, M., Zerbato-Poudou, M. T., & Velay, J. L. (2005). The influence of writing practice on letter recognition in preschool children: A comparison between handwriting and typing. *Acta Psychologica*, *119*(1), 67-79. https://doi.org/10.1016/j.actpsy.2004.10.019

Lyon, G. R., Shaywitz, S. E., & Shaywitz, B. A. (2003). A definition of dyslexia. *Annals of Dyslexia*, *53*(1), 1-14. https://doi.org/10.1007/s11881-003-0001-9

Mayes, S. D., Breaux, R. P., Calhoun, S. L., & Frye, S. S. (2019). High prevalence of dysgraphia in elementary through high school students with ADHD and autism. *Journal of Attention Disorders*, *23*(8), 787-796. https://doi.org/10.1177/1087054717720721

McBride, C. A. (2016). Is Chinese special? Four aspects of Chinese literacy acquisition that might distinguish learning Chinese from learning alphabetic orthographies. *Educational Psychology Review*, *28*(3), 523-549.

McBride, C. (2019). *Coping with dyslexia, dysgraphia and ADHD: A global perspective*. Routledge. https://doi.org/10.4324/9781315115566

McBride, C., & Mohseni, F. (2020). *Biliteracy* [manuscript under review]. Department of Psychology, The Chinese University of Hong Kong, Hong Kong

McBride-Chang, C., Chung, K. K., & Tong, X. (2011). Copying skills in relation to word reading and writing in Chinese children with and without dyslexia. *Journal of Experimental Child Psychology*, *110*(3), 422-433.

McCloskey, M., & Rapp, B. (2017). Developmental dysgraphia: An overview and framework for research. *Cognitive Neuropsychology*, *34*(3-4), 65-82. https://doi.org/10.1080/02643294.2017.1369016

Naka, M., & Naoi, H. (1995). The effect of repeated writing on memory. *Memory & Cognition*, *23*(2), 201-212. https://doi.org/10.3758/BF03197222

Nalpon, L. A., & Chia, N. K. H. (2009). Does cursive handwriting have an impact on the reading and spelling performance of children with dyslexic dysgraphia: A quasi-experimental study. *Journal of Reading Literacy*, *1*, 66-106.

Nicolson, R. I., & Fawcett, A. J. (2011). Dyslexia, dysgraphia, procedural learning and the cerebellum. *Cortex, 47*, 117-127. https://doi.org/10.1016/j.cortex.2009.08.016

Overvelde, A., & Hulstijn, W. (2011). Handwriting development in grade 2 and grade 3 primary school children with normal, at risk, or dysgraphic characteristics. *Research in Developmental Disabilities*, *32*(2), 540-548.

Penso, D. E. (1990). *Keyboard, graphic and handwriting skills: Helping people with motor disabilities.* Chapman & Hall. https://doi.org/10.1007/978-1-4899-3162-7

Phelps, J., & Stempel, L. (1988). The Children's Handwriting Evaluation Scale for manuscript writing. *Reading Improvement, 25*(4), 247-254.

Purcell, J., Turkeltaub, P. E., Eden, G. F., & Rapp, B. (2011). Examining the central and peripheral processes of written word production through meta-analysis. *Frontiers in Psychology*, *2*, 239.

Rahim, N., & Jamaludin, Z. (2019). Write-rite: Enhancing handwriting proficiency of children with dysgraphia. *Journal of Information and Communication Technology*, *18*(3), 253-271. https://doi.org/10.32890/jict2019.18.3.8290

Rao, C., Vaid, J., Srinivasan, N., & Chen, H. C. (2011). Orthographic characteristics speed Hindi word naming but slow Urdu naming: Evidence from Hindi/Urdu biliterates. *Reading and Writing*, *24*(6), 679-695.

Rode, G., Pisella, L., Marsal, L., Mercier, S., Rossetti,Y., & Boisson, D. (2006). Prism adaptation improves spatial dysgraphia following right brain damage. *Neuropsychologia*, *44*(12), 2487-2493. https://doi.org/10.1016/j.neuropsychologia.2006.04.002

Rogers, J., & Case-Smith, J. (2002). Relationships between handwriting and keyboarding performance of sixth-grade students. *American Journal of Occupational Therapy*, *56*(1), 34-39.

Rosenblum, S., Epsztein, L., & Josman, N. (2008). Handwriting performance of children with attention deficit hyperactive disorders: A pilot study. *Physical & Occupational Therapy in Pediatrics*, *28*(3), 219-234.

Rosenblum, S., Parush, S., & Weiss, P. L. (2003). Computerized temporal handwriting characteristics of proficient and non-proficient handwriters. *American Journal of Occupational Therapy*, *57*(2), 129-138.

Smits-Engelsman, B. C., &Van Galen, G. P. (1997). Dysgraphia in children: Lasting psychomotor deficiency or transient developmental delay? *Journal of Experimental Child Psychology*, *67*(2), 164-184.

Sterling, C., Farmer, M., Riddick, B., Morgan, S., & Matthews, C. (1998). Adult dyslexic writing. *Dyslexia*, *4*(1), 1-15.

Tafti, M. A., & Abdolrahmani, E. (2014). The effects of a multisensory method combined with relaxation techniques on writing skills and homework anxiety in students with dysgraphia. *International Journal of Psychology and Behavioral Sciences*, *4*(4), 121-127.

Tal-Saban, M., & Weintraub, N. (2019). Motor functions of higher education students with dysgraphia. *Research in Developmental Disabilities*, *94*, 103479. https://doi.org/10.1016/j.ridd.2019.103479

Thiel, L., Sage, K., & Conroy, P. (2015). Retraining writing for functional purposes: A review of the writing therapy literature. *Aphasiology*, *29*(4), 423-441.

Tseng, M. H. (1998). Development of pencil grip position in preschool children. *The Occupational Therapy Journal of Research*, *18*(4), 207-224.

Tseng, M. H., & Chow, S. M. (2000). Perceptual-motor function of school-age children with slow handwriting speed. *American Journal of Occupational Therapy*, *54*(1), 83-88. https://doi.org/10.5014/ajot.54.1.83

Tseng, M. H., & Hsueh, I. P. (1997). Performance of school aged children on a Chinese handwriting speed test. *Occupational Therapy International*, *4*(4), 294-303. https://doi.org/10.1002/oti.61

Van Waelvelde, H., Hellinckx, T., Peersman, W., & Smits-Engelsman, B. C. (2012). SOS: A screening instrument to identify children with handwriting impairments. *Physical & Occupational Therapy in Pediatrics*, *32*(3), 306-319. https://doi.org/10.3109/01942638.2012.678971

Vaughn, S., Schumm, J. S., & Gordon, J. (1992). Early spelling acquisition: Does writing really beat the computer? *Learning Disability Quarterly*, *15*(3), 223-228.

Walker, J. E. (2012). QEEG-guided neurofeedback for remediation of dysgraphia. *Biofeedback*, *40*(3), 113-114. https://doi.org/10.5298/1081-5937-40.3.03

Wang, R., Huang, S., Zhou, Y., & Cai, Z. G. (2020). Chinese character handwriting: A large-scale behavioral study and a database. *Behavior Research Methods*, *52*(1), 82-96.

Wang, M., Ko, I.Y., & Choi, J. (2009). The importance of morphological awareness in Korean–English biliteracy acquisition. *Contemporary Educational Psychology*, *34*(2), 132-142.

Wang,Y., & McBride, C. (2017). Beyond copying: A comparison of multi-component interventions on Chinese early literacy skills. *International Journal of Behavioral Development*, *41*(3), 380-389

Wang,Y., McBride-Chang, C., & Chan, S. F. (2014). Correlates of Chinese kindergarteners' word reading and writing: The unique role of copying skills? *Reading and Writing*, *27*(7), 1281-1302. https://doi.org/10.1007/s11145-013-9486-8

Woodward, S., & Swinth,Y. (2002). Multisensory approach to handwriting remediation: Perceptions of school-based occupational therapists. *American Journal of Occupational Therapy*, *56*(3), 305-312.

Wu, X., Li, W., & Anderson, R. C. (1999). Reading instruction in China. *Journal of Curriculum Studies*, *31*(5), 571-586. https://doi.org/10.1080/002202799183016

Zhang, Q., & Feng, C. (2017). The interaction between central and peripheral processing in Chinese handwritten production: Evidence from the effect of lexicality and radical complexity. *Frontiers in Psychology*, *8*, 334.

Ziegler, J. C., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: A psycholinguistic grain size theory. *Psychological Bulletin*, *131*(1), 3-29. https://doi.org/10.1037/0033-2909.131.1.3

Validity and Judgment Bias in Visual Analysis of Single-Case Data

Jürgen Wilbert[1], Jannis Bosch[1], and Timo Lüke[2]
[1]University of Potsdam
[2]University of Graz

Abstract

Analysis of data from single-case intervention studies commonly involves visual analysis. Previous research indicates that visual analysis may suffer from low reliability and unpromising error rates. We investigated the reliability and validity of visual analysis and explored to what extent data trends affect judgments. We administered a within-subject experiment in which 186 teacher-education students visually analyzed specifically constructed single-case graphs that included either an intervention effect, a trend effect, both effects, or no effect. Participants identified intervention effects in 75% of the graphs, regardless of the existence of a trend. Type I error rates were low (5%) in graphs without a trend but increased fivefold (25%) for graphs with a trend. Inter- and intra-rater reliability was low, particularly when a trend was present in the data.

Keywords: Single-case research, visual analysis, judgment, trend, curriculum-based measurement, reliability, validity

Single-case research has become an important and broadly accepted way to gain insight into educational processes (Gast & Ledford, 2018; Horner et al., 2005). Particularly in the field of special education, single-case research has been adopted as an appropriate method of evaluating the effectiveness of an intervention or the developmental processes underlying difficulties in acquiring academic skills (Kratochwill et al., 2010, 2012). Furthermore, single-case methods can be used by teachers and educators who are interested in evaluating the effects of their interventions or the learning progress of their students (e.g., in combination with curriculum-based measurements). The information resulting from single-case research designs is helpful for decision-making regarding future teaching processes for an individual student but also helps to decide whether or how to implement certain teaching methods in the classroom.

One of the major concerns with single-case studies is the validity of conclusions drawn from the data, with respect to both internal and external validity. Internal validity addresses the question whether a correct causal relation can be inferred from an intervention applied during a single-case study, whereas external validity refers to the generalizability of results across persons, settings, and measurements found in the study (Shadish et al., 2002).

Several strategies have been developed to counter these two methodological issues. These strategies focus either on aspects of the design or on methods for analyzing single-case data. In the present paper, we take a closer look at visual inspection, one of the major methods for analyzing single-case data. We specifically focus on aspects of the internal and external validity of conclusions derived from visual inspections.

Visual Inspection

Visual inspection (or visual analysis; Barton et al., 2018) is one of the most common strategies for analyzing single-case data (Davis et al., 2013; Lane & Gast, 2014). However, it is also one of the most controversial. In visual analysis, a person, usually the

Wilbert, J., Bosch, J., & Lüke, T. (2021). Validity and Judgment Bias in Visual Analysis of Single-Case Data. *International Journal for Research in Learning Disabilities, 5*(1), 13-24. https://doi.org/10.28987/ijrld.5.1.13

investigator, draws a conclusion about the effectiveness of an intervention solely based on inspection of a diagram comprising the measurement times on the x-axis and the measured values on the y-axis with a vertical line indicating the beginning of the intervention (Spriggs et al., 2018).

Proponents of visual inspection argue that this procedure is practitioner friendly, no further in-depth statistical knowledge is required, and the results are directly and easily understandable (Parsonson & Baer, 2015). They assume that any effect large enough to be practically significant will be detected with visual analysis and that advanced statistical procedures sensitive enough to detect smaller effects do not provide additional clinically significant information (Kazdin, 2011; Parsonson & Baer, 2015).

Critics, on the other hand, argue that visual analyses yield low interrater reliabilities (Danov & Symons, 2008; Ottenbacher, 1990; Park et al., 1990; van den Bosch et al., 2017). Furthermore, the presence of a data trend beginning prior to the intervention (i.e., a positive lag 1 autocorrelation) substantially increases the rate of type I judgment errors (i.e., an unsuccessful intervention is erroneously judged as being successful; Allison et al., 1992; Greenwald, 1976; Jones et al., 1978; Matyas & Greenwood, 1990). However, the proclaimed conservative nature of visual analysis, which should lead to an increase in type II error rates (i.e., a successful intervention is erroneously judged as being unsuccessful; Kazdin, 2011; Parsonson & Baer, 2015), has not been corroborated empirically (Matyas & Greenwood, 1990).

Previous research exploring the strategies used by inservice (Espin et al., 2017) and preservice teachers (Wagner et al., 2017) as well as board-certified behavior analysts (Normand & Bailey, 2006) when interpreting single-case data has demonstrated that the same errors occurred independent of previous experience with visual analysis of single-case data (Espin et al., 2017). This cannot be explained by poor general graph-reading skills alone, as Zeuch et al. (2017) found a correlation of $r = .45$ between general graph-reading skills (extracting information from pie, bar, line and other charts) and the accuracy of visual judgments of learning-progress charts (interpreting the first and last data point of the graph and judging the development throughout all data points).

A major challenge in determining the validity of visual analysis involves selecting a standard with which to compare raters' judgments. Most studies compare visual judgments to the results of statistical procedures (e.g., Brossart et al., 2006; Brossart et al., 2014). This approach, estimating the correctness of visual judgment by comparing it to the results of a statistical analysis, implies that the respective statistical procedures are the best possible ways to analyze the data and that raters cannot be more efficient than such statistical analysis. Both assumptions are highly problematic. Hence, statistical and visual analyses cannot be compared properly. It gets even more complicated when different statistical procedures are applied, with some corroborating an effect while others reject it (Parsonson & Baer, 2015).

Model-Based Data Generation as a Standard of Comparison

One way to overcome these problems is to simulate single-case data with highly controllable and known statistical properties. These properties are systematically varied, and single-case graphs are provided to raters with which to judge the presence of an effect or other criteria of interest to the researcher (e.g., Matyas & Greenwood, 1990; Ximenes et al., 2009).

The challenge here is determining which model to base the data generation on. An inappropriate model might threaten the ecological validity of a study (i.e., the extent to which the material approaches the conditions of real-world data). While all models are reductions of the complexity of real-world events, an oversimplified model impairs the generalizability of the conclusions and diminishes the external validity. Moreover, the data-generation process must be deduced from a model that is based on a theory of the factors influencing the measurements across time. Without a sound theoretical foundation, inference from the results of a particular study to the higher-order construct it presents (Shadish et al., 2002) is not possible. That is, a study lacks construct validity.

Huitema and McKean (2000) suggested a general model for single-case data: two-phase single-case designs with a pre-intervention phase comprising measurements before the start of the intervention (Phase A) and an intervention phase containing measurements beginning at the intervention's start and lasting throughout the intervention (Phase B). In this model, four factors predict the outcome at a specific measurement point: the performance at the beginning of the study (intercept), a developmental effect leading to a continuous increase throughout all measurements of both phases (trend effect), an immediate intervention effect leading to an immediate and enduring increase in the level of performance (level effect), and a continuous intervention effect that leads to a continuous increase in the slope of the learning curve (slope effect).

Most investigations that have used artificially constructed single-case graphs to study the quality of visual analyses have focused on participants' accuracy in detecting a level effect of an intervention under varying circumstances (e.g., Brossart et al., 2006; Matyas & Greenwood, 1990; Ximenes et al., 2009). For example, Normand and Bailey (2006) systematically varied level and slope effects in a study on visual aids in visual analysis; however, they did not present completely controlled data but manipulated data of two real-world single case graphs.

Espin et al. (2018) explored the evaluation accuracy and difficulty (i.e., response time) of preservice teachers comparing different "graph patterns;" that is, combinations of level and slope effects and display of goal lines (slope-to-goal and slope-to-slope comparison). However, the stimulus material appeared to be "error-free" as it used straight lines within phases. While their findings are new to the field and relevant as they show that even comparing straight lines was not easy for the participants, the material was rather artificial. In practice, virtually no single-case experiment results in a graph with straight lines for Phases A and B – ideally with a level *and* an additional slope effect recognizable.

Research Questions

In an attempt to fill the above research gaps, the present study addressed the following research questions:

1. How accurate are judgments on single-case graphs? To what extent do baseline trends influence the accuracy of judgments on single-case graphs?
2. How reliable are judgments on single-case graphs? To what extent do baseline trends influence the reliability of judgments on single-case graphs?

For the study, graphs were created using naturalistic – though simulated – data, based on a model including several factors (Huitema & McKean, 2000). Hence, we examined the reliability (intra- and inter-rater) and judgment correctness (power and type I error probability) of visual inspections. Additionally, we examined the impact of a baseline trend on judgment accuracy and reliability regarding the effectiveness of an intervention. The current research focused on an intervention effect that exerts its influence as a continuous increase in performance, starting with the beginning of the intervention (a slope effect). Therefore, we wanted to determine the influence of a trend effect (a positive lag 1 autocorrelation) on judgment correctness and reliability.

Hypotheses

We expected that judgments on the effectiveness of an intervention based on visual analyses would yield high power and low type I error rates when no trend effect is present (Hypothesis 1a). In contrast, when a trend effect is present, we expected increased type I error rates. (Hypothesis 1b). Moreover, we expected a high consistency of judgments between raters and low uncertainty within each rater when no trend effect is present. Hence, both inter- and intra-rater judgment reliabilities should be high (Hypothesis 2a). However, judgments should become unstable and inconsistent between raters resulting in decreased inter- and intra-rater reliabilities when a trend effect is present (Hypothesis 2b).

Furthermore, we wanted to differentiate between a technical judgment on the existence of an intervention effect (intervention effectiveness) and a pedagogical judgment on the efficacy of the intervention (intervention efficacy). The terms *intervention effectiveness* and *intervention efficacy* are used throughout this manuscript in order to distinguish between these two judgment processes. This distinction has not been made before and might provide further insights into the topic.

Method

To answer the research questions and test the hypotheses, we implemented a computer-based within-subject experiment, in which teacher-education students conducted visual analyses of graphs from a fictitious single-case research intervention study on reading. Their judgments are then compared to the graphs' underlying statistical properties.

Participants

A sample of 186 first-year teacher-education majors (89% female) from a research university in Germany participated in our study, ranging in age from 18 to 37 (M = 22.3, SD = 4.6). In self-report ratings 147 (79%), participants reported having no prior knowledge of assessing learning development; 36 (19%) reported having basic and only three (2%) reported having substantial knowledge in this field. Similarly, participants reported having little previous knowledge about single-case data analysis. Specifically, a majority, 143 (77%), had no prior knowledge, 39 (21%) had basic knowledge, and 4 (2%) report-

ed having substantial knowledge. All participants attended the lecture Introduction to Inclusive Education and received partial course credit as compensation. Nevertheless, participation in the study was voluntary as participants had the option of completing an assignment instead of participating. Only two students picked the assignment.

Procedure

After giving written informed consent, participants were instructed and tested in groups of (up to) four in a lab located on campus, seated in front of computers, separated by panel screens. All instruction and the test were computer-administered. To make sure participants understood the principles underlying visual analysis, participants learned about the difference between baseline and intervention phase and the difference between trend and intervention effect in single-case research designs. Further, the instruction included a cover story about single-case research on reading speed.

Afterwards, participants evaluated 80 single-case graphs. The graphs were presented in a randomized order. With the current graph visible, they answered three questions:

1. Does the reading speed of this child change throughout the data? Response options: "it declined," "no change," and "it increased."
2. Does the intervention have an effect? Response options: *Negative effect, No effect*, and *Positive effect* (technical judgment or *intervention effectiveness*).
3. Do you think it is useful to apply this intervention to a child with similar skills? Rated on a 5-point Likert scale with *Certainly not* (0) and *Certainly* (4) as semantic anchors (pedagogical judgment or *intervention efficacy*).

No time limit was set for answering the questions. Participants responded to Question 1 in $M = 4.0$ seconds ($SD = 5.2$), to Question 2 in $M = 3.0$ seconds ($SD = 3.0$), and to Question 3 in $M = 7.2$ seconds ($SD = 6.4$).

Three weeks later, 87 participants, randomly drawn from the first sample, were again presented with a random sample of 40 single-case graphs (10 per condition; details on the four conditions follow) drawn from the original item pool to determine test-retest reliability. The procedure was identical to the first measurement.

Design and Materials

We generated AB single-case graphs using a regression-based method. We adopted a common method to visualize single-case data (Spriggs et al., 2018; see Figure 1 for an example). To distinguish between trend effect (i.e., lag 1 autocorrelation throughout all data points) and intervention effect (i.e., an additional slope effect in Phase B), we implemented a 2 x 2 within-subject design. Both, trend and intervention effect, could be either present or non-present, resulting in the following conditions: trend effect (T^+I^0), trend and intervention effect (T^+I^+), intervention effect (T^0I^+), and no effect (T^0I^0).

A linear model applying the following formula created each of the 80 single-case graphs (20 per condition):

$y_i = \beta_{0i} + \beta_{trend|condition} \times MT + \beta_{intervention|condition} \times (MT-9) \times D + \varepsilon$, where β_{0i} is the intercept (i.e., the starting value) of case *i*, $\beta_{trend|condition}$ is the trend effect size, $\beta_{intervention|condition}$ is the intervention effect size, *MT* is the measurement time, *D* is a dummy-vector showing whether or not an intervention was present, and ε a measurement error.

Although each case was randomly created, simulation parameters were set according to empirical values reported by Klicpera and Schabmann (1993), who investigated the reading speed (words per minute) of German primary school students. The starting value (β_{0i}) was randomly chosen from a normal distribution with $M = 130$ and $SD_{between} = 20$ for each case. Trend effect size was set to one standard deviation across all 30 measurement points of a single case. Therefore, changes per measurement (β_{trend}) for conditions with a trend (T^+I^0 & T^+I^+) was and zero for conditions without trend (T^0I^+ & T^0I^0). The intervention effect size was set to three standard deviations across the 20 Phase B measurements (representing a shift from a very weak to an average reader based on the values reported by Klicpera and Schabmann, 1993). Accordingly, for conditions with an intervention effect (T^0I^+ & T^+I^+) $\beta_{intervention}$ was and for conditions without an intervention effect (T^+I^0 & T^0I^0) $\beta_{intervention}$ was zero.

Variability was introduced as a measurement error affecting each single measurement. The measurement error ε_{ij} for each data point was randomly drawn from a normal distribution with $M = 0$ and $SD = \sqrt{\frac{(1-r_{tt})}{r_{tt}} \times SD_{between}^2}$ with the measurement reliability r_{tt} set to .80 and $SD_{between} = 20$ (the standard deviation of the intercept β_{0i} between cases). Please compare Figure 1 for single-case graphs for each condition prior to and after the addition of measurement errors.

All graphs were created using the package *scan* (Wilbert & Lüke, 2019) for *R* (R Core Team, 2018).

As a data check, we reanalyzed the resulting 80 single-case data sets. For each data set (and phase) we calculated a regression with the criteria (words per second) regressed on measurement time, providing the slope for

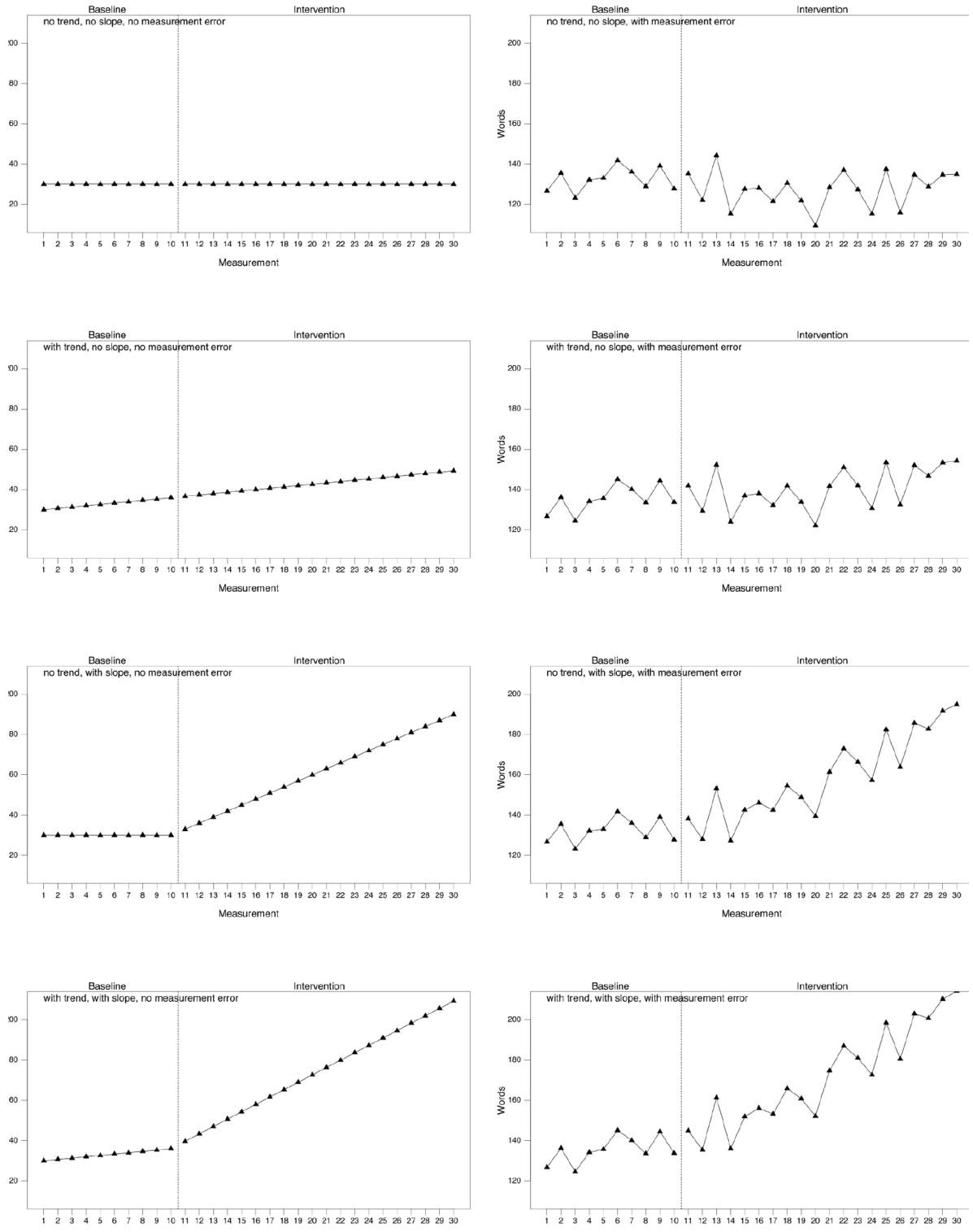


Figure 1.
Sample Items for Each of the Four Conditions Prior to the Addition of Measurement Errors and After the Addition of a Measurement Error

each phase. Additionally, we calculated the difference between the Phase B and Phase A slope for each graph. The values for Phase A indicate the trend effect, and the values for Phase B – Phase A indicate the intervention effect. (See Table 1 for further information on mean regression weights of the 20 items [single-case graphs] per condition.) Overall, the intended effects are represented by the 80 items: The two conditions with intervention effect showed an intervention effect (B-A) of 3.28 and 3.06 words per measurement (d = 3.3 and d = 3.1 for the complete intervention phase), while there was a trend effect of 0.43 and 0.77 words per measurement (d = 0.65 and d = 1.15 for all phases) for the condition with trend effect and nearly no trend effect (-0.14 and 0.11 words per measurement; that is, d = -0.21 and d = 0.17) for the conditions without trend effect.

Table 1
Mean Regression Weights for Each Condition and Phase (Dependent Variable Regressed on Measurement Time)

Condition[a]	PHASE			
	A	B	ALL[b]	B-A[c]
T^+I^+	0.43	3.72	2.97	3.28
T^0I^+	-0.14	2.92	2.26	3.06
T^+I^0	0.77	0.63	0.67	-0.15
T^0I^0	0.11	-0.08	-0.04	-0.19

Note. T^+I^+ *= trend and intervention effect;* $T+I0$ *= trend effect only;* T^0I^+ *= intervention effect only;* T^0I^0 *= no effect.*

[a] *N = 20 items per condition.*

[b] *All values, ignoring phase separation.*

[c] *Difference between regression weights of Phases B and A.*

Item presentation order was randomized and a second list was created with an inverse order. Each participant was randomly assigned to one of the two orders. Participants' judgments for an item were not influenced by the presentation order; therefore, we ruled out an influence of the serial position or participants fatigue on the results and did not include the presentation order in further analyses.

Data Analyses

Because judgments of both intervention effectiveness and intervention efficacy are ordinal data, cumulative link models were applied. As we were not only interested in the overall impact of the trend effect and the intervention effect manipulations, but also the variability of this impact between subjects, we implemented multilevel regression models. Trials on level-1 were nested within subjects on level-2. All predictors were modeled as fixed and random effects. Each regression model included two dummy variables representing presence of trend and intervention effects, and the interaction term, as predictor variables and judgment of intervention effectiveness and rating of intervention efficacy, respectively, as the criterion variable.

We calculated likelihood ratio chi-square tests (Winship & Mare, 1984) to determine the significance of the random slope effects. Thus, the complete model was compared to a model without the target random slope (e.g., in order to calculate the significance of the random slope of the trend effect, the full model was compared to a model with all predictors except the random slope trend effect).

We used intraclass correlation coefficients (ICC) to determine the degree of inter-rater agreement (i.e., the extent of agreement between raters). The ICC conceptualizes inter-rater agreement as the proportion of variance determined by the object of observation (Shrout & Fleiss, 1979). Because we were interested in the degree of absolute agreement rather than consistency of ratings, we used case 2 ICC (2, 1), based on the formalization of McGraw and Wong (1996). To determine whether a trend effect impacts inter-rater agreement, we used separate ICCs for trials with and without trend effect and an F-test based on the procedure suggested by Donner (1986). Additionally, we calculated Fleiss' Kappa, which only assumes categorical data to check for the stability of the results.

Because judgments of both intervention effectiveness and intervention efficacy are ordinal data, we calculated intra-rater reliability (i.e., the stability of ratings within a person) by means of non-parametric correlation coefficients. Average correlations across participants were Fisher's z-transformed in order to account for their skewed distribution.

Results

First, we checked if participants perceived the trend and intervention effect manipulation. As shown in Table 2, on average, more than 90% of the graphs with an intervention effect (T^+I^+ & T^0+I^+) were correctly identified as displaying an overall increase in reading performance (Question 1); about 6% were rated as showing no change. In the condition without intervention effect but with a positive trend effect (T^+I^0), the average ratings identified no change (40%) or an increase (50%). Likewise, 10% attested a decrease in performance.

Table 2
Average Percentage of Judgments on Overall Development of Change in Reading Performance by Condition

Condition	Judgment		
	Decline	No Change	Increase
$T^{+}I^{+}$	0.3	5.7	94.0
$T^{0}I^{+}$	0.6	6.2	93.2
$T^{+}I^{0}$	10.3	39.4	50.3
$T^{0}I^{0}$	42.3	48.0	9.7

Note. $T^{+}I^{+}$ *= trend and intervention effect; T+I0 = trend effect only;* $T^{0}I^{+}$ *= intervention effect only;* $T^{0}I^{0}$ *= no effect.*

Table 3
Average Percentage of Judgment on Intervention Effectiveness by Condition

Condition	Judgment		
	Negative effect	No effect	Positive effect
$T^{+}I^{+}$	0.7	23.6	75.5
$T^{0}I^{+}$	1.4	23.8	74.8
$T^{+}I^{0}$	15.8	57.6	26.6
$T^{0}I^{0}$	44.8	49.9	5.3

Note. $T^{+}I^{+}$ *= trend and intervention effect; T+I0 = trend effect only;* $T^{0}I^{+}$ *= intervention effect only;* $T^{0}I^{0}$ *= no effect.*

Table 4
Average Percentage of Judgments on Intervention Efficacy by Condition

Condition	Judgment				
	Certainly not (efficacious)	Rather not (efficacious)	Uncertain	Rather (efficacious)	Certainly (efficacious)
$T^{+}I^{+}$	0.5	9.6	12.1	42.8	35.0
$T^{0}I^{+}$	0.3	1.3	14.2	46.6	28.5
$T^{+}I^{0}$	12.9	38.5	27.5	17.9	3.2
$T^{0}I^{0}$	35.8	45.0	14.7	3.8	0.7

Note. $T^{+}I^{+}$ *= trend and intervention effect; T+I0 = trend effect only;* $T^{0}I^{+}$ *= intervention effect only;* $T^{0}I^{0}$ *= no effect.*

Graphs in the condition without intervention or trend effect ($T^{0}I^{0}$) were considered as showing no change (48%) or even a decline in performance (42%). Ten percent were rated as increasing reading performance. Hence, participants were able to correctly identify the total increase in reading fluency in the vast majority of graphs with an intervention effect. Judgments on graphs without an intervention effect were also correct in a majority of cases. However, they were a little less accurate than those on graphs with an intervention effect.

Descriptive Analyses

Participants' ratings of the intervention effectiveness (Question 2) are depicted in Table 3. As illustrated, when an intervention effect was present ($T^{+}I^{+}$ and $T^{0}I^{+}$), it was detected on about three out of four occasions with about one fourth of the average ratings identifying *no effect*. Similar to the results on Question 1, there were only marginal differences between judgments in the $T^{+}I^{+}$ and $T^{0}I^{+}$ conditions. When only a trend effect was present ($T^{+}I^{0}$), graphs were mainly regarded as representing no intervention effect (58%), but a substantial proportion was judged as showing a negative (16%) or even positive (27%) intervention effect. For the condition with no effects ($T^{0}I^{0}$), the majority of participants responded *no effect* (50%) or a *negative effect* (45%). However, 5% of graphs were interpreted as showing a *positive effect*.

Participants' ratings on the intervention efficacy (Question 3) are depicted in Table 4. Once again, ratings were similar for the conditions with ($T^{+}I^{+}$) and without ($T^{0}I^{+}$) trend effect if an intervention effect was present: Support for further implementation of this intervention was on the same level as the identification of a positive intervention effect. As expected, support for the intervention was lower for the conditions without intervention effect ($T^{+}I^{0}$ & $T^{0}I^{0}$) and corresponded with the portion of ratings indicating positive intervention effects for these graphs.

Taken together, these results show that the addition of a smaller trend effect had almost no effect on participants' judgments when an intervention effect was already present in the data. When no intervention effect was present, however, a trend effect had a stronger influence on participants' judgment. This pattern was similar for participants' ratings on both intervention efficacy and intervention effectiveness.

We then applied ordinal regression models to further investigate potential interferences of a trend effect on judgment accuracy.

Hypotheses 1a and 1b: Intervention Effectiveness and Intervention Efficacy

Results of the multilevel ordinal regression models are presented in Table 5 (intervention effectiveness)

and Table 6 (intervention efficacy). With respect to intervention effectiveness, the odds ratios suggested that the presence of a trend effect led to a 5.8 times higher chance for the choice of a higher category (i.e., from *negative* to *no* intervention effect or from *no* to *positive* intervention effect) in trials without an intervention effect. However, in trials with an intervention effect, the presence of a trend effect only very slightly increased the chance of a higher category answer than the intervention effect itself, as shown by the odds ratio of .2 for the trend x intervention effect interaction. The intervention effect itself led to a 81.5 times higher chance for the choice of a higher category. Random slope effects documented that all presented effects showed considerable and significant variations, suggesting differentiated influences of trend and intervention effects on the effectiveness judgment between persons.

Table 5

Multilevel Ordinal Regression Model (Logit) for Participants' Judgment on Intervention Effectiveness (Negative Effect, No Effect, Positive Effect)

	β	*SE*	OR[a]	*p*
Fixed				
Trend effect	1.76	0.06	5.8	<.001
Intervention effect	4.40	0.11	81.5	<.001
Trend x intervention effect	-1.61	0.10	0.20	<.001
Thresholds				
0 | 1[b]	-0.22	0.05	0.8	<.001
1 | 2[c]	2.97	0.06	19.5	<.001
Random	*SD* ß	LR	*df*	*p*
Intercept	0.49			
Trend effect	0.50	75.2	4	<.001
Intervention effect	1.18	306.1	4	<.001
Trend x intervention effect	0.39	3.1	4	<.001
Model fit				
LogLik	-10541			
AIC	21113			

Note. Analyses were conducted with the R package ordinal (Christensen, 2019). Trend and intervention effect were dummy-coded (0 and 1).

[a]Odds ratio.

[b]Intercept for judgment negative effect to no effect.

[c]Intercept for judgment no effect to positive effect.

Regression models for intervention efficacy showed a similar pattern (see Table 6). Odds ratios indicated that the trend effect influenced intervention efficacy ratings in trials without an intervention effect (odds ratio of 5 for the trend effect), but not in trials with intervention effect (odds ratio of .3 for the trend x intervention effect interaction). Once again, the intervention effect increased the chance of the choice of a higher category by a factor of 83.7. All effects showed significant random slopes, suggesting that trend and intervention effects also influenced efficacy judgments differentially from person to person.

Table 6

Multilevel Ordinal Regression Model (Logit) for Rating Intervention Efficacy on a 5-Point Likert Scale (0 – Certainly not to 4 – Certainly). Judgments Nested in Individuals

	β	*SE*	OR[a]	*p*
Fixed				
Trend effect	1.62	0.06	5.1	<.001
Intervention effect	4.43	0.13	83.7	<.001
Trend x intervention effect	-1.30	0.04	0.3	<.001
Thresholds				
0 | 1	-0.67	0.08	0.5	<.001
1 | 2	1.76	0.09	5.8	<.001
2 | 3	3.10	0.07	22.1	<.001
3 | 4	5.53	0.10	253.1	<.001
Random	*SD* ß	LR	*df*	*p*
Intercept	1.04			
Trend effect	0.58	107.5	4	<.001
Intervention effect	1.52	88.2	4	<.001
Trend x intervention effect	0.65	24.2	4	<.001
Model fit				
LogLik	-17364			
AIC	34762			

Note. Analyses were conducted with the R package ordinal (Christensen, 2019). Trend and intervention effect were dummy-coded (0 and 1).

[a]Odds ratio.

In summary, results of the regression models showed that in trials without an intervention effect the addition of a trend effect led to a roughly five times higher chance for the choice of a higher category answer for both the intervention effectiveness and the intervention efficacy ratings. In contrast, in trials with an intervention effect, the presence of a trend had only minor effects on participants' answers.

Hypotheses 2a and 2b: Reliability of Visual Judgments

Inter-Rater Reliability (Consistency of Judgments Between Raters)

The overall intraclass correlation between participants was ICC = .50 (95% CI: .43 – .58, F[79, 14378] = 203, $p < .001$) for intervention effectiveness and ICC = .54 (95% CI: .47 – .62, F[79, 14378] = 251, $p < .001$) for intervention efficacy, indicating a relatively low inter-rater reliability (see Table 7). For intervention effectiveness, trials without a trend effect (T^0) showed an ICC of .59 (95% CI: .49 – .70, F[39, 7098] = 283, $p < .001$), while the ICC was .34 (95% CI: .25 – .46, F[39, 7098] = 107, $p < .001$) for trials with a trend effect (T^+). For intervention efficacy, trials without a trend effect showed an ICC of .62 (95% CI: .52 – .73, F[39, 7098] = 340, $p < .001$) while the ICC was .42 (95% CI: .33 – .55, F[39, 7098] = 162, $p < .001$) for trials with a trend effect. Hence, trials with a trend effect had lower inter-rater reliability than trials without a trend for both intervention effectiveness and intervention efficacy.

Intra-Rater Reliability (Stability of Judgments Within Raters)

Median intra-rater reliability coefficients (Kendall's Tau) for the 40 items administered at Measurement Times 1 and 2 are presented in Table 8 for both intervention effectiveness and intervention efficacy. Intra-rater reliability was relatively low overall, with an average Kendall's Tau of .56 for intervention effectiveness and .57 for intervention efficacy. In line with the results for inter-rater reliability, trials with a trend effect showed lower intra-rater reliabilities for both questions. For intervention effectiveness, trials with a trend showed an average Kendall's Tau of .43 compared to .66 for trials without a trend. For intervention efficacy, trials with a trend showed an average Kendall's Tau of .53, while trials without a trend showed an average Kendall's Tau of .60. Hence, similar to the pattern observed for inter-rater reliability, trials with a trend effect had lower intra-rater reliability than trials without a trend for both intervention effectiveness and intervention efficacy.

Table 7
Inter-Rater Reliability and Agreement

	Intervention effectiveness				Intervention efficacy			
	ICC2.1	ICC2.K	r_{wg}	Fleiss' K[a]	ICC2.1	ICC2.K	r_{wg}	Fleiss' K[a]
T^+I^+	.045	.898	.980	.042	.068	.932	.965	.020
T^0I^+	.067	.930	.978	.064	.107	.957	.969	.033
T^+I^0	.145	.969	.946	.086	.170	.974	.962	.044
T^0I^0	.151	.970	.962	.128	.088	.947	.976	.041
T^1	.335	.989	.983	.223	.423	.993	.981	.121
T^0	.587	.996	.985	.356	.616	.997	.986	.193
ALL	.499	.995	.992	.304	.538	.995	.992	.163

Note. r_{wg} = Within-group correlation comparing the variance of all raters to random variance.
[a]Fleiss' Kappa for nominal and ordinal data (Bliese, 2000).

Table 8
Mean Correlation (Kendall's Tau) Between Measurements 1 and 2

Condition	Intervention effectiveness				Intervention efficacy			
	N	*MD*	1st quartile	3rd quartile	*N*	*MD*	1st quartile	3rd quartile
No trend (T^+I^+ & T^+I^0)	76	.66	.55	.75	76	.60	.50	.68
Trend (T^0I^+ & T^0I^0)	76	.43	.31	.59	76	.53	.39	.64
All	76	.56	.47	.67	76	.56	.48	.64

Note. MD is the median correlation for all participants; 1st quartile and 3rd quartile for all participants.

Discussion

In this study we addressed two central questions: First, how reliable are students' evaluations of single-case graphs? Second, to what extent do baseline trends impact judgment of an intervention's efficacy and effectiveness? We conducted a computer-based within-subject experiment, in which students judged 80 AB single-case graphs. As suggested by Ximenes et al. (2009), artificial data sets were created to enable us to vary intervention and trend effects independently.

In line with Matyas and Greenwood (1990) and corroborating Hypotheses 1a and 1b, judgments were found to be quite accurate when no baseline trend was present, with accuracy dropping considerably (type I errors rates increased fivefold) in the presence of a baseline trend. Unfortunately, the most common areas of application for single-case research – and accordingly visual analysis of the resulting graphs – are interventions targeting learning processes where baseline trends are common (e.g., reading fluency, basic arithmetic). Indeed, our findings support the argument that the presence of a data trend is indeed a major reason for type I errors in the visual analysis of single-case graphs.

However, the baseline trend did not reduce type II error rates (i.e., the power remained about 80%). This might be because the intervention effects used in this study were much larger than the trend effects, therefore, possibly overshadowing them in trials where an intervention effect was present. Future research should investigate whether the increase in type I error rates is replicable even when trend and intervention effect sizes are comparable.

In line with Hypotheses 2a and 2b and previous work by other researchers (Jones et al., 1978; Park et al., 1990), inter- and intra-rater reliabilities dropped for items including a baseline trend compared to those without a trend. However, contrary to our expectations in Hsypothesis 2a, reliabilities were low even for items where no trend was present. Judgments appeared both inconsistent across raters and unstable over time. This pattern was similar for judgments on the effectiveness as well as the efficacy of the intervention.

It is often argued that any intervention effect large enough to be relevant in practice is detectable by visual inspection (e.g., Kazdin, 2011; Parsonson & Baer, 2015). However, in line with other studies (e.g., Danov & Symons, 2008; Ottenbacher, 1990; Park et al., 1990), our results indicate that even under relatively clear conditions, with a large intervention effect size and the intervention effect exceeding the Phase A trend, visual judgments were not reliable both within and between raters.

Generally, the distinction between intervention effectiveness and intervention efficacy did not yield insight into the decision process. Trend effects on judgments were slightly more pronounced for intervention effectiveness, and reliabilities were a bit higher for intervention efficacy but overall, the results were very similar for all analyses. Thus, either the experimental variations exerted a consistent effect on both dependent variables or participants did not differentiate between effectiveness and efficacy and both measured practically the same.

Conclusion

In summary, the results of our study suggest that first-year teacher-education majors' visual judgments are unreliable and highly prone to type I errors in the presence of a baseline trend in the data. However, this conclusion must be balanced by several limitations: First, participants were university students with limited experience of visual analysis of single-case data. Note, however, that previous studies resulted in comparable reliabilities and error rates – even in experienced raters (Espin et al., 2017; Normand & Bailey, 2006).

Second, although we did our best to explain the difference between a trend and an intervention effect to the students, no empirical evidence verifies that they truly understood the distinction.

Third, the results were based on an arbitrary decision with regard to the relation between the effect size of the intervention, the trend effect, and the measurement error. Arguably, a change in the proportion between these effects sizes might have led to different results.

Finally, we assumed a linear trend and a linear intervention effect. While we think this is appropriate for a reading intervention, other kinds of interventions (e.g., behavioral modifications or medical treatments) might be better represented with non-linear developments or even performance shifts. Our results, therefore, are not directly applicable to these contexts.

Despite these limitations, the present study provides a novel approach to investigating these effects and reveals that, given certain defined conditions, visual inspections are only of limited value. Two ways to overcome such a limitation have been proposed: First, enriching visual graphs with lines (Kratochwill et al., 2010). A widely applied method consists of drawing a mean line of the A phase extrapolated across the B phase to improve visual inspection accuracy. Similarly, Kadzin (2011) recommended inserting a split-middle line, a type of regression line, for the A phase extrapolating across the B phase. However, Fisher, Kelley,

and Lomas (2003) found an increased type I error rate when inserting a split-middle line. Instead, they proposed a combination of a mean and a split-middle line (the dual criterion, DC) and showed that this procedure leads to an improved visual inspection accuracy. Evaluation studies on variations of this procedure (the conservative dual criterion with 0.25 standard deviations raised lines) have corroborated these findings (Stewart et al., 2007; Young & Daly, 2016).

A second way to improve the accuracy of visual inspections is to validate the interpretations with the results from statistical analyses (Harrington & Velicer, 2015, Park et al., 1990). However, this approach raises the question if a statistical analysis is the most reliable and valid approach to analyzing single-case data in the first place. From this perspective, visual graphs have an important but mere illustrative function.

References

Allison, D. B., Franklin, R. D., & Heshka, S. (1992). Reflections on visual inspection, response guided experimentation, and type I error rate in single-case designs. *The Journal of Experimental Education*, *61*, 45–51. https://doi.org/10.1080/00220973.1992.9943848

Barton, E. E., Lloyd, B. P., Spriggs, A. D., & Gast, D. L. (2018). Visual analysis of graphic data. In J. R. Ledford & D. L. Gast (Eds.), *Single case research methodology: Applications in special education and behavioral sciences* (3rd ed., pp. 179–214). Taylor and Francis.

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Frontiers of industrial and organizational psychology. Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 349–381). Jossey-Bass.

Brossart, D. F., Parker, R., Olson, E. A., & Mahadevan, L. (2006). The relationship between visual analysis and five statistical analyses in a simple AB single-case research design. *Behavior Modification*, *30*, 531–563. https://doi.org/10.1177/0145445503261167

Brossart, D. F., Vannest, K. J., Davis, J. L., & Patience, M. A. (2014). Incorporating nonoverlap indices with visual analysis for quantifying intervention effectiveness in single-case experimental designs. *Neuropsychological Rehabilitation*, *24*, 464–491. https://doi.org/10.1080/09602011.2013.868361

Christensen, R.H.B. (2019). *Ordinal: Regression models for ordinal data* (Version 2019.3-9). R Package.

Danov, S. E., & Symons, F. J. (2008). A survey evaluation of the reliability of visual inspection and functional analysis graphs. *Behavior Modification*, *32*, 828–839. https://doi.org/10.1177/0145445508318606

Davis, D. H., Gagné, P., Fredrick, L. D., Alberto, P. A., Waugh, R. E., & Haardörfer, R. (2013). Augmenting visual analysis in single-case research with hierarchical linear modeling. *Behavior Modification*, *37*, 62–89. https://doi.org/10.1177/0145445512453734

Donner, A. (1986). A review of inference procedures for the intraclass correlation coefficient in the one-way random effects model. *International Statistical Review*, *54*, 67–82. https://doi.org/10.2307/1403259

Espin, C. A., Saab, N., Pat-El, R., Boender, P.D.M., & van der Veen, J. (2018). Curriculum-based measurement progress data: Effects of graph pattern on ease of interpretation. *Zeitschrift Für Erziehungswissenschaft*, *21*, 767–792. https://doi.org/10.1007/s11618-018-0836-9

Espin, C. A., Wayman, M. M., Deno, S. L., McMaster, K. L., & Rooij, M. de (2017). Data-based decision-making: Developing a method for capturing teachers' understanding of CBM graphs. *Learning Disabilities Research & Practice*, *32*, 8–21. https://doi.org/10.1111/ldrp.12123

Fisher, W. W., Kelley, M. E., & Lomas, J. E. (2003). Visual aids and structured criteria for improving visual inspection and interpretation of single-case designs. *Journal of Applied Behavior Analysis*, *36*(3), 387. https://doi.org/10.1901/jaba.2003.36-387

Gast, D. L., & Ledford, J. R. (2018). Research approaches in applied settings. In J. R. Ledford & D. L. Gast (Eds.), *Single case research methodology: Applications in special education and behavioral sciences* (3rd ed., pp. 1–26). Taylor and Francis.

Greenwald, A. G. (1976). Within-subjects designs: To use or not to use? *Psychological Bulletin*, *83*, 314–320. https://doi.org/10.1037/0033-2909.83.2.314

Harrington, M., & Velicer, W. F. (2015). Comparing visual and statistical analysis in single-case studies using published studies. *Multivariate Behavioral Research*, *50*(2), 162. https://doi.org/10.1080/00273171.2014.973989

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children*, *71*, 165–179.

Huitema, B. E., & McKean, J. W. (2000). Design specification issues in time-series intervention models. *Educational and Psychological Measurement*, *60*, 38–58. https://doi.org/10.1177/00131640021970358

Jones, R. R., Weinrott, M. R., & Vaught, R. S. (1978). Effects of serial dependency on the agreement between visual and statistical inference. *Journal of Applied Behavior Analysis*, *11*, 277–283. https://doi.org/10.1901/jaba.1978.11-277

Kazdin, A. E. (2011). *Single-case research designs: Methods for clinical and applied settings* (2nd ed.). Oxford University Press.

Klicpera, C., & Schabmann, A. (1993). Do German-speaking children have a chance to overcome reading and spelling difficulties? A longitudinal survey from the second until the eighth grade. *European Journal of Psychology of Education*, *8*, 307–323. https://doi.org/10.1007/BF03174084

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). *Single-case design technical documentation*. What Works Clearinghouse. https://ies.ed.gov/ncee/wwc/Docs/ReferenceResources/wwc_scd.pdf

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2012). Single-case intervention research design standards. *Remedial and Special Education*, *34*, 26–38. https://doi.org/10.1177/0741932512452794

Lane, J. D., & Gast, D. L. (2014). Visual analysis in single case experimental design studies: brief review and guidelines. *Neuropsychological Rehabilitation*, *24*, 445–463. https://doi.org/10.1080/09602011.2013.815636

Matyas, T. A., & Greenwood, K. M. (1990). Visual analysis of single-case time series: Effects of variability, serial dependence, and magnitude of intervention effects. *Journal of Applied Behavior Analysis*, *23*, 341–351. https://doi.org/10.1901/jaba.1990.23-341

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*, 30–46. https://doi.org/10.1037//1082-989X.1.1.30

Normand, M. P., & Bailey, J. S. (2006). The effects of celeration lines on visual data analysis. *Behavior Modification*, *30*, 295–314. https://doi.org/10.1177/0145445503262406

Ottenbacher, K. J. (1990). Visual inspection of single-subject data: An empirical analysis. *Mental Retardation*, *28*, 283–290.

Park, H.-S., Marascuilo, L., & Gaylord-Ross, R. (1990). Visual inspection and statistical analysis in single-case designs. *The Journal of Experimental Education*, *58*, 311–320. https://doi.org/10.1080/00220973.1990.10806545

Parsonson, B. S., & Baer, D. M. (2015). The visual analysis of data and current researh into the stimuli controlling it. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 15–40). Routledge.

R Core Team. (2018). *R: A language and environment for statistical computing*. Author. http://www.R-project.org/

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Wadsworth Cengage Learning.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–428. https://doi.org/10.1037//0033-2909.86.2.420

Spriggs, A. D., Lane, J. D., & Gast, D. L. (2018). Visual representation of data. In J. R. Ledford & D. L. Gast (Eds.), *Single case research methodology: Applications in special education and behavioral sciences* (3rd ed., pp. 157–178). Taylor and Francis.

Stewart, K. K., Carr, J. E., Brandt, C. W., & McHenry, M. M. (2007). An evaluation of the conservative dual-criterion method for teaching university students to visually inspect Ab-design graphs. *Journal of Applied Behavior Analysis*, *40*(4), 713–718. https://doi.org/10.1901/jaba.2007.713-718

van den Bosch, R. M., Espin, C. A., Chung, S., & Saab, N. (2017). Data-based decision-making: Teachers' comprehension of curriculum-based measurement progress-monitoring graphs. *Learning Disabilities Research & Practice*, *32*, 46–60. https://doi.org/10.1111/ldrp.12122

Wagner, D. L., Hammerschmidt-Snidarich, S. M., Espin, C. A., Seifert, K., & McMaster, K. L. (2017). Pre-service teachers' interpretation of CBM progress monitoring data. *Learning Disabilities Research & Practice*, *32*, 22–31. https://doi.org/10.1111/ldrp.12125

Wilbert, J., & Lüke, T. (2019). *scan: Single-case data analyses for single and multiple designs* (Version 0.40). https://cran.r-project.org/package=scan

Winship, C., & Mare, R. D. (1984). Regression models with ordinal variables. *American Sociological Review*, *49*, 512–525. https://doi.org/10.2307/2095465

Ximenes, V. M., Manolov, R., Solanas, A., & Quera, V. (2009). Factors affecting visual inference in single-case designs. *The Spanish Journal of Psychology*, *12*, 823–832. https://doi.org/10.1017/S1138741600002195

Young, N. D., & Daly, E. J. (2016). An evaluation of prompting and reinforcement for training visual analysis skills. *Journal of Behavioral Education*, *25*(1), 95–119. https://doi.org/10.1007/s10864-015-9234-z

Zeuch, N., Förster, N., & Souvignier, E. (2017). Assessing teachers' competencies to read and interpret graphs from learning progress assessment: Results from tests and interviews. *Learning Disabilities Research & Practice*, *32*, 61–70. https://doi.org/10.1111/ldrp.12126

The Effects of a Comprehensive and Supplemental Middle School Reading Program

Irma F. Brasseur-Hock[1], Whitney Miller[2], Jocelyn Washburn[1], Alyson J. Chroust[3], and Michael F. Hock[1]
[1]The University of Kansas, [2]Virginia Tech, [3]East Tennessee State University

Abstract

We present results of an evaluation of the first year of a multi-year comprehensive middle school reading program. Four public middle schools in rural Virginia with large populations of students with limited reading proficiency participated in a study to determine the reading program's impact. We evaluated 235 students with low reading achievement scores, including students with disabilities, to determine reading gains. The multi-year curriculum consisted of multiple components (word-level instruction, comprehension and vocabulary, motivation and engagement, and assessment) and seven related instructional units, each taught using explicit instruction. A quasi-experimental design was used to determine the intervention's effectiveness. Statistically significant differences were found between the experimental and comparison conditions on a standardized measure of reading achievement with some scores favoring the experimental condition. Results support, in part, the reading program's promise to improve middle school students' reading achievement scores at a level that may narrow the reading achievement gap.

Keywords: Adolescent reading, reading disabilities, reading interventions

In response to identified needs related to the limited reading proficiency (LRP) exhibited by many middle school students, a state agency and several district leaders from rural school districts in the southwestern region of the state of Virginia in the United States contacted the researchers for assistance in exploring possible solutions. Building on an existing partnership, the state and rural district leaders and the researchers decided to implement and evaluate the Fusion Reading (FR) program, a comprehensive intervention for struggling adolescent readers (Hock et al., 2012).

The state started by providing several schools with materials and professional development on FR for several reasons. First, they believed that the intervention could provide LRP students with improvements in the basic skills they need (e.g., decoding, fluency, vocabulary knowledge, comprehension). Second, previous reading interventions for these students had had little or small effects. And third, they believed that the intervention's use of literature that was engaging and relevant to the lives of adolescents would increase student motivation and desire to engage in reading.

The overarching goal of the project was to conduct a rigorous evaluation of FR in rural schools and determine the level of impact on students with low scores on the state Standards of Learning (SOL) reading assessment (VDOE, 2017b). The primary research question was whether or not SOL scores and scores on a standardized reading measure would improve for the students with LRP who were taught FR.

The Challenge of Limited Reading Proficiency for Adolescents

A significant discrepancy exists between the reading abilities of adolescents with limited reading proficiency (LRP) and proficient readers, a discrepancy that has been growing. For example, in 2019, the average eighth-grade reading score on the National Assessment of Educational Progress (NAEP) was 263 points, a significant decline in scores from 2017. For

Brasseur-Hock, I. F., Miller, W., Washburn, J., Chroust, A. J., & Hock, M. F. (2021). The Effects of a Comprehensive and Supplemental Middle School Reading Program. *International Journal for Research in Learning Disabilities, 5*(1), 25-41. https://doi.org/10.28987/ijrld.5.1.25

eighth-grade students living in poverty, the average score was 250 points; for students with disabilities, the average score was 229 points; and for English learners, the average score was 221 points (National Center for Education Statistics [NCES], 2019). More significant is how these point differentials translate into basic reading ability. For eighth-grade students living in poverty, 35% are reading below basic proficiency. For students with disabilities, 68% are reading below basic proficiency, and among English learners, 61% are reading below the basic level (NCES, 2019). Thus, a large number of students are not proficient in the reading skills needed for success in school.

For many students, limited reading proficiency can be a chronic condition. For example, by high school, students with limited reading proficiency are, on average, three years below grade level in reading (Cortiella & Horowitz, 2014). Students who score at below basic skill levels are unable to use prior knowledge to make a comparison, describe the central problem faced by a main character in a text, use context to identify meaning of vocabulary, provide text information to support a generalization, read across text to provide an explanation, or support an opinion with text information or related prior knowledge (NAEP, 2019). Consequently, students reading significantly below a basic level are unable to comprehend much of the written material they encounter in school.

The Magnitude of the Literacy Challenge

We previously conducted a descriptive study to bring clarity to the nature of the reading skills of adolescents, including students with disabilities (Hock et al., 2009). Entering ninth-grade students were administered 11 standardized reading tests across five reading domains: alphabetics, word-level reading, fluency, vocabulary, and comprehension. The results of the study described the differences across reading domains between proficient readers and readers with limited reading proficiency. Students with limited reading proficiency scored statistically significantly lower than their proficient reader counterparts in each domain and 20 or more standard score points lower than the proficient reader group. Sixty-one percent of the limited reading proficiency group scored low in all five reading domains.

In a latent class analysis of the same data set, we found five statistically unique subgroups of adolescent readers with low reading achievement: (a) readers with severe global weaknesses, (b) readers with moderate global weaknesses, (c) dysfluent readers, (d) weak language comprehenders, and (e) weak reading comprehenders (Brasseur-Hock et al., 2011). The profiles of these subgroups demonstrate considerable diversity and are distinguished by their specific strengths and weaknesses. For example, two of the subgroups scored from one to two standard deviations below the mean on almost all reading measures. Another subgroup demonstrated weaknesses only on the measure of fluency.

Other researchers have identified similar reading skill profiles and have extended the research to include related cognitive skill profiles. For example, examining the reading skills and cognitive attributes of middle school students, Miciak et al. (2014) found that measures of phonological awareness, listening comprehension, rapid naming, processing speed, verbal knowledge, and nonverbal reasoning identified three groups of inadequate responders to reading instruction. The three groups included students with (a) comprehension deficit; (b) decoding, fluency, and comprehension deficit; and (c) poor fluency skills. All groups had distinct score clusters for the six measures. Other researchers have found through multigroup confirmatory factor analyses that about 85% of the struggling readers had weaknesses in comprehension, decoding, and fluency (Cirino et al., 2013).

Given the significant and comprehensive needs of LRP students and the diversity of subgroups or clusters of poor comprehenders, increasing student literacy to the level required by more rigorous standards will be a significant challenge for teachers whose students lack basic reading skills.

The Evidence We Have

Literature reviews, meta-analyses, and recent studies of reading interventions, programs, as well as instructional methodology aimed at improving reading proficiency among adolescent struggling readers inform our understanding of what works with whom and under what conditions (e.g., Slavin et al., 2008; Torgesen et al., 2006; Vaughn & Wanzek, 2014). For example, the Center on Instruction's Practice Brief (Boardman et al., 2008) recommends that interventions designed for adolescents include instruction in the following components: word study, fluency, vocabulary, comprehension, and motivation. In addition, based on our studies, we suggest that secondary curricular demands and learner profiles of adolescent struggling readers be taken into consideration when designing and delivering reading interventions (e.g., Brasseur-Hock et al., 2011; Hock et al., 2009).

In the following, we have organized our review of the literature moving from broader instructional approaches to instruction with specific interventions and groups.

Instructional Approaches

In a comparison of four approaches to reading programs for adolescent struggling readers, Slavin and colleagues (2008) found that instructional-process programs, which improve daily teaching practices and are accompanied by professional development, had greater research support than mixed approaches and programs that focus on technology alone. The Slavin review included 33 separate studies, all using randomized or matched control groups.

In a synthesis of 69 experimental research studies across 51 reading programs for secondary students, Baye et al. (2018) found that instructional approaches that used one-to-one and small-group tutoring, cooperative learning, whole-school and writing-focused approaches showed positive outcomes. The researchers also found that reading instruction in social studies/science classes, teaching structured reading strategies, and personalized rotation learning models were effective. However, programs providing an extra hour of reading time and those utilizing technology were no more effective than programs without those features. Thus, across the 69 programs reviewed, the effects were relatively small (i.e., ES = +0.09 to +0.13).

Components of Reading Interventions

In a review of 22 randomized controlled trials (RCTs) on reading interventions for children and adolescents with reading disabilities, Galuschka et al. (2014) evaluated 49 comparisons of experimental and control groups that included reading fluency, phonemic awareness, reading comprehension, phonics instruction, auditory training, medical treatments, and interventions with colored overlays or lenses. A key finding showed that phonics instruction was statistically confirmed as the only approach to affect the reading and spelling performance of children and adolescents with reading disabilities. Specifically, this meta-analysis demonstrated that severe reading and spelling difficulties could be treated with appropriate instructional methods. The authors concluded that systematic instruction of letter–sound correspondences and decoding strategies was the most effective method for improving the literacy skills of children and adolescents with severe reading disabilities. Corroborating these conclusions, the Center on Instruction recommends phonics instruction for older readers to focus on advanced word study and decoding multisyllabic words (Boardman et al., 2008; Torgesen et al., 2007).

Scammacca et al. (2015) examined the findings from 82 studies of interventions for adolescent struggling readers in Grades 4-12. This meta-analysis was conducted as an extension of an earlier meta-analysis (Scammacca et al., 2007) with similar research questions on the level of intervention effectiveness and use of reading comprehension measures. In both literature reviews, the researchers included interventions designed to impact reading fluency, vocabulary, and reading comprehension. Results showed that teachers could influence reading outcomes for older students with reading difficulties and that adolescents, including those with learning disabilities, could benefit from interventions that target both word-level and reading comprehension strategies (Scammacca et al., 2007, 2015). In the latter review, the researchers found that effect sizes in studies of more recent years (1980-2011) showed lower effect sizes, likely due to increased use of standardized measures as the outcome variable for reading comprehension (Scammacca et al., 2015). Additionally, the authors identified three other causes of lower effect sizes: (a) improved "business-as-usual" (BAU) instruction typically serving as the comparison in intervention studies, (b) use of more rigorous research designs, and (c) changes in participant characteristics.

Another synthesis of 14 studies of reading comprehension interventions for middle school students with learning disabilities, conducted between 1979-2009, found large effect sizes for researcher-developed comprehension measures and medium effect sizes for standardized comprehension measures (Solis et al., 2012). All but one intervention in these studies related to strategy instruction on main idea or summarization. However, 12 of the 14 interventions were implemented by researchers, somewhat limiting the generalizability of the findings.

In an effort to determine which features of vocabulary instruction have an influence on adolescents' comprehension, Wright and Cervetti (2017) conducted a systematic review of 36 studies of vocabulary interventions with comprehension as their outcome measure. One key finding from their analysis was that instruction focusing on the strategies for learning new words had a larger impact than teaching definitions of new words. Another finding indicated that there was no evidence to support one particular strategy for solving word meanings, but that students who actively used a strategy showed increased understanding of text.

Severe Reading Disability

In a recent article describing the impact of a two-year randomized control trial study with 194 fourth-grade students with severe reading disabilities, researchers found no statistically significant differences between students in treatment and those in a BAU condition on measures of word identification, vocabulary, and comprehension (Al Otaiba et al., 2018).

However, while there were no significant statistical differences, there were promising effect size gains (ES = 0.14 to 0.19). Given these gains, the researchers suggest that even more intense intervention for students with chronic and severe reading disabilities may be required. For example, the intensive reading program implemented, called Passport to Literacy, was a multicomponent year-long Tier 2 reading program that met four days a week for about 30 min a day. Program components included instruction in word reading skills, vocabulary, and comprehension. Thus, intensive, comprehensive, and multicomponent reading programs may be required for students with severe reading disability (Al Otaiba et al., 2018).

Examining the effects of a year-long, small-group, intensive intervention for 41 eighth graders who persistently had inadequate response to previous reading interventions, Vaughn and colleagues (2012) found that students showed growth but still lacked grade-level proficiency. Students receiving intensive intervention demonstrated significantly higher scores than comparison students on standardized measures of comprehension (ES = 1.20) and word identification (ES = 0.49). However, most students in the treatment condition continued to lack grade-level proficiency in reading despite three years of intervention.

Further, Vaughn et al. (2013) reported the results of a longitudinal study of reading comprehension interventions for adolescents with learning disabilities receiving support within a response-to-intervention framework. In this study, the researchers developed interventions across three tiers of instruction, with increasing levels of intensity for students who were nonresponsive to less intense instruction. The influence of the interventions with added intensity (i.e., Tier 2 and Tier 3 interventions) on student reading achievement scores showed larger gains for the experimental groups than comparison students.

However, the magnitude was considered small (d = 0.16). The results of this study and the previous study by Vaughn and colleagues (2012) show that even with explicit and intensive reading instruction, students with severe reading disabilities demonstrated limited reading improvement and suggest the need for intensive instruction for middle school students with severe reading disabilities to close their reading proficiency gap.

Rural Contextual Factors

To fully examine what reading instruction works for students with LRP, context must also be explored (Eppley et al., 2018). Some unique contextual challenges to literacy improvement efforts in rural school districts include lack of resources, skepticism about externally led initiatives, and limited teacher-collaboration opportunities due to geographic isolation and small school size (Azano, 2015; Hamann & Meltzer, 2005). Additionally, rural schools often experience teacher shortages and high turnover rates (Azano & Stewart, 2016; Holloway, 2002).

In one study of the effectiveness of two commercially available explicit instruction approaches used to address the LRP needs of 49 sixth, seventh, and eighth graders living in rural areas, Shippen et al. (2014) found that students with more skills at pretest demonstrated more growth at posttest, underscoring the importance of paying attention to the initial capabilities of students when evaluating program effectiveness. Additionally, intervention placement procedures must be carefully implemented due to underlying factors that may not parallel placement practices in other places (Callahan et al., 2020). Lastly, in an effort to resist polarizing and rigid conceptions common to socially, culturally, and economically marginalized spaces (Peine et al., 2020), literacy improvement efforts must engage in a partnership and strengths-based approach (Knight et al., 2016).

Amalgamating these findings, we conclude that explicit, comprehensive reading strategy instruction is effective, to varying degrees, for students with LRP. Furthermore, the findings support the need to learn more about the instructional conditions that could close the reading gap for these readers in rural settings. Evidence showing that teachers are able to deliver interventions in real-world settings with as much efficacy as researchers is also needed. Finally, additional research is needed on the impact of multiyear, intensive, and comprehensive reading instruction designed to address all the critical reading component skills identified as essential to have high-impact on the reading achievement of LRP.

A Response to the Challenge: Fusion Reading

In an attempt to address this challenge, several rural Virginia schools adopted FR (Hock et al., 2008), a comprehensive intervention for struggling adolescent readers. Plans were developed for professional development, implementation, and rigorous evaluation that targeted student reading outcomes.

Briefly, FR is an intensive reading class designed to meet for 45-, 60-, or 90-min class periods daily or every other day. The course does not replace language arts or other core classes but is supplemental to core classes and is usually offered in special education classrooms or as an elective. Classes consist of 12-15 nonproficient readers in Grades 6-8 who typically score two or more grade levels below grade placement on a standard reading assess-

ment measure. A major goal of FR is to increase student motivation, engagement, and reading outcomes.

FR consists of seven instructional units. Both teacher and student materials (three workbooks) are provided in hard copy and electronic forms. FR units include (a) Classroom Structure – Establish the Course; (b) Thinking Reading Process; (c) Possible Selves for Readers; (d) Word-Level and Fluency Strategies; (e) Comprehension Strategies; (f) a Vocabulary Strategy; and (c) a Test-Taking Strategy. Throughout the program, daily and unit assessment is provided. Each unit is described in more detail in the Methods section.

Findings From Previous Studies of FR

Multiple studies have shown the impact of FR. For example, as part of an Institute of Education Sciences (2006) grant, an underpowered random assignment study of struggling 9th- and 10th-grade readers in urban high schools was conducted to bolster claims of promise for the intervention. Comparison condition students received Second Chance Reading (Showers et al., 1998). All students were administered the Group Reading and Diagnostic Evaluation (GRADE; Williams, 2001). An independent analysis of the data was conducted by the University of Houston's TIMES Center. Thirty-four students received FR and 35 students received Second Chance Reading.

The data were analyzed using a hierarchical linear modeling approach as implemented in SAS PROC MIXED. The dependent variables were the standard and raw scores on the GRADE comprehension composite test score. A significant interaction was found between treatment and measurement occasion for the standard score on the GRADE Comprehension Composite score, $F(2, 88) = 3.53$, $p = .03$. The pre- to posttest gain for the experimental group was statistically significant, $F(2, 88) = 4.59$, $p = .01$. The within-subjects effect size for this subtest score is Hedges's $g = .70$; $F(2, 93) = 3.06$; $p = .05$ raw score and Hedges's $g = .66$; $F(2, 93) = 3.73$; $p = .03$ for standard scores (Hock, Bulgren et al., 2017).

Another study, a quasi-experimental matched comparison group study, was conducted using FR and Corrective Reading (Hock, Brasseur-Hock et al., 2017). Forty middle school students with learning disabilities were included, 20 in each condition. Students attended a suburban school district. The GRADE (Williams, 2001) was administered pre- and posttest, and the Measure of Academic Progress (MAP) (Northwest Evaluation Association, 2011) was administered at multiple time points.

The difference in GRADE Total Test reading score was statistically significant. Given the nested nature of the data, a repeated-measures analysis of covariance (ANCOVA) was conducted of the overall GRADE total scores. Significant differences were found between the intervention and comparison group over time; $F(1, 32) = 6.67$, $p = .015$, Hedges's $g = 1.66$. A second repeated-measures ANCOVA was conducted on MAP scores. Significant differences between the experimental and comparison groups were found over time; $F(1, 27) = 5.16$, $p = .031$, Hedges's $g = 1.04$.

In another analysis of the same data set, an independent-samples t test was conducted to compare the difference in Total Test scores of the GRADE. The mean score for the experimental group posttest ($M = 33.60$, $SD = 10.29$) was significantly greater than the mean score for the comparison group posttest ($M = 21.70$, $SD = 7.31$), $t(38) = 4.216$, $p < .001$). The standardized effect size index, Cohen's d, was 1.35 (Hock, Brasseur-Hock et al., 2017).

Expanding the Evidence

The evaluation reported in the current paper extends the research on the FR intervention to include students with limited reading proficiency from impoverished rural school districts and cultures. Specific research questions included:

1. What is the impact of FR on the reading achievement scores of middle school adolescents with limited reading proficiency in rural schools?
2. What is the magnitude of the gain score difference (effect size) for students in FR compared to students in a business-as-usual (BAU) reading comparison condition?
3. What is the level of fidelity of implementation for intervention dosage and curriculum implementation?

Methods

Setting

This study took place in two rural school divisions (districts) located in southwest Virginia. One division was a medium-size school division with a total student population of 9,182 across 10 elementary schools, 4 middle schools, and 4 high schools. The other division was a smaller school division with a total student population of 2,042 across two elementary schools, one middle school, and one high school (Virginia Department of Education [VDOE], 2017a).

A total of four schools participated in the study, with three schools from the larger division and one school from the smaller division. Students were from

sixth-, seventh-, or eighth-grade middle school classrooms (see Table 1 for specific student demographics). Three schools taught LRP students the intervention program, FR; the fourth school served as a BAU comparison condition.

In Virginia, where this study took place, a discrepancy similar to the national challenge exists. That is, on a statewide basis, 52% of all LRP students are reading below proficiency compared to 20% of their peers (VDOE, 2017a). The schools participating in this study reported that state Standards of Learning (SOL) scores were also low. For example, one middle school's SOL eighth-grade reading scores were lower than 90.7% of the middle schools in Virginia. Another middle school reported that only 60% of all students scored at the proficient level in reading (VDOE, 2017a).

Student Participants

A total of 235 students participated in the study; 153 students in the experimental condition and 82 in the comparison condition. All students in the study were considered to be LRP, defined as students with documented disabilities and reading goals, English language learners with low reading achievement scores, or students living in poverty. All students had low reading achievement scores. Students (a) were enrolled in Grades 6, 7, or 8; (b) scored between the 15th and 36th percentile on a standardized reading assessment; (c) scored below proficient on the division's reading screening test; and (d) scored below proficient on the Virginia reading Standards of Learning (SOL) test. See Table 1 for additional information on student participants.

Of the 153 students in the experimental group, 54 were in sixth grade, 54 were in seventh grade, and 45 were in eighth grade. A total of 82 students from these same grade levels were in the comparison condition. All students were required to have parent or guardian consent to participate in the study, and students provided their assent to participate.

Teacher Participants

All teachers in both the experimental and comparison groups were VDOE-licensed. The experimental group was taught by three teachers who were new FR program implementers but were experienced teachers in the districts (see Table 2). None of the teachers, experimental or comparison, had prior FR teaching experience. The comparison group teacher was responsible for either directly teaching students in the BAU in enrichment/intervention classes or in enlisting the support of English or social studies teachers to support students as they worked on completing class assignments.

Experimental Condition

Students in the experimental condition received FR. The curriculum includes seven units, each taught using explicit instruction. Bundled into the program are four major components: (a) Engagement and Motivation, (b) Word-Level Instruction, (c) Comprehension, and (d) Ongoing Assessment (Hock et al., 2008).

Components of the Comprehensive Intervention

The Engagement and Motivation component includes the use of highly engaging teen literature, lessons designed for student success through explicit instruction, multilevel reading material, positive and corrective feedback, ongoing performance assessment, and Possible Selves for Readers (PSR) (Hock et al., 2012). PSR is used to focus students' attention on the importance of becoming expert readers and how being expert readers can help them reach their hopes and dreams as learners, individuals, and in career areas. For

Table 1
Student Demographics

School	Students Enrolled	Free/Reduced-Price Lunch	% IEPs [a]	% ELL	Black	Hispanic	White
Exp. 1	448	69.4%	15.6%	1.1%	29.7%	4%	66.3%
Exp. 2	648	53.%%	13.3%	1.2%	11.9%	9.3%	78.7%
Exp. 3	517	65.4%	17.2%	1.4%	29%	7.2%	63.8%
Comp.	576	57.5%	10.2%	9.4%	23.6%	18.4%	58%

Note. IEP= Individualized Education Program; ELL = English-language learner.
[a]Students with disabilities and IEP goals for reading made up 10.2% to 17.2% of the students in the study. Overall, about 35% of the students were classified as LD, 17% as speech-language hearing, 21% as other health impaired, and about 9% as autistic.

Table 2
FR Teacher Demographics

Teacher	Degree	Certifications	Number of Years Teaching	Gender	Race	Age
1	Bachelor's Degree	Early Education NK-4th Middle Education 4th-8th	26 years	Female	White	54
2	Master's in Education	Emotional Disabilities K-12th Specific Learning Disabilities K-12th Elementary 4th-7th Reading Specialist	27 years	Female	White	59
3	Bachelor's Degree	Elementary NK-8th	16 years	Female	White	38
4	Bachelor's Degree	Intellectual Disabilities K-12th Specific Learning Disabilities K-12th	25 years	Female	Black	57

example, students participate in structured interviews in which they describe themselves as an individual, as a learner, and as a worker. They also identify their hopes, expectations, and fears for the future in each of these areas. From this examination of what is possible for each student, an action plan is developed that clearly shows the linkage between reading and the attainment of the student-identified goals. PSR is an ongoing experience and reflects the dynamic nature of student goals.

Word-Level Instruction is taught through The Bridging Strategy (TBS) (Brasseur et al., 2012). TBS consists of four core units: phonics, decoding, word identification, and reading fluency. When students apply TBS, they use multiple skills and strategies to quickly and accurately recognize words in connected text. When they encounter an unfamiliar multisyllabic word, they learn to apply a four-step strategy in which they break unrecognized multiple syllabic words into pronounceable word parts. These word-level skills are explicitly taught to a level of automaticity and practiced with expository and narrative text using multilevel text. Teachers provide positive and corrective feedback to small cooperative groups and, as needed, to individual students.

The Comprehension component of FR consists of four key strategies: (a) Summarization, (b) Prediction, (c) Vocabulary, and (d) Strategy Integration (Brasseur et al., 2012; Hock et al., 2012). With the Summarization Strategy, students learn to identify important clues in the text, link the material to prior knowledge, read short chunks of information, find main ideas, and summarize major sections of text. In the Prediction Strategy, students learn how to make predictions and draw inferences within their reading. With the Vocabulary Strategy, students learn a seven-step process that allows them to determine the meaning of unknown vocabulary through analysis of affixes and context clues and extensive classroom discussion of multiple word meanings, word usage in different contexts, and similarities of the target word to other words. Finally, and most important, through Strategy Integration, students learn how to apply and adapt all the reading strategies they have learned to reading materials in their math, science, language arts, and social studies core classes. They practice application of strategies in the FR class using the core class text materials and receive feedback from their teacher. Core class teachers and co-teachers then cue students to use the strategies during core class activities. About 60% of FR instruction focuses on core class reading material.

Two activities embedded in the Comprehension component, Thinking Reading and Book Study (Brasseur et al., 2012), are designed to increase the amount of time disengaged readers spend engaged in the reading process. Thinking Reading is an instructional process used to demonstrate expert reading behaviors, to forecast strategy application, and to provide opportunities to practice strategy application in the context of authentic reading material. Thinking Reading is similar to Reciprocal Teaching (Palincsar & Brown, 1984) in that the teacher eventually transfers the role of expert reader to students. In Thinking Reading, however, teachers use highly engaging reading materials in an effort to get disengaged readers reengaged with text.

Book Study is designed for extension and application of learned reading strategies outside the classroom. Students select books in their areas of interest and at their independent reading level. Then they complete assignments that are directly related to the strategies and vocabulary being taught. The goals of these activities are to get disengaged readers'"eyes on print" (Chamberlain, 2006, p. 172), provide multiple exposures to expert reader models, offer readers opportunities to practice new reading strategies, and extend reading practice beyond the classroom.

Finally, the Assessment component provides individualized data that inform and personalize instruction. Individual student progress is carefully documented in each instructional unit. Formative assessment data are gathered daily for each strategy's instructional lesson and during the various practice activities. Thus, regular measurement of motivation, engagement, word-level skills, and comprehension is embedded in the program and collected regularly by the teacher. This information helps assess individual student progress and provides immediate, individualized, positive, and corrective feedback to students.

Progress measures are embedded within each major unit of the curriculum. These measures inform the learner and teacher about the level of student mastery of a particular reading strategy, mastery of skills being taught, and comprehension of reading material. The measures are also used to make future program curriculum decisions for individuals or groups of students. Overall achievement gains are documented by division end-of-grade assessments and/or standardized reading measures.

The Instructional Process and Procedures

A key structure of FR is the Daily Lesson Format (DLF), which provides a structure for the class that ensures all critical instructional activities are included in each class session. For example, during a 60-min class, teachers and students rotate through five activities: Warm-Up (5 min), Thinking Reading (12 min), Explicit Instruction (20 min), Vocabulary (18 min), and Wrap-Up (5 min). The instructional activities are as follows:

1. Students do a Warm-Up activity as soon as they enter the classroom. The Warm-Up is usually a vocabulary question related to the novel the class is reading. Students earn points for completing the activity.
2. Students transition to Thinking Reading, where the teacher models metacognitive strategies and the thinking of an expert reader. Students read highly engaging novels and eventually demonstrate and practice the reading strategies they have been taught.
3. During Explicit Instruction teachers explain a strategy, model the strategy, guide student application of the strategy, have students practice the strategy, and then provide feedback to students. Students are taught the individual course reading strategies during this time.
4. Next, students study Vocabulary, and are guided through the seven-step strategy applying affix meanings and discussion to define the meaning of the word; opportunities are provided to locate other words that contain the same affixes.
5. Finally, during lesson Wrap-Up, students are given a quick assessment of the main skill taught. Usually, this involves having students complete an exit ticket assignment. Also, the upcoming lesson is previewed.

The DLF structure helps ensure that each class has instructional variety and that every minute possible is an opportunity for explicit instruction.

Explicit Instruction. Teachers follow explicit instruction practices for all reading and strategy instruction. In FR, the procedure includes the following steps: First, teachers clearly explain each skill or strategy that will be learned during each lesson. Then teachers provide an expert model of how the skill is applied or how the strategy works in the context of narrative or expository text. Once teachers have provided an expert model of the skill or strategy, they engage students in guided practice.

Guided practice scaffolds teachers' support, with the students taking responsibility for application of the reading skill or strategy. Guided practice is a recursive process with the teacher providing additional modeling and supports as needed. Once students demonstrate some level of initial proficiency in guided practice, they work with a partner and continue to practice with reading material that moves from easy to more difficult levels. During partner practice, the teacher works with individual students to assess proficiency and provide support to students who require elaborated feedback. Finally, new skills and strategies are applied to actual core class materials. These processes and procedures are followed for each of the daily lessons in the program.

Professional Development

Each participating FR teacher received extensive blended PD from one of the program developers and two certified FR trainers. That is, face-to-face PD was provided in combination with online modules designed to provide personalized professional learning. Blended professional development for this study is defined as consisting of both online digital media and face-to-face PD and

coaching from FR coaches. In addition, building and district-level administrators responsible for curriculum and instruction also received PD. The importance of including building and district leaders in secondary school PD plans is well documented (e.g., Bredeson, 2000; McDonald et al., 2009). The model employed to provide all PD was based on validated practices for professional learning (e.g., Darling-Hammond et al., 2017; Fullan, 2005; Knight & van Nieuwerburgh, 2012; Kurz et al., 2017).

The specialized PD provided to FR teachers was scheduled based on the pace of their implementation. Summer PD for Year 1 was conducted over two consecutive days. Training included information on attributes of struggling readers, theoretical underpinnings of FR, classroom routines and set-up, instructional methodology, student grouping strategies, progress monitoring, and an overview of the instructional materials. In addition, the FR teachers were taught how to instruct students during the first unit of the curriculum entitled Establish the Course.

During the fall semester of Year 1, each FR teacher received three additional days of PD over the course of three months that included instruction on the Prediction Strategy, Possible Selves for Readers, and The Bridging Strategy. PD for the Prediction Strategy and Possible Selves for Readers was conducted over two half-day sessions on different dates in the form of a professional learning community (PLC). This PLC watched the online modules for each of the strategies and engaged in in-depth conversations about implementation and next steps in combination with their FR coaches. The Bridging Strategy PD was implemented differently due to the content of this strategy. Since past FR teachers had found PD for the Bridging Strategy to be more challenging than other strategies, Bridging Strategy PD was conducted by certified trainers over an entire professional learning day.

Coaching. FR coaches and professional developers were in frequent contact with the curriculum director and special education coordinator to respond to questions and monitor progress. Furthermore, FR coaches provided monthly coaching to each of the FR sites to ensure fidelity of implementation. During the coaching sessions, the FR coaches employed strategy checklists, classroom modeling requested by the FR teacher of specific strategy components, problem-solving and comparing checklists during their planning period, and provided encouragement and motivation for each of the FR teachers. Coaching techniques followed the principles of Partnership Instructional Coaching (e.g., Knight, 2007, 2009).

End of Year 1. In June 2017 of Year 1, FR teachers received a full-day review of the following information from Year 1: data analysis from each of the FR sites, student success stories, review and refresher of Thinking Reading and administration and scoring of the TOSCRF-2, FR alignment with Virginia's SOL, and teacher survey review of Year 1 FR; in addition, they began planning the launch of FR Year 2 for the 2017-2018 school year.

Comparison Condition

The BAU teacher was described as not using a specific intensive and explicit reading program. Instead, BAU instruction was teacher-designed remediation lessons using the grade-level English curriculum. For example, students in need of reading assistance were scheduled for 45-minute sessions during an academic resource period that met five days per week throughout the school year. During this time, the students received tutoring support for their English language course assignments by the special education teacher. No formal reading program was universally provided. In addition, some core class teachers tutored these same students in the general education curriculum based on out-of-class assignments or homework assignments. Thus, students with disabilities were primarily instructed or tutored by the special education teacher who followed the tutorial model described previously. Adolescents with LRP and without disabilities were instructed or tutored by their grade-level English teachers.

Measures

Two measures were used in this study: the state of Virginia Reading Standards of Learning (SOL) assessment (VDOE, 2017b) and the Test of Silent Contextualized Reading Fluency-2 (TOSCRF-2; Hammill et al., 2014). The SOL assessment contains two types of tests, the online passage-based computer adaptive test (CAT) and the traditional test. A passage-based CAT is a customized assessment where each student receives a unique set of passages and items. The passages are fictional and nonfictional taken from the state's core class curricula. For example, questions from *The Monkey's Paw* (W.W. Jacobs, 1902) are included in the middle school English language arts test. This is in contrast to the traditional test in which all students who take a particular version of the test receive the same passages and respond to the same test questions. The reading test covers the SOL in the reading strand of the English SOL. The SOL are grouped into categories, labeled as reporting categories, that address related content and skills. For example, a reporting category for the read-

ing SOL test is: *Use word analysis strategies and word reference materials*. Each SOL in this reporting category addresses skills using word analysis strategies or word reference materials. When the results of the SOL tests are reported, the scores are presented for each reporting category and as a total test score. The Virginia Reading SOL assessments provide no data on reliability or validity. However, the tests are developed with teacher input and are aligned with the state standards, which provides some measure of validity.

The second measure, The Test of Silent Contextual Reading Fluency-2 (TOSCRF-2; Hammill et al., 2014), is an updated version of the TOSCRF (Hammill et al., 2006) and was normed on a nationally representative sample of 2,375 students ranging in age from 7 to 24 years. The test measures the speed with which students can recognize the individual words in a series of passages that become progressively more difficult in content, vocabulary, and grammar. The TOSCRF-2 measures a variety of reading skills, including recognizing print words and knowing their meaning, use of syntax and morphology, using word knowledge and grammar to grasp the meaning of words, sentences, paragraphs, contextual material, and to understand contextual material with silent fluency. The TOSCRF-2 also measures fluency.

Authors of the TOSCRF-2 report very large correlations with popular measures of reading comprehension (mean corrected correlation .75; range .41–.92). For example, the average correlation between TOSCRF-2 and the Oral Reading Index from *Gray Oral Reading Tests—Fifth Edition* (GORT-5; Wiederholt & Bryant, 2012) was .73. The tests also correlated .75 with the *Tests of Silent Reading Efficiency and Comprehension* (TOSREC; Wagner et al., 2010). The TOSCRF-2 has evidence of high reliability (median .87; range .84–.90), sensitivity (median .78; range .73–.84), specificity (median .79; range .71–.84), and receiver operating characteristic/area under the curve (ROC/AUC; median .88; range .85–.89).

Research Design

The research design for this study was a quasi-experimental comparison group design involving intact groups. One division with three middle schools was selected by VDOE to implement and evaluate the FR program. In order to strengthen the evaluation, a comparison middle school from another division agreed to participate as the comparison condition. The comparison school was given the opportunity to adopt FR after the study was completed. Table 1 compares the characteristics of the participating schools receiving FR.

The three schools implementing FR were labeled Fusion Reading 1, Fusion Reading 2, and Fusion Reading 3. These three schools made up the experimental condition. The fourth middle school was the comparison condition, which offered LRP readers BAU support for reading instruction.

Fidelity of Implementation

Instructional checklists designed to measure implementation of the FR program and a Fusion Reading Teacher Reflection (FRTR) form were developed to measure fidelity of implementation for the experimental condition. Fidelity was conceptualized as the difference between the intended program model, based on FR lesson plans, and the FR program actually implemented by the teacher.

The first checklist, What's Fusion Reading Looking Like?, was divided into two major sections: global fidelity to the lesson format and fidelity to specific instructional procedures. The fidelity checklist measured how closely the FR teacher followed the design of the DLF and instructional practices. The fidelity checklist observation measure was administered for all six lessons: (a) Classroom Climate, (b) Daily Warm-Up Activity, (c) Thinking Reading, (d) Explicit Instruction, (e) Vocabulary, and (f) Wrap-Up. The second checklist, Vocabulary Instruction, was intended to help guide FR teachers through the entire seven-step vocabulary strategy, which in turn allowed students to be engaged with meaningful discussions and make decisions about the meaning of a given word. The third checklist, Thinking Reading, evaluated how well FR teachers implemented the Thinking Reading process. That is, did the teachers model how a strategic reader reads as well as how a strategic reader thinks while making sense of the text? The final form, How's Fusion Reading Going?, was given to FR teachers a week prior to a scheduled monthly coaching visit. This form allowed FR teachers to provide FR coaches with feedback on how they had been progressing with FR. The information helped FR coaches to plan their visit and address any fidelity issues or barriers a FR teacher may have documented.

FR coaches learned how to utilize each of the checklists through online modules prepared by the FR program authors. FR coaches met monthly to compare checklists and feedback given to FR teachers to ensure consistent decisions about fidelity of instruction among FR coaches. Furthermore, FR coaches met with the FR authors using a virtual conferencing tool to deepen the understanding and generalization of coaching FR with each of the FR schools. All information gathered during these meetings was deliberated and shared amongst all FR coaches.

FR coaches made observational notes regarding fidelity of FR implementation based on instructional

component checklists. Additional fidelity information was gathered directly from the FR teachers when they completed the FRTR form prior to a scheduled monthly coaching visit. Checklists and observational notes were the foundation of coaching conversations, and were given directly to FR teachers by the coaches at the request of the teachers. Goals for the next coaching session were established and grounded on these checklists and any anecdotal information provided to the FR teacher. Additionally, FR coaches demonstrated specific components of FR when requested by the FR teacher.

Qualitative analysis of all data gathered indicated that two of three FR teachers implemented FR with a high level of fidelity. The remaining teacher had numerous absences during the 2016-2017 school year due to documented medical reasons, and consequently was unable to focus her attention on the new intervention being implemented.

Analysis and Results

An analysis of outlier status using percentiles and boxplots (using SPSS version 22; Tukey, 1977) was conducted in accordance with standard practice to protect against inflated error rates and distortions of statistical estimates. The scores of zero students were outliers; thus, all scores were included in subsequent analyses.

To determine whether there were differences in performance between comparison and FR students, an ANCOVA was conducted on students' TOSCRF-2 scores with grade level (sixth, seventh, and eighth grade) and group/school (Comparison, Fusion Reading 1, Fusion Reading 2, and Fusion Reading 3) as between-subjects variables and 2016 scores from the Virginia Standards of Learning measure (VA SOL) as a covariate. VA SOL scores will be identified as just scores in the following text. Partial eta-squared was used as a measure of effect size (Richardson, 2011). Effect sizes for partial eta squared ($\eta_p^{2)}$ are generally considered small 0.01, medium 0.06, or large 0.14 (Murphy & Myous, 2004).

The grade-level group difference between sixth-, seventh-, and eighth-grade students' mean TOSCRF-2 scores at pretest was nonsignificant after statistically controlling for 2016 scores, $F(2, 222) = 0.27$, $p = 0.773$, $\eta_p^2 = 0.08$. There was insufficient evidence to indicate a difference in performance between the three grade levels. In contrast, the group difference between comparison and FR school students' mean TOSCRF-2 scores at posttest was significant after statistically controlling for 2016 scores, $F(3, 222) = 8.67$, $p = 0.01$, $\eta_p^2 = 0.81$. However, this main effect was qualified by a significant interaction between grade level and group, $F(6, 222) = 4.07$, $p = 0.001$, $\eta_p^2 = 0.10$.

To investigate this interaction, the data from each grade level were examined separately. Results showed that the sixth-grade students' average TOSCRF-2 scores differed significantly as a function of group, $F(3, 69) = 11.92$, $p < 0.01$, $\eta_p^2 = 0.34$, after statistically controlling for 2016 scores. Using Bonferroni correction (adjusted $\alpha = .0167$), pairwise comparisons of sixth-grade students' data revealed that the scores of students who received the FR intervention were higher than the scores of students who received the comparison intervention (BAU) (see Table 3). The seventh-grade students' average TOSCRF-2 scores differed significantly as a function of group, $F(3, 71) = 10.30$, $p < .001$, $\eta_p^2 = 0.30$, after statistically controlling for 2016 scores. Using Bonferroni correction (adjusted $\alpha = .0167$), pairwise comparisons of seventh-grade students' data revealed that the scores of students who received the FR intervention were higher than the scores of students who received the comparison intervention (see Table 3). Finally, the eighth-grade students' average TOSCRF-2 scores also differed significantly as a function of group, $F(3, 80) = 20.35$, $p < .001$, $\eta_p^2 = 0.43$, after statistically controlling for 2016 scores. Using Bonferroni correction (adjusted $\alpha = .0167$), pairwise comparisons of eighth-grade students' data revealed that the scores of students at two out of the three FR intervention schools were higher than the scores of students who received the comparison intervention (see Table 3).

To investigate this interaction, the data from each grade level were examined separately. The following pairwise comparisons controlled for multiple comparisons through the Bonferroni adjustment for multiple comparisons ($\alpha = 0.0167$). The sixth-grade students' average VA SOL scores were marginally significant as a function of school/group, $F(3, 81) = 2.16$, $p < 0.10$, $\eta_p^2 = 0.07$, after statistically controlling for 2016 VA SOL scores. Pairwise comparisons of sixth-grade students' data failed to reach statistical significance; however, inspection of individual schools indicated that student scores at two of three FR schools improved (see Table 4). The seventh-grade students' average VA SOL scores differed significantly as a function of school/group, $F(3, 77) = 7.01$, $p < .001$, $\eta_p^2 = 0.22$, after statistically controlling for 2016 scores. Similar to sixth-graders' data, student scores at FR intervention schools were higher than the scores of students who received the comparison intervention; however, only one pairwise comparison reached statistical significance (see Table 4). There was insufficient evidence to indicate a difference in eighth-grade students' average VA SOL scores as a function of school/group, $F(3, 100) = 0.47$, $p = 0.71$, $\eta_p^2 = 0.01$, after statistically controlling for 2016 scores.

A post hoc power analysis was conducted with G*Power (Erdfelder et al., 1996) to determine whether

Table 3
Mean (and Standard Error) TOSCRF-2 Scores

Grade Level	Group	*N*	Mean TOSCRF-2 (SE)	*p*-value[a] (vs. comparison school)
6th Grade	Comparison	20	18.35 (5.00)	
	Fusion 1	16	48.85 (4.36)	< 0.001
	Fusion 2	21	53.50 (4.60)	< 0.001
	Fusion 3	17	57.29 (5.20)	< 0.001
7th Grade	Comparison	22	24.42 (4.65)	
	Fusion 1	15	58.27 (5.66)	< 0.001
	Fusion 2	24	50.09 (4.46)	0.001
	Fusion 3	15	56.98 (5.64)	< 0.001
8th Grade	Comparison	40	28.32 (3.12)	
	Fusion 1	10	64.92 (6.24)	< 0.001
	Fusion 2	15	68.30 (5.10)	< 0.001
	Fusion 3	20	36.12 (4.41)	0.919

[a]*p*-values reflect Bonferroni adjustment for multiple comparisons.

the design had sufficient power to detect an interaction between grade level and group. The effect size *f* (based on the partial eta-squared of 0.05) was 0.23. The power to detect an effect of this size with four groups, one covariate, and a total sample size of 273 was determined to be 0.81. In contrast, power analyses for the pairwise comparisons for sixth and seventh graders indicated that the contrasts between the comparison and Fusion 1 were underpowered to detect an effect (d = 0.58, alpha = 0.0167, df = 42, one-tailed, power = 37.9%; d = 0.06, alpha = 0.0167, df = 39, one-tailed, power = 2.53%) whereas the other two contrasts had 99.99% power to detect an effect.

Discussion

Regarding Research Questions 1 and 2, the findings from this study of middle school students with LRP indicated that students who received the FR program performed significantly higher on a standardized measure of reading skills than students in a comparison middle school. Specifically, when reading skills were assessed using the TOSCRF-2, FR students, across the three grade levels, scored significantly higher than comparison students. The TOSCRF-2 measures a variety of reading skills, including recognizing print words and knowing their meaning; use of syntax and morphology; and using word knowledge and grammar to grasp the meaning of words, sentences, paragraphs, contextual material, and to understand contextual material with silent fluency. In previous studies, we have found the TOSCRF and TOSCRF-2 to be sensitive to the FR program.

The impact or effect size of the differences in scores between the FR and comparison groups on the TOSCRF-2 was large favoring the FR condition. In addition, the effect on scores between grade-level groups favored the FR groups as well, indicating a more moderate effect. However, the eighth-grade group comparison only favored two of the three FR groups.

The VA SOL assessment requires sixth-grade students to be able to discuss the impact of setting on plot development; describe character development; differentiate between first- and third-person point of view; differentiate between free verse and rhymed poetry; explain how an author's choice of vocabulary contributes to the author's style; skim materials to develop a general overview of content and to locate specific information; identify transitional words and phrases that signal an author's organizational pattern; identify organizational pattern(s); identify the elements of narrative structure, including setting, character, plot, conflict; describe how word choice and imagery contribute to the meaning of a text; identify and analyze the author's use of figurative language; and analyze ideas within and between selections providing textual evidence (VDOE, 2017b).

Table 4
Mean (and Standard Error) VA SOL Scores

Grade Level	Group	*N*	Mean VA SOL (SE)	*p*-value[a] (vs. comparison school)
6th Grade	Comparison	27	381.66 (6.76)	
	Fusion 1	17	377.28 (8.29)	0.999
	Fusion 2	23	392.25 (7.02)	0.999
	Fusion 3	19	402.56 (7.81)	0.309
7th Grade	Comparison	26	387.15 (6.03)	
	Fusion 1	15	387.56 (7.99)	0.999
	Fusion 2	25	422.66 (6.16)	0.001
	Fusion 3	16	394.13 (7.71)	0.999
8th Grade	Comparison	54	373.16 (4.55)	
	Fusion 1	12	384.63 (9.66)	0.999
	Fusion 2	19	378.57 (7.68)	0.999
	Fusion 3	20	378.66 (7.48)	0.999

[a]*p*-values reflect Bonferroni adjustment for multiple comparisons.

Most of these skills were not a focus of the supplemental FR Year 1 program. These skills are typically addressed in English language arts classes and supported by multiple occasions to integrate FR skills and strategies with English language arts content materials. Thus, significant statistical differences were not found between the FR students and the comparison students on the state reading SOL measure. The effects of FR on student outcomes for sixth grade was small, while seventh-grade students showed moderate to large impact. Effects for eighth-grade students were not significant.

In sum, FR shows promise as a supplemental and comprehensive reading program for adolescent readers with LRP whose low reading achievement is related to a lack of basic word level, vocabulary, and reading comprehension strategies. How, and if, FR can address the specific language arts skills on the VA SOL assessment (or other state reading assessments) is unclear. Thus, while FR does focus on supported generalization and integration of reading skills and strategies necessary for success in core classes, enhancements to the integration process seem warranted. That is, more explicate instruction and extended practice with elaborated feedback as students apply reading skills and strategies to actual English language arts course material may help them acquire the language arts skills measured by state SOL assessments. Additionally, some language arts standards may need to be woven into the FR program.

Fidelity of Implementation

Our third research question addressed fidelity of implementation of the FR program. Measures and checklists of fidelity, developed during previous studies, were used to measure fidelity across several domains, including (a) global fidelity, (b) instructional procedures, (c) Thinking Reading procedures, and (d) the vocabulary instructional process. While we were unable to retrieve all the measures and conduct a statistical analysis of fidelity data, we were able to make informed decisions about the overall level of fidelity from coaching notes and logs, concluding that two of three experimental teachers had a high level of fidelity and one had a low level of fidelity due, in large measure, to chronic health and absentee issues. The low-fidelity classroom may have suffered from extensive use of substitute teachers, who were not formally taught how to teach the FR program. Instead, the substitutes focused on Thinking Reading and learning vocabulary words by an independent study activity. Given these data limitations, the study goal of measuring fidelity of implementation could not be fully documented.

Limitations

There are several limitations to this study. First, the lack of direct data on fidelity of implementation of the

BAU program limits the comparison. Whether BAU was fully implemented as designed and whether and where instructional overlap occurred between BAU and FR is unknown. For example, both the BAU and FR could have had elements of explicit instruction, and explicit instruction has been found to positively impact reading outcomes for students with disabilities (e.g., Swanson, 1999). Not knowing if BAU skills and strategies were taught explicitly, or were taught at all, limits our understanding of what works. In addition, and as explained above, much of the fidelity of implementation data for the FR condition was missing at the end of the study, and the statistical analysis of fidelity of implementation was limited. This was due to the desire of teachers to receive documentation of written checklist feedback during the coaching session and our decision to honor this request. Beyond the consensus of FR coaches, the extent to which the FR program was implemented with fidelity in its totality is unknown.

Second, compared to the experimental groups, the sample size of the comparison condition was smaller and from only one school. This could impact findings as groups could be impacted by factors not related solely to reading achievement. Thus, while the comparison school was matched on several key points, the quasi-experimental design limits the strength of the findings.

Third, FR is a comprehensive and intensive adolescent reading program. In this study, we report only on the results of one year of a multi-year program. Thus, during this period, students received only a portion of what is designed to be a program that builds upon mastery of seven core reading strategy units. It may be, therefore, that the more distal effect on the SOL test scores had not yet occurred after just one year. Other researchers have concluded that more than one year might be needed for some students with LRP. For example, Vaughn et al. (2012) suggested that multiple-year reading interventions might be needed to close the reading achievement gap, Thus, it is unknown what change or impact the program might have on students who participate in the instructional activities beyond one year. Our goal for this study was to measure the promise of the FR program to improve reading outcomes after one year of instruction and to respond to the school district to evaluate impact after one year of instruction. This information would be used to help determine if the program should continue.

Implications

Supplemental reading programs can be effective if certain systems and structures are in place (Bemboom & McMaster, 2013). For example, a supplemental course requires scheduling support, extensive PD and coaching, a dedicated classroom, and instructional materials. Teachers need extended time to teach; FR requires that students attend the class five times a week for at least 50 min each day. Scheduling challenges in middle and high schools need to be addressed before effective supplemental instruction can be delivered to all adolescent LRP readers. For example, since FR is supplemental, students may have to use an elective class option to participate in the class, forcing them to miss another elective.

We are convinced that there is no short-term solution to the challenge of improving the reading outcomes of adolescent struggling readers, and for that reason we have designed FR as a multi-year curriculum. Other researchers are developing comprehensive reading programs that move beyond six- or eight-week courses. As additional data are becoming available from rigorous studies of adolescent reading programs, there is some consistency in the difficulty of obtaining high-impact outcomes that document closing of the achievement gap (e.g., Al Otaiba et al., 2018; Vaughn et al., 2012). Finally, the direct link of supplemental reading courses to core class material is critical for generalization of reading skills. Supplemental reading programs that are decontextualized from core class text materials may be one reason for the limited long-term effects of some current reading programs. We believe that additional focus is needed to support integration and application of reading skills and strategies to authentic core class materials. While comprehensive, intensive adolescent reading programs may be part of the solution to the challenges facing adolescents with LRP, the integration of instruction that makes practice and elaborative feedback more personalized may be another.

References

Al Otaiba, S. A., Petscher, Y., Wanzek, J., Lan, P., & Rivas, B. (2018). I'm not throwing away my shot: What Alexander Hamilton can tell us about standard reading interventions. *Learning Disabilities Research & Practice, 33*(3), 156-167. https://doi.org/10.1111/ldrp.12179

Azano, A. P. (2015). Addressing the rural context in literacies research: A call to action. *Journal of Adolescent & Adult Literacy*, *59*(3), 267-269. https://doi.org/10.1002/jaal.480

Azano, A. P., & Stewart, T. T. (2016). Confronting challenges at the intersection of rurality, place, and teacher preparation: Improving efforts in teacher education to staff rural schools. *Global Education Review*, *3*(1), 108-128.

Baye, A., Lake, C., Inns, A., & Slavin, R. E. (2018). A synthesis of quantitative research on reading programs for secondary students. *Reading Research Quarterly*, *54*(2), 133-166. https://doi.org/10.1002/rrq.229

Bemboom, C., & McMaster, K. (2013). A comparison of lower- and higher-resourced tier 2 reading interventions for high school sophomores. *Learning Disabilities Research and Practice*, *28*(4), 184-195. https://doi.org/10.1111/ldrp.12020

Boardman, A. G., Roberts, G., Vaughn, S., Wexler, J., Murray, C. S., & Kosanovich, M. (2008). *Effective instruction for adolescent struggling readers: A practice brief*. RMC Research Corporation, Center on Instruction.

Brasseur, I. F., Hock, M. F., & Deshler, D. D. (2012). *The bridging strategy.* McGraw-Hill Education.

Brasseur-Hock, I. F., Hock, M. F., Kieffer, M. J., Biancarosa, G., & Deshler, D. D. (2011). Adolescent struggling readers in urban schools: Results of a latent class analysis. *Learning and Individual Differences*, *21*(4), 438-452. https://doi.org/10.1016/j.lindif.2011.01.008

Bredeson, P. V. (2000). The school principal's role in teacher professional development. *Journal of In-Service Education*, *26*(2), 385-401. https://doi.org/10.1080/13674580000200114

Callahan, C. M., Azano, A. P., Park, S., Brodersen, A. V., Caughey, M., Bass, E. L., & Amspaugh, C. M. (2020). Validation of instruments for measuring affective outcomes in gifted education. *Journal of Advanced Academics*, *31*(4), 470-505. https://doi.org/10.1177/1932202X20929963

Chamberlain, S. P. (2006). Sharon Vaughn: The state of reading research and instruction for struggling readers. *Intervention in School and Clinic, 41*(3), 169-174. https://doi.org/10.1177/10534512060410030701

Cirino, P. T., Romain, M. A., Barth, A. E., Tolar, T. D., Fletcher, J. M., & Vaughn, S. (2013). Reading skill components and impairments in middle school struggling readers. *Reading and Writing*, *26*(7), 1059-1086. https://doi.org/10.1007/s11145-012-9406-3

Cortiella, C., & Horowitz, S. H. (2014). *The state of learning disabilities: Facts, trends and emerging issues*. National Center for Learning Disabilities.

Darling-Hammond, L., Hyler, M. E., & Gardner, M. (2017). *Effective teacher professional development*. Learning Policy Institute. https://learningpolicyinstitute.org/product/teacher-prof-dev

Eppley, K., Azano, A. P., Brenner, D. G., & Shannon, P. (2018). What counts as evidence in rural schools? Evidence-based practice and practice-based evidence for diverse settings. *The Rural Educator*, *39*(2), 36-40. https://doi.org/10.35608/ruraled.v39i2.208

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers*, *28*(1), 1-11. https://doi.org/10.3758/BF03203630

Fullan, M. (2005). *Leadership and sustainability: System thinkers in action*. Corwin Press.

Galuschka, K., Ise, E., Krick, K., & Schulte-Korne, G. (2014). Effectiveness of treatment approaches for children and adolescents with reading disabilities: A meta-analysis of randomized controlled trials. *PLOS ONE*, *9*(2), Article e89900. https://doi.org/10.1371/journal.pone.0089900

Hamann, E. T., & Meltzer, J. (2005). *Multi-party mobilization for adolescent literacy in a rural area: A case study of policy development and collaboration*. Education Alliance at Brown University. https://www.brown.edu/academics/education-alliance/publications/multi-party-mobilization-adolescent-literacy-rural-area-case-study-policy-development-a

Hammill, D. D., Wiederholt, J. L., & Allen, E. A. (2006). *TOSCRF: Test of silent contextual reading fluency: Examiner's manual*. Pro-Ed.

Hammill, D. D., Wiederholt, J. L., & Allen, E. A. (2014). *TOSCRF-2: Test of silent contextual reading fluency – Second edition: Examiner's manual*. Pro-Ed.

Hock, M. F., Brasseur, I. F., & Deshler, D. D. (2008). *Fusion reading*. McGraw-Hill Education.

Hock, M. F., Brasseur-Hock, I. F., & Deshler, D. D. (2012). *Possible selves for readers.* McGraw-Hill Education.

Hock, M. F., Bulgren, J. A., & Brasseur-Hock, I. F. (2017). The strategic instruction model: The less addressed aspects of effective instruction for high school students with learning disabilities. *Learning Disabilities Research and Practice, 32*(3), 166-179. https://doi.org/10.1111/ldrp.12139

Hock, M. F., Brasseur-Hock, I. F., Hock, A. J., & Duvel, B. (2017). The effects of a comprehensive reading program on reading outcomes for middle school students with disabilities. *Journal of Learning Disabilities*, *50*(2), 195-212. https://doi.org/10.1177/0022219415618495

Hock, M. F., Brasseur, I. F., Deshler, D. D., Catts, H. W., Marquis, J. G., Mark, C. A., & Stribling, J. W. (2009). What is the reading component skill profile of adolescent struggling readers in urban schools? *Learning Disability Quarterly*, *32*(1), 21-38. https://doi.org/10.2307/25474660

Holloway, D. (2002). Using research to ensure quality teaching in rural schools. *Journal of Research in Rural Education*, *17*(3), 138-153.

Institute of Education Sciences. (2006). *IES; PR/Award R305G040011.* U.S. Department of Education.

Knight, J. (2007). *Instructional coaching: A partnership approach to improving instruction*. Corwin Press.

Knight, J. (Ed.). (2009). *Coaching: Approaches and perspectives*. Corwin Press.

Knight, J., & van Nieuwerburgh, C. (2012). Instructional coaching: A focus on practice. *Coaching: An International Journal of Theory, Research and Practice, 5*(2), 100-112. https://doi.org/10.1080/17521882.2012.707668

Knight, D. S., Hock, M. F., & Knight, J. (2016). Designing instructional coaching. In C. M. Reigeluth, B. J. Beatty, & R. D. Meyers (Eds.), *Instructional-design theories and models, The learner-centered paradigm of education, Vol. IV* (pp. 269-286). Routledge.

Kurz, A., Reddy, L. A., & Glover, T. A. (2017). A multidisciplinary framework of instructional coaching. *Theory into Practice*, *56*(1), 66-77. https://doi.org/10.1080/00405841.2016.1260404

McDonald, J. P., Klein, E. J., & Riordan, M. (2009). *Going to scale with new school designs: Reinventing high school*. Teachers College Press.

Miciak, J., Stuebing, K. K., Vaughn, S., Roberts, G., Barth, A. E., & Fletcher, J. M. (2014). Cognitive attributes of adequate and inadequate responders to reading intervention in middle school. *School Psychology Review*, *43*(4), 407-427. https://doi.org/10.1080/02796015.2014.12087413

Murphy, K. R., & Myors, B. (2004). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests* (2nd ed.). Lawrence Erlbaum.

National Center for Education Statistics. (2019). *2019 NAEP report card: Reading*. U.S. Department of Education, Institute of Education Sciences. Retrieved from https://www.nationsreportcard.gov/reading?grade=8

Northwest Evaluation Association. (2011). *Technical manual for Measures of Academic Progress (MAP) and Measures of Academic Progress for Primary Grades (MPG).* Author.

Palincsar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction*, *1*(2), 117-175. https://doi.org/10.1207/s1532690xci0102_1

Peine, E. K., Azano, A. P., & Schafft, K. A. (2020). Beyond cultural and structural explanations of regional underdevelopment: Identity and dispossession in Appalachia. *Journal of Appalachian Studies*, *26*(1), 40-56. https://doi.org/10.5406/jappastud.26.1.0040

Richardson, J.T.E. (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review*, *6*(2), 135-147. https://doi.org/10.1016/j.edurev.2010.12.001

Scammacca, N. K., Roberts, G., Vaughn, S., & Stuebing, K. K. (2015). A meta-analysis of interventions for struggling readers in grades 4–12: 1980–2011. *Journal of Learning Disabilities*, *48*(4), 369-390. https://doi.org/10.1177/0022219413504995

Scammacca, N., Roberts, G., Vaughn, S., Edmonds, M., Wexler, J., Reutebuch, C. K., & Torgesen, J. K. (2007). *Interventions for adolescent struggling readers: A meta-analysis with implications for practice*. RMC Research Corporation, Center on Instruction.

Shippen, M. E., Miller, A., Patterson, D., Houchins, D. E., & Darch, C. B. (2014). Improving adolescent reading skills in rural areas using evidence-based practices. *Rural Special Education Quarterly*, *33*(2), 12-17. https://doi.org/10.1177/875687051403300203

Showers, B., Joyce, B., Scanlon, M., & Schnaubelt, C. (1998). A second chance to learn to read. *Educational Leadership*, *55*(6), 27-30.

Slavin, R. E., Cheung, A., Groff, C., & Lake, C. (2008). Effective reading programs for middle and high schools: A best-evidence synthesis. *Reading Research Quarterly*, *43*(3), 290-322. https://doi.org/10.1598/RRQ.43.3.4

Solis, M., Ciullo, S., Vaughn, S., Pyle, N., Hassaram, B., & Leroux, A. (2012). Reading comprehension interventions for middle school students with learning disabilities: A synthesis of 30 years of research. *Journal of Learning Disabilities*, *45*(4), 327-340. https://doi.org/10.1177/0022219411402691

Swanson, H. L. (1999). Instructional components that predict treatment outcomes for students with learning disabilities: Support for a combined strategy and direct instruction model. *Learning Disabilities Research & Practice*, *14*(3), 129-140. https://doi.org/10.1207/sldrp1403_1

Torgesen, J., Myers, D., Schirm, A., Stuart, E., Vartivarian, S., Mansfield, W., Stancavage, F., Durno, D., Javorsky, R., & Haan, C. (2006). *National assessment of Title I: Interim report. Vol. II: Closing the reading gap: First year findings from a randomized trial of four*

reading interventions for striving readers. U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Torgesen, J. K., Houston, D. D., Rissman, L. M., Decker, S. M., Roberts, G., Vaughn, S., Wexler, J., Francis, D. J., Rivera, M. O., & Lesaux, N. (2007). *Academic literacy instruction for adolescents: A guidance document from the Center on Instruction*. RMC Research Corporation, Center on Instruction.

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley Publishing Company.

Vaughn, S., & Wanzek, J. (2014). Intensive interventions in reading for students with reading disabilities: Meaningful impacts. *Learning Disabilities Research and Practice, 29*(2), 46-53. https://doi.org/10.1111/ldrp.12031

Vaughn, S., Swanson, E., & Solis, M. (2013). Reading comprehension for adolescents with significant reading problems. In H. Swanson, K. Harris, & S. Graham (Eds.), *Handbook of learning disabilities* (2nd ed., pp. 373-387). Guilford.

Vaughn, S., Wexler, J., Leroux, A., Roberts, G., Denton, C., Barth, A., & Fletcher, J. (2012). Effects of intensive reading intervention for eighth-grade students with persistently inadequate response to intervention. *Journal of Learning Disabilities, 45*(6), 515-525. https://doi.org/10.1177/0022219411402692

Virginia Department of Education. (2017a). *School quality profiles*. Retrieved from https://schoolquality.virginia.gov

Virginia Department of Education. (2017b). *Standards of learning (SOL) & testing*. Retrieved from https://www.doe.virginia.gov/testing/index.shtml

Wagner, R. K., Torgesen, J. K., Rashotte, C. A., & Pearson, N. A. (2010). *Test of silent reading efficiency and comprehension (TOSREC), examiner's manual*. Pro-Ed.

Wiederholt, J. W., & Bryant, B. R. (2012). *GORT-5: Gray oral reading tests-fifth edition, examiner's manual.* Pro-Ed.

Williams, K. T. (2001). *Group reading assessment and diagnostic evaluation: Teacher's scoring & interpretive manual*. American Guidance Service.

Wright, T. S., & Cervetti, G. N. (2017). A systematic review of the research on vocabulary instruction that impacts text comprehension. *Reading Research Quarterly, 52*(2), 203- 226. https://doi.org/10.1002/rrq.163

CBM Language Measures as Indicators of Foreign-Language Learning: Technical Adequacy of Scores for Secondary-School Students

Laura Hoefnagel, Christine A. Espin, and Ralph Rippe
Leiden University

Abstract

Students with and without learning disabilities often struggle to learn a foreign language (FL). Teachers could benefit from a measure designed to screen and identify students at risk for FL learning difficulties. In this study, we examined the reliability and validity of scores from four curriculum-based measures (CBM) as potential indicators of English FL learning: reading aloud, maze selection, and English-to-Dutch and Dutch-to-English word translation. Participants were 133 Dutch students in Grade 8. Criterion variables were English course grades and scores on a standardized achievement test (Cito-VAS). Alternate-form reliability ranged from r = .77 to .87. Correlations between CBM and criterion measure scores ranged from r = -.04 to .65. Scores from maze selection and reading aloud alone predicted English-language proficiency better than a combination of scores from the four measures, explaining 29.7% and 23.6% of the variance, respectively. Implications for the use of CBM for FL screening and progress-monitoring are discussed.

Keywords: Foreign-language learning, curriculum-based measurement, progress monitoring, technical adequacy, secondary school

In our globalized society, mastering languages other than one's native language is essential. For example, many universities in the United States require foreign-language credits for admission and/or graduation (see Campus Explorer, 2019; Grove, 2019). In 2016, more than 1.4 million students were enrolled in foreign-language courses at institutions of higher education in the U.S. (Looney & Lusin, 2018). In Europe, the ambition is to have 75% of young citizens master two foreign languages (Dutch Education Council, 2008). In 2015, 98.6% of lower secondary-level students in the European Union studied at least one foreign language, of which English was by far the most common (Eurostat, 2017).

Difficulties in Foreign-Language Learning

Although many students learn a foreign language without difficulty, others struggle. One of the best predictors of foreign-language (FL) learning ability is native language ability (Domínguez de Ramírez & Shapiro, 2007; Ganschow et al., 1998; Sparks, 2008; Sparks et al., 2006). It is perhaps not surprising, then, that students with learning disabilities (LD) often are considered to be at risk for FL learning difficulties (Skinner & Smith, 2011), and are granted waivers or substitutions for FL courses. However, not all students with LD experience difficulties with FL learning, and not all students who have difficulties with FL learning have LD (DiFino & Lombardino, 2004; Sparks, 2006, 2016; Wight, 2015).

Sparks (2006, 2009) suggested that FL learning ability be viewed as occurring along a continuum, and that identification of students at risk for FL difficulties be made on the basis of performance rather than labels. Students who are identified as being at risk could then be monitored and, if necessary, provided specialized teaching methods and accommodations to enhance their FL learning (Skinner & Smith, 2011; Sparks et al., 2002). A tool that could be used to screen and identify

Hoefnagel, L., Espin, C. A., & Rippe, R. (2021). CBM Language Measures as Indicators of Foreign-Language Learning: Technical Adequacy of Scores for Secondary-School Students. *International Journal for Research in Learning Disabilities, 5*(1), 42-57. https://doi.org/10.28987/ijrld.5.1.42

at-risk students, monitor their progress, and evaluate the effectiveness of specialized methods and accommodations is Curriculum-Based Measurement (CBM).

Curriculum-Based Measurement

CBM is a simple procedure for repeated measurement of student growth toward long-range instructional goals in academic areas (Deno, 1985). Using CBM, teachers measure student progress on a frequent basis (e.g., once a week) using brief samples of work, and place the scores on a graph that depicts progress. They subsequently examine the progress graph to determine the effectiveness of instruction.

CBM measures are designed to be practical (simple, time efficient, easy to administer and score) and to produce scores that serve as valid and reliable indicators of performance and progress in an academic area (Deno, 1985; Espin & Deno, 2016). A considerable body of research has examined the validity and reliability of scores from CBM measures in reading and writing (see McMaster & Espin, 2007; Wayman et al., 2007), but this research has primarily been conducted with students in their native language. And while research has been carried out on the development of CBM measures for English Learners (EL; e.g., Baker & Good, 1995; Campbell et al., 2013; Domínguez de Ramírez & Shapiro, 2006, 2007; Sandberg & Reschly, 2011), the findings cannot automatically be generalized to FL learning as the situations under which EL and FL students learn a second language differ.

In searching the literature, we located only one study that examined CBM measures of FL learning (Chung & Espin, 2013). Chung and Espin examined the technical adequacy of scores from maze selection, Dutch-to-English word translation, and English-to-Dutch word translation, both alone and in combination, as indicators of FL learning for middle-school students. For each measure, different time frames and scoring procedures were compared. Criterion variables in the study were English course grades and scores on a standardized English reading test. Results varied somewhat across grade and skill level but provided tentative support for scores from maze-selection (2 min, correct minus incorrect choices) and word-translation measures (English-to-Dutch- or Dutch-to-English, 2 min, correct translations) as indicators of FL performance. In addition, results demonstrated that a combination of scores from maze and English-to-Dutch word translation accounted for a greater proportion of variance in English course grades than scores from either measure alone.

Despite these important findings, the Espin and Chung study (2013) has limitations. First, the study did not include a reading-aloud measure. CBM reading aloud often is used to monitor progress in one's native language (Wayman et al., 2007). Second, for some of the analyses, sample sizes were small because different CBM and criterion measures were used across grade and educational levels. Finally, there was a ceiling effect for the maze scores.

The Present Study

Given the importance of FL learning in today's globalized society, and the number of students who struggle to learn an FL, it seems important to replicate the Chung and Espin (2013) study, addressing the limitations of the study wherever possible.

The present study was a replication and extension of Chung and Espin (2013). Specifically, the study examined the reliability and validity of scores from four CBM measures, alone and in combination, as potential indicators of FL learning. The four measures were maze selection, Dutch-to-English word translation, English-to-Dutch word translation, and reading aloud. To avoid some of the limitations of the Chung and Espin study, we increased the length of the maze passages to avoid ceiling effects and, to the extent possible, used identical measures across educational levels.

Two research questions were addressed in the study:

1. What are the reliability and validity of scores from four CBM measures as potential indicators of English FL performance?
2. Does a combination of scores predict English FL performance better than scores from a single measure?

Based on the results of Chung and Espin (2013), we expected that scores from maze-selection and word-translation measures (both English-to-Dutch and Dutch-to-English) would be reliable and would significantly relate to scores on the criterion measures. Further, we expected that a combination of scores from maze selection and English-to-Dutch word translation would account for a greater proportion of variance in the criterion variables than scores from either measure alone. Because Chung and Espin (2013) did not include a reading-aloud measure, we had no expectations related to scores from the reading-aloud measures.

The present study can be characterized as Stage 1 CBM research (Fuchs, 2004), where the focus is on the technical adequacy of scores as indicators of performance. Given the focus on technical adequacy, partici-

pants in our study represented a range of performance levels. Results provide information on the extent to which the CBM scores accurately rank order students on their English FL performance, and have implications for the use of the measures to screen and identify at-risk students. The findings also inform future Stage 2 research, where the focus is on the use of the measures for progress monitoring.

Method

Participants

Participants were 133 eighth-grade students (67 males, 66 females; M_{age} = 13.60, SD = .69, age range 12–16 years) from 15 classrooms in three secondary schools in The Netherlands. Schools were located in three middle-large to large cities in the west and central part of the country. Schools were selected via the researchers' networks. Participants were recruited via their English-language courses. All students were invited to participate.

Secondary education in The Netherlands is divided into different educational levels, which are (from the lowest to highest): vocational-low, vocational-high, professional, and university preparation. English as a FL is mandatory in all Dutch secondary schools, and the national curriculum for English consists of reading, writing, listening, and speaking. The curriculum follows the Common European Framework of Reference for Languages (CEFR), with CEFR target levels set for the end of secondary school for each educational level (www.erk.nl).

Participants in the study were in their second year of formal English education and represented all educational levels: vocational-low (12.8%), vocational-high (11.3%), professional (16.5%), university (36.1%), and combined professional/university (23.3%) levels. Fourteen students (10.5%) were enrolled in a bilingual education program in which at least 50% of core courses were provided in English for the first three years of secondary school. These students were from the university education levels. Home-language information was available for 40% of the students. For all these students, Dutch was spoken at home.

Fifteen percent of participants were students with dyslexia. The diagnosis of dyslexia in The Netherlands is based on significant delays in reading and/or spelling, despite systematic and frequent intervention, where delays are not due to an intellectual or sensory disability, or to inadequate education (De Jong et al., 2016). The prevalence of dyslexia is estimated to be 3.6% at the end of primary school; that is, sixth grade (Blomert, 2005). However, the Dutch Inspectorate of Education (2019) reports that the percentage of students actually labeled with dyslexia in sixth grade is 7.5%. This number increases sharply at secondary school, to 11.9% and 13% of seventh- and ninth-grade students, respectively. These percentages are similar to the 15% of students with dyslexia in the current sample.

Predictor Variables

Predictor variables were scores from four CBM FL measures: reading aloud, maze selection, English-to-Dutch word translation, and Dutch-to-English word translation.

Reading Aloud

Reading-aloud passages were two English narrative texts selected from Children's Educational Services passages (Deno & Marston, 1987). Passages were 488 and 498 words in length, written for students in Grade 4, and were non-culturally specific. Students read aloud from each passage for 1 min, whereupon the number of correct (WRC) and correct minus incorrect (WRCI) words read were scored. Incorrect words included mispronunciations, word substitutions, omissions, reversals, and words supplied by the examiner when a student did not know a word.

Maze Selection

Maze selection passages were constructed from the same English narrative texts used for reading aloud to minimize differences in results due to text effects. To create the maze, the first sentence was left intact, after which every seventh word was deleted and replaced by the correct word and two distractors. The three choices were placed in bold print and underlined in the text, and were not split across lines. If the seventh word was a proper noun, it was left and, instead, the next word was deleted. The placement of the correct choice varied. Distractors were approximately equal in length (within one letter) to the correct choice, and were clearly incorrect (for guidelines, see Conoyer et al., 2017; Fuchs & Fuchs, 1992). Students read each maze text silently for 2 min, circling the word that restored meaning to the passage. Scores were the number of correct (MCC) and correct minus incorrect maze (MCCI) choices. Scoring was carried out with and without a guessing rule. With the guessing rule, scoring was stopped after three consecutive incorrect choices.

CBM maze selection is similar to the modified- or multiple-choice cloze measures often used in FL assessment (Hale et al., 1989; Porter, 1976) with one key difference. In typical FL modified-cloze measures, distractors are similar in meaning and syntax to the target word (Porter, 1976). In CBM mazes, on the other hand, distractors are selected to be clearly different in meaning and syntax from the target word so that one answer is obviously correct (Fuchs & Fuchs, 1992). Chung and Espin (2013) reported alternate-form reliability for CBM maze scores ranging from r = .69 to .78 and validity from r = .20 to .79, with higher reliabilities reported for MCCI than for MCC.

Word-Translation Measures

Dutch-to-English and English-to-Dutch word-translation measures consisted of a list of 50 words (25 words per page) with a blank next to each word. Words were randomly selected from an English-language curriculum used in Dutch secondary schools. All parts of speech were represented on each measure. Students wrote as many translations as possible in 2 min. Scores were the number of correct (WTC) and correct minus incorrect (WTCI) translations.

Based on the results of Chung and Espin (2013), a decision was made to count translations as correct only if they were spelled correctly. Chung and Espin (2013) reported alternate-form reliability for word translation scores ranging from r = .76 to .88 for WTC and from r = .59 to .78 for WTCI, and validity from r = .44 to .77 for WTC. Validity for WTCI was not examined in Chung and Espin (2013) because the reliabilities were low.

Criterion Variables

Criterion variables in the study were English course grades and scores on a standardized English-language test (Cito-VAS).

English Course Grades

English course grades were average grades across three grading periods in the school year. Grades were based on the students' performance in reading, listening, writing, speaking, vocabulary, and grammar within the individual student's educational level. Grades ranged from 1 (low) to 10 (high), and were reported to one decimal point. A grade of 5.5 was passing. Grades are assigned within educational level; thus, a grade of 7 in English in a vocational-low level program was not equivalent to a grade of 7 in a university-preparation level program. Analyses involving grades were, therefore, carried out within educational level.

Cito-VAS Scores

The Cito-VAS test (Cito, 2015) is a standardized achievement test administered in many Dutch secondary schools in the middle of the school year. In our study, two of the three participating schools administered the Cito-VAS. (The school with students in combined professional/university levels did not administer the test.) Scores from the English reading and English vocabulary subtests of the Cito-VAS were used in the study.

The English reading subtest consisted of expository passages, each with 1-3 multiple-choice questions, for a total of 35 questions. The English vocabulary subtest consisted of multiple-choice items in which students had to choose (a) the correct Dutch translation of the underlined word in an English sentence, (b) the correct English word to complete a sentence, (c) a synonym or an antonym for an English word, or (d) the word that did not belong in a set of words. The vocabulary subtest included a total of 45 items. Each subtest took approximately 50 minutes to complete. Different forms of the Cito-VAS were administered at different educational levels. Therefore, standard scores were used in the analysis, enabling comparisons across test levels.

Technical adequacy information for the Cito-VAS was available only for an earlier version of the test that did not include the English vocabulary subtest (Van Til & Van Boxtel, 2015). Cronbach's alphas for the English reading subtest were reported to be .76, .78 and .80. With regard to validity, a consistent increase in mean scores across grade and educational levels was reported, and correlations between subtests measuring different constructs were found to be weaker (r = .25 - .42) than between subtests measuring overlapping constructs (r = .58 to .72). Finally, standard scores on the English reading subtest for eighth-grade students were found to predict educational level one year later (r = .63).

Procedure

Participants completed the measures in the following order: maze selection, English-to-Dutch word translation, Dutch-to-English word translation, and reading aloud. For all CBM measures, two parallel forms were administered, with the order of the forms counterbalanced. Maze-selection and word-translation measures were administered in a group setting. Reading aloud was administered individually on the same day or within the same week. If a student was absent for part of the data collection, every attempt

was made to schedule a make-up session. Data were collected and scored by four master's-level students who were trained in two 1.5-hour training sessions. Two data collectors were present for all data collection. English course grades, Cito-VAS scores, and student background information were collected from the schools at the end of the school year.

Scoring

All measures were scored by two coders. Interscorer agreement was calculated by dividing the smaller by the larger score and multiplying by 100. Agreement was calculated separately for each score. For reading aloud, agreement was 99.2% (WRC) and 98.9% (WRCI). For maze selection, agreement was 99.8% (MCC) and 99.8% (MCCI). For English-to-Dutch word translation, agreement was 99.0% (WTC) and 97.4% (WTCI). For Dutch-to-English word translation, agreement was 98.2% (WTC) and 94.7% (WTCI). Disagreements were discussed and resolved before coming to a final score.

Results

Data Inspection

Data inspection indicated normal distributions for all independent variables and no substantial univariate outliers. To check for bivariate outliers, multivariate scatterplots were inspected. The patterns in the scatterplots revealed approximately linear associations between the independent and dependent variables. One possible bivariate outlier was detected in nearly every scatterplot. For this student, who was at the university preparation level and was diagnosed with dyslexia, Cito-VAS scores and English course grades were relatively high, whereas scores on the CBM measures were relatively low. Removal of this outlier yielded a change in explained variances (for example from $R^2 = .35$ to $R^2 = .42$ for the relation between maze selection MCC and Cito-VAS scores). Because of the disproportionally large effects of the student's scores on the strength of the correlations, analyses were conducted both with and without the outlier. The association patterns were the same with and without the outlier, but the results were somewhat stronger (i.e., correlation coefficients increased) when the outlier was removed. We report results *without* the outlier for the validity analyses.

Handling of Missing Data and Assumptions

Patterns of missing observations were checked. Little's MCAR test (Little, 1988) showed that no patterns in missingness could be detected; $\chi 2(8, N = 133) = 5.83$, $p = .666$; therefore, any missingness was considered to be completely at random. Analyses were based on full information maximum likelihood (FIML) estimation (Graham et al., 1996), with which missingness is commonly handled within the analysis model (Dempster et al., 1977) as it yields the most likely parameter values given all available data in the model, regardless of their level of completeness.

Descriptive Analyses

Means and standard deviations for scores on the CBM measures (alternate forms and combined) are reported in Table 1. On average, students read aloud approximately 150 correct and 5 incorrect words in 1 min, made approximately 22 correct and 0.5 to 1.0 incorrect maze choices in 2 min (depending on whether a guessing rule was applied or not), translated approximately 25 words correct and 4 incorrect from English to Dutch and approximately 19 words correct and 5 incorrect from Dutch to English. Means and standard deviations broken down by gender are reported in Table 2. Girls tended to score higher on the CBM measures than boys, but differences were not large.

There were significant differences in mean scores between Forms A and B for reading aloud (WRC, $t(121) = 6.55$, $p < .001$; WRCI, $t(121) = 6.42$, $p < .001$) and English-to-Dutch translation (WTC, $t(129) = 11.42$, $p < .001$; WTCI, $t(129) = 9.72$, $p < .001$; Bonferroni correction applied), but not for maze selection or Dutch-to-English translation. Further, no significant differences in mean scores were found between scoring with or without use of a guessing rule for the maze.

Means and standard deviations for Cito-VAS English vocabulary and English reading subtests, broken down by educational level, are reported in Table 3. (Recall that scores were not available for one school.) Means and standard deviations for the English course grades are reported by educational level in Table 3. Individual grades ranged from 4.30 to 9.27 ($M = 7.14$, $SD = 1.03$).

Table 1
Means and Standard Deviations of the CBM Measures Form A, Form B, and Mean of A and B

	Form A		Form B		Mean (A+B)	
Measure / Score	*N*	*M (SD)*	*N*	*M (SD)*	*N*	*M (SD)*
Reading Aloud						
WRC	122	155.46 (37.72)	122	144.29 (31.75)	122	149.87 (33.56)
WRCI	122	150.43 (39.46)	122	139.22 (33.08)	122	144.82 (35.11)
Maze, guessing rule						
MCC	131	21.91 (8.70)	132	21.96 (7.53)	132	21.88 (7.84)
MCCI	131	21.35 (9.00)	132	21.36 (7.75)	132	21.30 (8.10)
Maze, no guessing rule						
MCC	131	22.34 (7.99)	132	22.30 (7.04)	132	22.26 (7.24)
MCCI	131	21.37 (8.97)	132	21.31 (7.91)	132	21.28 (8.13)
English-to-Dutch						
WTC	130	26.68 (7.08)	130	23.15 (5.71)	130	24.92 (6.19)
WTCI	130	23.33 (8.42)	130	18.80 (6.71)	130	21.05 (7.13)
Dutch-to-English						
WTC	131	19.70 (8.61)	131	18.63 (8.16)	131	19.17 (8.09)
WTCI	131	14.16 (9.92)	131	13.30 (9.63)	131	13.73 (9.31)

Note. WRC = words read correct. WRCI = words read correct minus incorrect. MCC = maze choices correct. MCCI = maze choices correct minus incorrect. WTC = words translated correct. WTCI = words translated correct minus incorrect.

Table 2
Means and Standard Deviations of the CBM Measures by Gender

	Males		Females	
Measure / Score	*N*	*M (SD)*	*N*	*M (SD)*
Reading Aloud				
WRC	63	148.58 (35.27)	58	151.76 (31.92)
WRCI	63	143.67 (36.52)	58	146.51 (33.92)
Maze, guessing rule				
MCC	66	20.95 (8.43)	65	22.88 (7.17)
MCCI	66	20.31 (8.57)	65	22.35 (7.54)
Maze, no guessing rule				
MCC	66	21.65 (7.42)	65	22.94 (7.08)
MCCI	66	20.27 (8.65)	65	22.36 (7.54)
English-to-Dutch				
WTC	64	24.17 (6.52)	65	25.71 (5.82)
WTCI	64	20.44 (7.38)	65	21.70 (6.92)
Dutch-to-English				
WTC	65	18.45 (7.90)	65	19.95 (8.31)
WTCI	65	13.48 (8.70)	65	14.01 (10.01)

Note. WRC = words read correct. WRCI = words read correct minus incorrect. MCC = maze choices correct. MCCI = maze choices correct minus incorrect. WTC = words translated correct. WTCI = words translated correct minus incorrect.

Table 3
Means and Standard Deviations for Cito-VAS English Vocabulary and English Reading Subtests and for English Course Grades by Educational Level

		Cito-VAS Vocabulary	Cito-VAS Reading	English course grades	
Educational level	*N*	*M (SD)*	*M (SD)*	*N*	*M (SD)*
Vocational-low	11	152.91 (21.49)	136.36 (17.83)	17	7.27 (.79)
Vocational-high	15	173.83 (31.59)	155.19 (17.23)	15	7.03 (.72)
Professional	21	173.95 (28.46)	150.95 (14.41)	22	6.64 (1.13)
Professional/university	-	-	-	31	6.57 (.95)
University	24	195.47 (32.66)	169.80 (19.69)	48	7.73 (.88)
Total	71	177.94 (32.55)	155.96 (20.62)	133	7.14 (1.03)

Note. WRC = words read correct. WRCI = words read correct minus incorrect. MCC = maze choices correct. MCCI = maze choices correct minus incorrect. WTC = words translated correct. WTCI = words translated correct minus incorrect.

Reliability Analyses

To assess alternate-form reliability, Pearson correlations between scores on parallel forms of each measure were computed. Reliability coefficients ranged from r = .77 to .87, with all but one coefficient (English-to-Dutch translation) above .82 (see Table 4). All correlations were statistically significant, with all p-values < .001. Alternate-form reliability coefficients were high despite significant mean differences between Forms A and B for reading aloud and English-to-Dutch translation, indicating that, even though students scored higher on Form A than on Form B, the rank ordering of students remained similar across the forms. Mean scores across Forms A and B were used for the subsequent validity analyses to increase the stability of the scores.

Validity Analyses

To reduce the number of statistical tests, a limited number of scores were carried forward for the validity analysis. Selection of scores was based on the reliability coefficients, the efficiency of scoring procedure, and on whether the scoring procedure was typically used in other CBM research. The following scores were selected for the validity analysis: WRC for reading aloud, MCCI with use of a guessing rule for maze selection, and WTC for both word-translation measures. Mean scores across forms A and B were used for all analyses.

Correlations With Criterion Variables

Correlations between CBM scores and the Cito-VAS scores were statistically significant, ranging from r = .31 to .65 (see Table 5). In general, correlations for reading aloud and maze selection were higher than for the translation tasks; the lowest correlations were found for English-to-Dutch translation. Correlations tended to be somewhat higher with the Cito-VAS reading subtest than the vocabulary subtest, but differences were small.

Correlations with English course grades were computed within educational level, resulting in samples ranging from 14 to 48 students per subgroup. Means and standard deviations for the CBM scores, broken down by educational level, are reported in Table 6. In general, as illustrated, mean scores increased across educational level although scores for combined professional/university were higher than for university only. Correlations between CBM scores and English course grades ranged from r = -.04 to .65 (see Table 7). Across educational levels, correlations tended to be lowest for the English-to-Dutch translation, but patterns for the other measures differed somewhat. For example, for vocational and professional educational levels, coefficients tended to be higher for reading aloud and maze selection than for word-translation measures, and for professional/university and university levels, coefficients for word-translation measures were as high as or higher than for reading aloud and maze.

Regression Analyses

To examine whether a combination of measures predicted English-language proficiency better than a single measure, latent linear regression models were tested in different stages, evaluating performance through different compositions of a latent performance score. Multilevel models revealed negligible intra-class correlations on educational level (ICCs for average Cito scores = .06, .03 for Cito vocabulary, and .08 for Cito

Table 4
Alternate-Form Reliability Coefficients of Scores From the CBM Measures

	Reading Aloud				Maze selection	
Scoring	*N*	*r*	Guessing rule	Scoring	*N*	*r*
WRC	122	.87	Rule	MCC	131	.85
WRCI	122	.87		MCCI	131	.85
			No rule	MCC	131	.83
				MCCI	131	.84

English-to-Dutch translation			Dutch-to-English translation		
Scoring	*N*	*r*	Scoring	*N*	*r*
WTC	130	.87	WTC	131	.86
WTCI	130	.77	WTCI	131	.82

Note. All correlations significant at $p < .001$ level.

Table 5
Correlations Between CBM Scores and Cito-VAS Scores

Measure	Cito Vocabulary	Cito Reading
Reading aloud WRC (*N* = 65)	.56***	.65***
Maze selection MCCI (*N* = 69)	.63***	.63***
English-to-Dutch WTC (*N* = 68)	.31*	.34**
Dutch-to-English WTC (*N* = 69)	.50***	.52***

* $p < .05$. ** $p < .01$. *** $p < .001$.

Note. WRC = words read correct; MCCI = maze choices correct minus incorrect; WTC = words translated correctly.

Table 6
Means and Standard Deviations for Selected CBM Scores by Educational Level

Educational level	WRC	MCCI	WTC E-D	WTC D-E
Vocational-low	120.00 (31.57)	11.32 (7.86)	16.88 (5.22)	11.88 (5.85)
Vocational-high	139.23 (33.53)	17.11 (8.78)	19.62 (5.25)	13.36 (6.59)
Professional	133.89 (31.11)	22.52 (6.93)	26.36 (4.08)	17.02 (6.18)
Combined professional/university	169.69 (25.95)	25.33 (5.81)	28.15 (5.14)	24.31 (6.94)
University	152.07 (28.49)	22.23 (6.10)	25.47 (4.65)	19.02 (7.44)
Total	150.15 (33.56)	21.42 (8.06)	24.93 (6.20)	19.12 (8.09)

Note. WRC = words read correct; MCCI = maze choices correct minus incorrect; WTC E-D = words translated correct, English-to-Dutch; WTC D-E = words translated correct, Dutch-to-English. Standard deviations are in parentheses.

Table 7
Correlations Between CBM Scores and Average English Grades Within Educational Level

	Reading aloud WRC		Maze selection MCCI	
Educational level	*N*	*r*	*N*	*r*
Vocational low	16	.48	17	.57*
Vocational high	14	.65*	15	.54*
Professional	18	.51*	21	.63**
Professional/university	30	.38*	30	.40*
University	43	.26	47	.54***
	English-to-Dutch WTC		Dutch-to-English WTC	
Educational level	*N*	*r*	*N*	*r*
Vocational low	16	.43	16	.47
Vocational high	14	.35	15	.44
Professional	21	-.04	21	.30
Professional/University	31	.46*	31	.50**
University	48	.30*	47	.53***

* $p < .05$. ** $p < .01$. *** $p < .001$.

Note. WRC = words read correct; MCCI = maze choices correct minus incorrect; WTC = words translated correctly.

reading, respectively); thus, analyses did not account for variance attributable to educational-level characteristics (Luke, 2004).

The approach used to test the linear regression models was as follows. First, we compared a model in which performance was indicated by the Cito-VAS subtest scores only, hereafter Cito-Only Model, to a model in which a latent performance score was indicated by both the Cito-VAS subtest scores and the English course grades, hereafter Gold Standard Model, under planned missing data assumptions (Little & Rhemtulla, 2013; Rippe & Merkelbach, 2019). In the latter approach, latent performance scores for individuals with missing data on the Cito-VAS subtests were approximated based on the English course grades and their correlation with the Cito-VAS subscales. These latent scores were not constructed explicitly before being entered into the model; instead, they were "estimated" implicitly within the model itself based on maximum likelihood estimations of the latent variable regression coefficients, accounting for the covariance between the (two) observed outcome scores.

As a consequence of using the latent variable approach, most classical multicollinearity measures could not be computed. Through inspection of all pairwise correlations, no indication of multicollinearity was found. All correlations were .70 or lower, with only one exception: the correlation between Dutch-to-English and English-to-Dutch word translations was .73, meaning 53% of their variance was shared. Variance inflation factors were well below 10, ranging between 2.55 and 3.33.

In the first stage, the two latent performance variants were evaluated on overall model fit only with all possible combinations of predictors. Model parameters were not interpreted. Based on overall model fit and parameter effect size, for parameter interpretation a subset of models was reevaluated as observed-variable-only models using a proportional bootstrap with 1000 samples to obtain standard errors.

For model estimations, we used Lavaan (Rosseel, 2012) version 0.6-4 in R version 3.5.3. Full information maximum likelihood was used to handle missing data. The number of EM iterations for FIML was set at a maximum of 5,000. To determine the fit of each model, the comparative fit index (CFI), normed fit index (NFI), standardized root mean residual (sRMR), and root mean square error of approximation (RMSEA) were inspected. The CFI and NFI should be as high as possible (above .90), while the sRMR and the RMSEA should be as low as possible (below .06). The RMSEA and sRMR can yield contrasting conclusions, as the sRMR is a simple absolute fit index comparing observed and predicted correlations without accounting for complexity, while the RMSEA is based on the non-centrality parameter and can be considered more precise.

Preliminary analyses revealed that there was no intra-class effect of educational level and no differential effect of gender. Therefore, educational level and gender were not accounted for in the regression analyses. Moreover, preliminary analyses showed that the models with the Gold Standard latent performance score yielded a less favorable fit-complexity ratio than the models with the Cito-Only latent performance score, indicated by the Akaike information criterion (AIC). A lower AIC relative to the other model means a better trade-off between the fit and the complexity of the model. The English course grades did not add any unique information to the latent performance score beyond the scores from the two Cito-VAS subtests. Therefore, models with the dependent latent performance score consisting only of the two Cito-VAS subtests were used in the subsequent analyses.

As a first step in the regression analysis, standard assumptions were checkedfor the subsample of students with Cito-VAS scores (N = 63). No violations on multicollinearity, normality, and homoscedasticity were found, although correlations among the predictors were high, ranging from $r = .51$ to .78 (see Table 8), suggesting caution in interpretation. s

In Stage 2, after all combinations of predictors had been compared in the Stage 1 analyses, a selection of models were reevaluated as observed-variable-only models using bootstrapped standard errors with 1,000 runs. The models were selected based on their fit. An overview of the seven models selected is presented in Table 9, with their respective fit indices provided in Table 10. The AIC allows only for comparing Models 2 through 7 to Model 1, because these models are nested. The model with all four predictors (Model 1) had the best fit in terms of complexity trade-off, as indicated by the lowest AIC value (AIC = 1,103.00; see Table 10). Even though the model consisting of all four CBM measures as predictors is the most complex, it outperformed simpler one- and two-predictor models in terms of the balance between fit of the model and complexity. The two-predictor models (Models 6 and 7) were not favored over single predictor models. For Model 6, the AIC was high. For Model 7, although the AIC was low, the RMSEA was unfavorable (.14). Therefore, neither model qualified for further interpretation. Among the simpler single-predictor models, Model 5 (Reading Aloud) resulted in the smallest difference with Model 1 in terms of AIC value (1148.72 vs. 1103.00, respectively).

Further model evaluation was based on both absolute and comparative fit using the NFI, CFI, RMSEA and sRMR values (see Table 10). As shown, all models had high values (equal to or closely approaching 1.0) on the NFI and CFI, and low values (approaching 0) on the RMSEA and sRMR, indicating good fit with little error. The all-predictor Model 1 showed somewhat poorer values on some of the indices (NFI = .97, sRMR = .01). The single-predictor models (Models 2 to 5) showed the best fit (NFI and CFI = 1.00, RMSEA and sRMR < .001).

In the above models, contributions of both the English vocabulary subtest and the English reading subtest to the latent performance score were significant ($\beta = .76$ to .83, $z = 5.67$ to 15.58, $p < .001$ and $\beta = .89$ to .97, $z = 7.48$ to 17.93, $p < .001$, respectively), with one exception: In Model 3, the contribution of the English vocabulary subtest to the latent performance score was significant ($\beta = .78$, $z = 5.67$, $p < .001$), while the English reading subtest was not significant ($\beta = .95$, $z = 1.31$, $p = .192$). This discrepancy can be explained by the fact that the models describing the same amount of variance do not necessarily describe the same part of the variance in the outcome.

In Stage 3, Models 1, 2, and 5 were selected as final models. For these models, the contribution of the predictors within the model was evaluated. The estimated regression coefficients are displayed in Table 11. In the all-predictor model (Model 1), scores from maze selection ($\beta = 0.59$, $z = 3.99$, $p < .001$) and English-to-

Table 8
Correlations Among the CBM Measures

Measure	Reading aloud WRC	Maze selection MCCI	English-to-Dutch WTC	Dutch-to-English WTC
Reading Aloud WRC	-	.70	.51	.74
Maze selection MCCI		-	.67	.71
English-to-Dutch WTC			-	.78
Dutch-to-English WTC				-

Note. All correlations significant at $p < .001$.

Note. These are correlations for the subsample of students with Cito-VAS scores ($N = 63$). WRC = words read correct; MCCI = maze choices correct minus incorrect; WTC = words translated correctly.

Dutch word translation (β = -0.34, z = -2.49, p = .012) contributed significantly to the prediction of the latent performance score (see Table 11). Removing the overlapping contribution of the four CBM measures, maze selection had the strongest unique contribution to the prediction of the latent English performance score, explaining 21.9% of the remaining total variance. The unique contribution of English-to-Dutch word translation was in the negative direction, and explained 7.3% of the variance. Thus, after accounting for the overlap between the CBM measures in the prediction of the latent English performance score, maze selection had the strongest contribution in the positive direction whereas English-to-Dutch had the next strongest contribution, but in the negative direction.

In the single-predictor model with maze selection (Model 2), scores from maze selection contributed significantly to the prediction of the latent performance score (β = .71, z = 10.41, p < .001). Maze selection explained 29.7% of the total variance in the latent performance score when the other CBM measures were not accounted for. In the single-predictor model with reading aloud (Model 5), scores from reading aloud contributed significantly to the prediction of the latent performance score (β = .66, z = 6.32, p < .001). Finally, reading aloud explained 23.6% of the total variance in the latent performance score when the other CBM measures were not accounted for.

Discussion

The goal of this study was to examine the technical adequacy of scores from four CBM measures – reading aloud, maze selection, Dutch-to-English, and English-to-Dutch word translation – as potential indicators of FL learning, replicating and extending an earlier stusdy by Chung and Espin (2013). In general, our results provided the greatest support for maze selection and reading aloud scores as general indicators of FL performance.

The first research question addressed the alternate-form reliability and validity of scores from the CBM measures. Reliability coefficients were high, with all but one coefficient falling between .82 and .87. The coefficients for maze selection and for English-to-Dutch word translation were higher than those found by Chung and Espin (2013), where coefficients were below r = .80. Differences may be due to the fact that in the present study, the same word translation measure was administered across educational levels, whereas Chung and Espin used a different form of the measure for lower and higher educational levels. Variability in scores was greater in the present study (*SDs* = 5.71 to 9.92; see Table 1) than in the Chung and Espin study (*SDs* = 2.64 to 7.24).

The effects of different scoring procedures on alternate-form reliability were also examined. For reading aloud and maze selection, reliability was not affected by scoring procedure (correct vs. correct minus incorrect or with vs. without a guessing rule). For the word translation tasks, consistent with the findings of Chung and Espin (2013), higher reliabilities were found for correct than for correct minus incorrect scores.

A select number of scores were carried forward for validity analysis: WRC for reading aloud, MCCI for maze selection, and WTC for the word translation tasks. The patterns of correlations differed across criterion measure and across educational level. For Cito-VAS, correlations could be computed across educational level. These correlations ranged from r = .31 to .65, a range similar to that reported by Chung and Espin (2013; range r = .37 to .79). Correlations with the Cito-VAS were higher for reading aloud and maze selection (r =.56 to .65) than for word translation (r = .31 to .52). For English course grades, correlations had to be computed within educational level. The patterns of results differed by educational level: (a) at lower educational levels, stronger correlations

Table 9
Final Models With the Latent Performance Score From Cito-VAS English Vocabulary and English Reading Subtests as Dependent Variable

Model	Predictors
Model 1	Maze MCCI + English-Dutch WTC + Dutch-English WTC + Reading Aloud WRC
Model 2	Maze MCCI
Model 3	English-Dutch WTC
Model 4	Dutch-English WTC
Model 5	Reading Aloud WRC
Model 6	Maze MCCI + English-Dutch WTC
Model 7	Maze MCCI + Reading Aloud WRC

Table 10
Model Fit and Complexity Trade-Off for the Selected Models

Model	*N*	AIC	NFI	CFI	RMSEA	sRMR
1	63	1103.00	.97	1.00	< .001	.01
2	69	1210.06	1.00	1.00	< .001	< .001
3	68	1225.43	1.00	1.00	< .001	< .001
4	69	1227.99	1.00	1.00	< .001	< .001
5	65	1148.72	1.00	1.00	< .001	< .001
6	68	1188.82	1.00	1.00	< .001	.01
7	64	1125.33	.98	.99	.14	.01

Table 11
Coefficients From Final Models on the Latent Performance Score

Predictors	β	*SE*	*z*	*p*	95% CI	Total variance
Model 1 (*N* = 63)						
Maze MCCI	0.59	0.15	3.99	< .001	[0.30, 0.88]	.219
English-Dutch WTC	-0.34	0.14	-2.49	.012	[-0.61, -0.07]	.073
Dutch-English WTC	0.26	0.17	1.58	.115	[-0.06, 0.59]	-
Reading Aloud WRC	0.27	0.16	1.71	.089	[-0.04, 0.58]	-
Model 2 (*N* = 69)						
Maze MCCI	0.71	0.07	10.41	< .001	[0.58, 0.85]	.297
Model 5 (*N* = 65)						
Reading Aloud WRC	0.66	0.10	6.32	< .001	[0.45, 0.86]	.236

Note. CI = confidence interval. β = standardized estimate.

were found for reading aloud and maze selection than for word translation; (b) at higher educational levels, stronger correlations were found for Dutch-to-English word translation and maze selection than for the reading aloud. These results may reflect the importance of English vocabulary knowledge at more advanced levels of English-language learning and the greater sensitivity of a reading-aloud measure for beginning learners.

Consistent across all educational levels was the finding that scores on English-to-Dutch word translation tended to result in lower correlations with the criterion variables. Similarly, Chung and Espin (2013) found low correlations between English-to-Dutch word-translation and Cito-VAS scores (although not with English course grades). The lower validity coefficients for English-to-Dutch word-translation scores might be related to the fact that students had to spell the Dutch words correctly; thus, their scores on the task reflected both English-language knowledge and Dutch spelling ability. Although students also had to spell the English words correctly, perhaps if they knew what the English word was, they also knew how to spell it. A second, more likely, explanation might be the lower variability in English-Dutch translation scores leading to an attenuation in correlations. For example, standard deviations for English-to-Dutch translation were smaller than for Dutch-to-English translation.

Even though the English-to-Dutch translation produced the smallest validity coefficients, it may be prudent to not yet discard the measure as a potential CBM FL measure. English-to-Dutch translation requires recognition rather than production, and thus might serve as a good measure for students who are just beginning to learn English.

The second research question examined whether a combination of measures predicted FL proficiency in English better than a single measure. The sample size was relatively small for this analysis (*N* = 63), and the predictors correlated with each other; thus, the results should be considered suggestive. Findings showed that

a combination of measures did not predict FL proficiency in English better than a single measure. Although all models had adequate fit, the single-predictor models had the best fit. Thus, a single measure contributed more strongly to the prediction of the latent English performance score than a combination of four or two measures. The single-predictor models with either the reading-aloud or maze-selection measure performed the best. Without the overlap with the other CBM measures, maze selection alone accounted for 29.7% of the variance in the latent performance score, and reading aloud alone for 23.6% of the variance.

These findings thus suggest that scores from maze selection are valid indicators of general FL proficiency, a finding that is in line with the results of the simple correlational analyses. Scores from maze selection accounted for nearly 30% of the variance in the latent performance score constructed from scores on the Cito-VAS vocabulary and reading subtests. Given that maze selection takes only 2 minutes to administer, whereas the Cito-VAS subtests together take 100 minutes, 30% is a substantial amount. The slight advantage of the maze selection over reading aloud may be due to the fact that maze selection requires understanding of the text passage and recognition of the words used as choices to fluently progress through the text, perhaps making it a more robust FL indicator than reading aloud. Alternately, both the maze and Cito-VAS reflect silent FL reading, whereas reading aloud also reflects speaking skills.

The differences in model performance were small. The combination of all four measures – although slightly worse than the single-predictor models – yielded good model fit indices as well. In the model with all four CBM measures, maze selection and English-to-Dutch word translation were found to make significant unique contributions to the prediction of the latent performance score. Accounting for the overlapping contribution of the four CBM measures, maze selection still made a significant unique contribution to the prediction of the latent performance score, explaining 21.9% of the remaining variance. Apparently, after accounting for the common contribution of the CBM measures, the maze-selection task measures an additional, different aspect of the construct than the other measures. English-to-Dutch word translation also made a significant unique contribution after accounting for the overlap between the four measures (7.3% of the variance), but in a negative direction. This negative unique contribution, in combination with the lower validity coefficients for English-to-Dutch translation, suggests that the measure demands skills other than FL proficiency, such as spelling in the native language.

Our results diverge from those of Chung and Espin (2013), who found that a combination of maze selection and English-to-Dutch word translation resulted in better prediction than either measure alone. The results from the present regression analyses were based on a larger sample combining all educational levels, and used a latent FL performance score. Thus, although still suggestive, they provide a basis for somewhat firmer conclusions.

It was surprising that the two measures that represented the construct of FL reading showed the highest correlations with the criterion variables and the strongest contributions as single predictors to the prediction of the latent performance score, as opposed to the measures that represented the construct of FL vocabulary knowledge. Because scores from both Cito-VAS English reading and vocabulary subtests contributed to the latent performance score, one might have expected a combination of CBM measures representing reading and vocabulary to best predict student performance. Perhaps vocabulary knowledge is an integral part of reading in the FL. That is, beginning learners need a sufficient level of vocabulary knowledge in order to read a text in the FL (Wallace, 2007). Scores on CBM reading tasks may reflect not only FL reading proficiency but also vocabulary knowledge.

In sum, the results from the regression analyses indicate that a combination of measures does not predict FL proficiency better than a single measure. Practically speaking, this is "good news" in the sense that screening and CBM progress monitoring with a single measure is less time consuming and more feasible in the classroom than using a combination of measures. Determining which single measure to use may depend on practical considerations, however. Maze selection can be administered in a group setting, whereas reading aloud must be administered individually. Thus, although maze selection is more efficient, teachers still may prefer to administer reading aloud because it provides additional information related to the students' ability to pronounce words in the foreign language.

Limitations

One limitation of the present study relates to the criterion measures used. Although course grades and the Cito-VAS have social validity in the sense that both are used to make decisions about students' promotion to the next grade, technical adequacy data on the measures were limited. Although the Cito-VAS is the most widely used standardized achievement test in Dutch secondary education, reliability and validity data were available only for a previous version of the test, and that

version did not include the vocabulary subtest (technical adequacy for the reading subtest was good; see Method section). Although course grades are probably the most commonly used indicator of performance in secondary education and have a large impact on the student's school career, course grades are largely based on teacher judgment and have a restricted range. Nevertheless, the use of both grades and standardized test scores allowed for a convergence of evidence.

A second limitation of the study relates to the sample. First, analyses involving grades had to be conducted within educational level, thereby reducing sample sizes for these analyses. Second, it was not possible to examine whether results varied by language background because native-language information was available for only 40% of students. Finally, students in the university-preparation levels were overrepresented (47.7%) and students in vocational levels underrepresented (24.1%) compared to reported national levels (19% and 55%, respectively; Dutch Inspectorate of Education, 2018). Replication of the study with a larger, more representative sample, therefore, is in order.

Implications

The results of this study have implications for the use of CBM measures in FL instruction. If the results were to be replicated with a larger and more diverse sample, it would provide support for the use of CBM maze and reading aloud as screening measures to identify students who are likely to be at risk for FL learning difficulties. Such students could be provided with additional support and instruction before they begin to fail. In addition, if future Stage 2 progress-monitoring research supports the technical adequacy of scores, the measures could be used to monitor the progress of students with severe and persistent FL learning difficulties and to evaluate the effects of specialized, individualized interventions on that progress.

The increasing need for all students to learn English in our globalized society underscores the need for related screening and progress measures. This need is further underscored by the extent to which some students struggle to learn a foreign language. For example, recall that the percentage of students with dyslexia in The Netherlands increases from 7.5% in sixth grade to 13% in ninth grade. This increase may be related to the increase in language requirements at the secondary-educational level. All Dutch secondary students must learn both Dutch and English. At higher educational levels, students are required to learn up to five FLs (English, French, German, Latin and Greek). A label of dyslexia allows for accommodations in FL learning.

Conclusion and Future Research

In conclusion, our findings support the reliability and validity of (Dutch Inspectorate of Education, 2019) scores from reading-aloud and maze selection measures (and potentially word-translation measures) as potential CBM indicators of English-language learning. Future Stage 2 research must examine the reliability and validity of the growth rates produced by scores from these measures. An important aspect of this work will be to establish the equivalence of alternate forms of the measures. This may prove to be a challenge for reading-aloud and English-to-Dutch word-translation measures, where significant mean differences in scores were found between the alternate forms. Future research also must examine whether teachers' implementation of CBM progress monitoring in FL results in improved instruction and, ultimately, in improved learning for students who struggle to learn a foreign language.

References

Baker, S. K., & Good, R. (1995). Curriculum-based measurement of English reading with bilingual Hispanic students: A validation study with second grade students. *School Psychology Review, 24,* 561-578. doi:10.1080/02796015.1995.12085788

Blomert, L. (2005). *Dyslexie in Nederland* [Dyslexia in The Netherlands]. Retrieved from https://www.boomtestonderwijs.nl/media/14/boek_dyslexie_in_nederland.pdf

Campus Explorer. (2019). *College language requirements.* Retrieved from https://www.campusexplorer.com/college-advice-tips/16292AF6/College-Language-Requirements/

Campbell, H., Espin, C. A., & McMaster, K. (2013). The technical adequacy of curriculum- based writing measures with English learners. *Reading and Writing, 26,* 431-452. doi:10.1007/s11145-012-9375-6

Chung, S., & Espin, C. A. (2013). CBM progress monitoring in foreign language learning for secondary school students: Technical adequacy of different measures and scoring procedures. *Assessment for Effective Intervention*, *38*, 236-248. doi:10.1177/15345084134897

Cito. (2015). *Cito Volgsysteem Voortgezet Onderwijs. Toets 2* [Cito monitoring system for secondary education. Test 2]. Cito.

Conoyer, S. J., Lembke, E. S., Hosp, J. L., Espin, C. A., Hosp, M. K., & Poch, A. L. (2017). Getting more from your maze: Examining differences in distractors. *Reading & Writing Quarterly*, *33*, 141-154. doi:10.1080/10573569.2016.1142913

De Jong, P. F., De Bree, E. H., Henneman, K., Kleijnen, R., Loykens, E. H. M., Rolak, M., Struiksma, A. J. C, Verhoeven, L., & Wijnen, F. N. K. (2016). *Dyslexie: Diagnostiek en behandeling. Brochure van de Stichting Dyslexie Nederland* [Dyslexia: Diagnostics and treatment. Brochure of the Foundation Dyslexia the Netherlands.] Retrieved from http://www.stichtingdyslexienederland.nl/publicaties/brochures-sdn

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological), 39*, 1-22. doi:10.1111/j.2517-6161.1977.tb01600.x

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional children*, *52*, 219-232. doi:10.1177/001440298505200303

Deno, S. L., & Marston, D. (1987). *Standard reading passages: Measures for screening and progress monitoring.* Children's Educational Services.

DiFino, S. M., & Lombardino, L. J. (2004). Language learning disabilities: The ultimate foreign language challenge. *Foreign Language Annals, 37,* 390-400. doi:10.1111/j.1944-9720.2004.tb02697.x

Domínguez de Ramírez, R., & Shapiro, E. (2006). Cross-language relationship between Spanish and English oral reading fluency among Spanish-speaking English Language Learners in bilingual education classrooms. *Psychology in the Schools, 44,* 795-806. doi:10.1002/pits.20266

Domínguez de Ramírez, R., & Shapiro, E. (2007). Curriculum-based measurement and the evaluation of reading skills of Spanish-speaking English Language Learners in bilingual education classrooms. *School Psychology Review, 35,* 356-369. doi:10.1080/02796015.2006.12087972

Dutch Education Council. (2008). *Vreemde talen in het onderwijs* [Foreign languages in education, Advisory report]. Retrieved from https://onderwijsraad.archief-web.eu/?timestamp=20190923033712&url=https%3A%2F%2Fwww.onderwijsraad.nl%2Fpublicaties%2Farchief%2Fitem21#archive

Dutch Inspectorate of Education. (2018). *Rapport De Staat van het Onderwijs 2018: Onderwijsverslag over 2016/2017* [Report: The State of Education 2018: Education report about 2016/2017, Report]. Retrieved from https://www.onderwijsinspectie.nl/documenten/rapporten/2018/04/11/rapport-de-staat-van-het-onderwijs

Dutch Inspectorate of Education. (2019). *Dyslexieverklaringen. Verschillen tussen scholen nader bekeken* [Dyslexia statements. A closer look at differences between schools, Report]. Retrieved from https://www.onderwijsinspectie.nl/documenten/themarapporten/2019/04/10/dyslexieverklaringen-verschillen-tussen-scholen-nader-bekeken

Espin, C. A., & Deno, S. L. (2016). Oral reading fluency or reading aloud from text: An analysis through a unified view of construct validity. In K. D. Cummings & Y. Petscher (Eds.), *The fluency construct: Curriculum-based measurement concepts and applications* (pp. 365-384). Springer. doi:10.1007/978-1-4939-2803-3_13

Eurostat. (2017, February 23). *Foreign language learning: 60% of lower secondary level pupils studied more than one foreign language in 2015* [Press release]. Retrieved from http://ec.europa.eu/eurostat/web/products-press-releases/-/3-23022017-AP

Fuchs, L. S. (2004). The past, present, and future of curriculum-based measurement research. *School Psychology Review, 33,* 188-192.

Fuchs, L. S., & Fuchs, D. (1992). Identifying a measure for monitoring student reading progress. *School Psychology Review*, *21*, 45-58.

Ganschow, L., Sparks, R. L., & Javorsky, J. (1998). Foreign language learning difficulties: An historical perspective. *Journal of Learning Disabilities, 31,* 248-258. doi:10.1177/002221949803100304

Graham, J. W., Hofer, S. M., & MacKinnon, D. P. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research, 31*(2), 197-218. doi:10.1207/s15327906mbr3102_3

Grove, A. (2020, September 30). *Foreign language requirement for college admission.* Retrieved from https://www.thoughtco.com/foreign-language-requirement-college-admissions-788842

Hale, G. A., Stansfield, C. W., Rock, D. A., Hicks, M. M., Butler, F. A., & Oller, J. W. (1989). The relation of multiple-choice cloze items to the Test of English as a Foreign Language. *Language Testing, 6,* 47-76. doi:10.1177/026553228900600106

Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, *83*, 1198-1202. doi:10.2307/2290157

Little, T. D., & Rhemtulla, M. (2013). Planned missing data designs for developmental researchers. *Child Development Perspectives, 7,* 199-204. doi:10.1111/cdep.12043

Looney, D., & Lusin, N. (2018, February). Enrollments in languages other than English in United States Institutions of higher education, Summer 2016 and Fall 2016: Preliminary report. *Modern Language Association of America Web Publication.* Retrieved from https://www.mla.org/content/download/ 83540/2197676/2016-Enrollments-Short-Report.pdf.

Luke, D. (2004). *Multilevel modeling.* Sage.

McMaster, K., & Espin, C. A. (2007). Technical features of curriculum-based measurement in writing: A literature review. *The Journal of Special Education*, *41*(2), 68-84. doi:10.1177/00224669070410020301

Porter, D. (1976). Modified cloze procedure: A more valid reading comprehension test. *English Language Teaching Journal, 30*(2), 151-155. doi:10.2307/329673

Rippe, R.C.A., & Merkelbach, I. (2019). *Planned missing data in early literacy interventions: A replication study with an additional gold standard.* Manuscript submitted for publication.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1-36. doi:10.18637/jss.v048.i02

Sandberg, K. L., & Reschly, A. L. (2011). English learners: Challenges in assessment and the promise of curriculum-based measurement. *Remedial and Special Education*, *32*, 144-154. doi:10.1177/0741932510361260

Skinner, M. E., & Smith, A. T. (2011). Creating success for students with learning disabilities in postsecondary foreign language courses. *International Journal of Special Education*, *26*, 42-57. Retrieved from http://www.internationaljournalofspecialed.com

Sparks, R. L. (2006). Is there a "disability" for learning a foreign language? *Journal of Learning Disabilities, 39,* 544-557. doi:10.1177/00222194060390060601

Sparks, R. L. (2008). Evidence-based accommodation decision making at the postsecondary level: Review of the evidence for foreign language learning. *Learning Disabilities Research & Practice, 23,* 180-183. doi:10.111/j.1540-5826.2008.00276.x

Sparks, R. L. (2009). If you don't know where you're going, you'll wind up somewhere else: The case of "foreign language learning disability." *Foreign Language Annals, 42,* 7-26. doi:10.111/j.1944-9720.2009.01005.x

Sparks, R. L. (2016). Myths about foreign language learning and learning disabilities. *Foreign Language Annals*, *49*, 252-270. doi:10.1111/flan.12196

Sparks, R. L., Patton, J., Ganschow, L., Humbach, N., & Javorsky, J. (2006). Native language predictors of foreign language proficiency and foreign language aptitude. *Annals of Dyslexia, 56,* 129-160. doi:10.1007/s11881-0006-2

Sparks, R. L., Schneider, E., & Ganschow, L. (2002). Teaching foreign (second) language to at-risk learners. In J. A. Hammadou-Sullivan (Ed.), *Literacy and the second language learner* (pp. 55-84). Retrieved from http://www.infoagepub.com/

Van Til, A., & Van Boxtel, H. (2015). *Wetenschappelijke verantwoording Toets 0 t/m 3, tweede generatie* [Scientific justification Test 0 to 3, second generation]. Cito. Retrieved from http://www.cito.nl/

Wayman, M., Wallace, T., Wiley, H. I., Tichá, R., & Espin, C. A. (2007). Literature synthesis on curriculum-based measurement in reading. *The Journal of Special Education*, *41*, 85-120. doi:10.1177/00224669070410020401

Wallace, C. (2007). Vocabulary: The key to teaching English language learners to read. *Reading Improvement*, *44*, 189-194. Retrieved from http://www.projectinnovation.biz/reading_improvement

Wight, M.C.S. (2015). Students with learning disabilities in the foreign language learning environment and the practice of exemption. *Foreign Language Annals, 48,* 39-55. doi:10.1111/flan.12122

Manuscript Submission Guidelines

English is used for submissions to the journal, correspondence and publication. All submissions must be formatted consistent with the 7th edition of the *Publication Manual of the American Psychological Association* (APA). Manuscripts must include a 100- to 150-word abstract summarizing the contents.

A critical concern in the learning disabilities field is the definition of the population. Therefore, authors are expected to operationally define the study participants in accordance with professional standards (see CLD Research Committee: Rosenberg et al., 1993. Minimum standards for the description of participants in learning disabilities research. *Learning Disability Quarterly, 26*(4), 210-213). In addition, parameters of the setting in which the research took place are to be clearly delineated. Such descriptions facilitate replication and application of results. Manuscripts that fail to specify participant and setting variables will be rejected or returned to the authors for clarification. Authors of research manuscripts are encouraged to include brief (e.g. one to two sentences) explanations of why specific procedures and analysis methods were employed. Peer reviewers will evaluate the appropriateness of this and all aspects of the reported study.

Manuscripts are not to exceed 35 double-spaced pages, consistently employing a 12-point font (including references, tables, figures and appendices). Please limit tables and figures to those essential to conveyance of your content. Please present figures and tables in portrait format and use grey scale rather than colored images.

The manuscript submission and review process will be conducted electronically. To submit your manuscript digitally, you must use one of the following formats: Microsoft Word (.doc or .docx), RTF or PDF. Manuscripts must be saved as "letter" ("U.S. letter") page length/ paper size. Submissions in Word or RTF format will be converted into a PDF before being sent for blind peer review. If you submit a PDF file, it is your responsibility to ensure it can be read and printed by others. To avoid delays, please embed all fonts and use Adobe PDF Distiller instead of PDF Writer to ensure that others can view the article exactly as intended. No email attachment should exceed 15 MB.

Each manuscript must be accompanied by a cover letter that communicates (a) that the manuscript is an original work, (b) that the manuscript is not under consideration by any other journal, and (c) any other disclosures as required by the APA (e.g., ethical treatment of participants, financial relationship disclosures). A cover page that provides the name, affiliation, mailing address, phone number, fax and email address of each author must also accompany each manuscript. All communications will be with the lead author. No author identifying information should be included directly in manuscripts submitted for review (e.g., explicit reference to previous publications, acknowledgments, author biographies). Please send the required cover letter, separate cover page with author identifying information, and manuscript as three separate attachments to a single email to: **ijrld@bc.edu** (please do not send tables, figures, appendices or other supporting materials in separate files). Manuscripts may **not** be submitted by fax or in paper form.

For further information, please refer to:
http://www.iarld.com/wp-content/uploads/2011/08/IJRLD-Call-for-Manuscripts.pdf

International Journal for Research in Learning Disabilities
Lynch School of Education and Human Development
140 Commonwealth Avenue
Chestnut Hill, MA 02467
USA

The IARLD (International Academy for Research in Learning Disabilities) is an international professional organization dedicated to conducting and sharing research about individuals who have learning disabilities.

The IARLD is an elected group of premier scientists, educators and clinicians in the field of learning disabilities throughout the world. The Academy was formed in 1976 by Dr. William Cruickshank (United States of America) and Dr. Jacob Valk (The Netherlands), meeting in Canada with the intention of providing a forum for the exchange of information and the advancement of knowledge regarding learning disabilities.

Since its inception, the Academy has realized its mission of being a professional, international, interdisciplinary consortium of scientists. The Academy currently has a membership of nearly 200 distinguished scholars, representing 26 different countries and thirty disciplines.

IARLD members represent:

- distinguished researchers,
- distinguished practitioner/clinicians,
- young researchers, and
- promising doctoral students.

www.iarld.com

ISSN: 2325-565X (print)
ISSN: 2329-3764 (online)