medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

Characteristics of dynamic assessments of word reading skills and their implications for validity: A systematic review and meta-analysis

Emily Wood[1,2,] Kereisha Biggs[1], & Monika Molnar, [1,2]

[1] Department of Speech-Language Pathology, University of Toronto

[2] Rehabilitation Sciences Institute, University of Toronto

Author Note

Emily Wood https://orcid.org/0000-0002-1466-5615

Monika Molnar https://orcid.org/0000-0003-1337-9948

The authors declare no conflicts of interest.

Correspondence concerning this article should be sent to Emily Wood, Rehabilitation Sciences Institute, 500 University Avenue, Toronto, ON, M5G1V7,

Email: e.wood@utoronto.ca

Phone number: 289-983-1164

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

Abstract

Dynamic assessments (DAs) of word reading skills demonstrate strong criterion reference validity with word reading measures (WRMs). However, DAs vary in the skills they assess, their format and administration method, and the type of words and symbols used in test items. These characteristics may have implications on assessment validity. To compare validity of DAs of word reading skills on these factors of interest, a systematic search of five databases and the grey literature was conducted. We identified 35 studies that met the inclusion criteria of evaluating participants aged 4-10, using a DA of word reading skills and reporting a Pearson's correlation coefficient as an effect size. A random effects meta-analysis with robust variance estimation and subgroup analyses by DA characteristics was conducted. There were no significant differences in mean effect size based on administration method (computer vs. in-person) or symbol type (familiar vs. novel). However, DAs that evaluate phonological awareness or decoding (vs. sound-symbol knowledge), those that use a graduated prompt format (vs. test-teach-retest), and DAs that use nonwords (vs. real words) demonstrated significantly stronger correlations with WRMs. These results inform selection of DAs in clinical and research settings, and development of novel, valid DAs of word reading skills.

Keywords: Dynamic assessment, Validity, Literacy, Reading, Decoding, Phonological awareness, Alphabetic principle, Virtual assessment,

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

Introduction

Literacy is a complex construct requiring integration of multiple skills. Simply, it can be described as the product of the ability to recognize or decode words and to comprehend language (Hoover & Gough, 1990). In this review, we derive our definition of the construct of word reading skills from the subskills that comprise word recognition ability in the evidence-based model Scarborough's reading rope (Scarborough, 2001). These subskills – phonological awareness, knowledge of the alphabetic principle (or sound-symbol knowledge) decoding and sight word recognition have been consistently found to be among the strongest and most accurate predictors of reading ability for young children (e.g., Catts et al., 2005; Ehri, 1998; Hogan et al., 2005). In speech-language pathology (SLP), psychology and education, most word reading tools employ a static assessment (SA) paradigm (e.g., Phonological Awareness Test-2 (PAT-2:NU), Robertson & Salter, 2017; Woodcock Reading Mastery Test-III(WRMT-III), Woodcock, 2011). In SA, an examiner passively evaluates performance without providing prompts or feedback to measure a child's acquired knowledge in a given domain (Grigorenko & Sternberg, 1998). Children from, diverse linguistic backgrounds or those with fewer literacy experiences are prone to perform poorly because they have limited or different acquired knowledge compared to the English monolingual children for whom the tests are designed (Bedore & Peña, 2008; Ginsborg, 2006). When many children underperform, it leads to floor effects, rendering it difficult to discern those who are truly at-risk from those who have had insufficient linguistic or educational experiences. This can result in failure to identify word reading difficulties early (Catts et al., 2009).

Given these limitations, interest in alternative approaches, like dynamic assessments (DAs), has been increasing. While SAs measure a child's acquired skills DAs examine a child's ability to learn a skill with support in the form of teaching, feedback and prompting from the examiner in the test (Grigorenko & Sternberg, 1998). This approach reduces bias and misidentification of difficulty because the impact of previous linguistic or literacy experiences on test outcomes is minimized (Bedore & Peña, 2008; Petersen & Gillam, 2013). Reviews that have evaluated the use of DAs report promising findings

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

on their utility and validity. DAs demonstrate greater predictive validity than SAs across several domains (e.g., DAs of cognitive ability, literacy, and mathematics) (Caffrey et al., 2008). DAs of word reading can predict unique variance in later reading ability beyond SAs (Dixon et al., 2022a), can contribute to the accurate identification of reading difficulties (Dixon et al., 2022b), and demonstrate consistent validity with word reading outcome measures across typically developing, at-risk, bilingual, and monolingual children (Wood et al., 2023).

However, these DAs are also characterized by heterogeneity in terms of the word reading skills they assess, their format, administration method; in addition to the word and symbol type they use. Previous reviews have explored the impact of these factors in the domain of static assessment. For example, the factor of administration method (virtual vs. in-person) was considered by Alfano et al., (2022). Authors found that across the 7 included studies, there were no significant differences between online and in-person administration of pediatric language and literacy assessments. Outcomes such as these provide support for the development of novel SA tools that can be administered online. Whether administration method, or any other of the characteristics described below affect the validity of DAs has not been considered thus far. In this meta-analysis we will directly examine characteristics of DAs and determine which if any types of assessments are superior to others in terms of their criterion reference validity with word reading measures. Outcomes have implications for the revision of existing assessments and development of novel DAs. The characteristics of interest are described in more detail below.

Word Reading Skill Type

DAs have been developed to evaluate various skills associated with literacy development and ability including decoding (e.g., Cho et al., 2017), phonological awareness (e.g., Gellert & Elbro, 2017b), sound-symbol knowledge (e.g., Clayton et al., 2018), morphological awareness (e.g., Navarro et al., 2018), expressive vocabulary development (Peña et al., 2001), oral narratives (Peña et al., 2014), reading comprehension (e.g., Gruhn et al., 2020) and working memory (e.g., Swanson, 1994). As stated,

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

in this review, we focus exclusively on DAs of that evaluate the word reading skills of decoding, phonological awareness or sound-symbol knowledge because these skills have consistently been found to be among the strongest predictors of word reading ability (e.g., The National Literacy Panel (2008)).

Historically research in the validity of static measures of these word reading skills has found that all three correlate strongly with later word reading ability. The National Early Literacy Panel (2008) which included nearly 300 studies, and a meta-analysis of 60 effect sizes by Elbro and Scarborough (2003) both documented strong correlations between kindergarten letter(-sound) knowledge (r=.50 and r=.52 respectively) and decoding (r=.53 and r=.57 respectively) and later word reading ability. Similarly, the National Early Literacy Panel and a meta-analysis of 235 studies by Melby-Lervåg et al., (2012) both found moderate-strong correlations between kindergarten-aged phonemic awareness ability, a subskill of phonological awareness and later reading ability, (r=.42 and r=.43 respectively). These across-study results suggest that SAs of word reading skills are all similarly correlated with later reading abilities, with slight advantage to decoding and letter-sound knowledge over phonological awareness. Whether this is indeed the case in DA is investigated by the current study.

In the realm of DA there have been fewer systematic examinations into the capacity of these skills to predict later word reading. A recent systematic review of 18 studies found that across studies, DAs of phonological awareness and decoding predicted between 1-21% additional unique variance beyond traditional static measures in later word reading ability, but that a DA of paired associate learning (a task akin to learning sound-symbol knowledge) accounted for only 6% unique variance (Dixon et al., 2022a), suggesting a different pattern of prediction between DAs and SAs. Given that DAs evaluate ability to learn rather than acquired knowledge, it may be that assessments that evaluate more complex skills better permit children to demonstrate their ability to learn in the context of DA. Simple sound-symbol knowledge tasks require that a child learn the name that corresponds to a symbol. However, complex phonological awareness tasks like phoneme substitution require that a child identify a sound, delete it, replace it with a new sound and blend the new sounds together to form the new word,

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

and complex decoding tasks require integration of both sound-symbol knowledge and phonological awareness skills. These more complex tasks might be better suited to capturing learning potential in the context of DA. Given this, we expect that DAs of decoding and phonological awareness may demonstrate stronger correlational relationships with WRMs, relative to DAs of SSK.

Format

DAs come in many formats, but there are two primary approaches (Lantolf & Poehner, 2004). Interactionist DA is unscripted and endeavors to modify cognitive or skill ability. The examiner responds contingently to the individual examinee and their capacities. Interventionist DA, however, more closely parallels SA. The examiner provides pre-defined, non-contingent and increasingly explicit levels of support in response to student performance. Its scripted nature requires less clinical skill and time to administer, and its standardization permits evaluation of its psychometric validity (Poehner, 2008). Interventionist DAs across cognitive domains demonstrate stronger predictive validity than interactionist DAs (Caffrey et al., 2008). In the field of word reading assessment, all DAs can be characterized as interventionist. The studies included in this review focus on two common formats of interventionist DA.

The first, referred to as the (test)/teach/retest (TT) format in this paper, consists of a static pre-test, followed by a dynamic teaching phase, and a static re-test (Budoff, 1987). During the teaching phase, children receive feedback and instruction. Not all assessments incorporate the initial static pre-test. If one is conducted, post-test performance is compared to the initial score to assess the difference in performance following teaching. When no pre-test is administered, the post-test measures how a child performs after receiving explicit dynamic instruction in a task.

The second approach, referred to as the graduated prompts (GP) format in this paper, combines teaching and testing phases of the assessment within each item (Brown & Ferrara, 1985). Children are provided with feedback regarding their response. If incorrect, a hierarchy or series of increasingly explicit prompts are provided, until the child answers correctly, or all prompts are exhausted. The greater number of prompts required, the lower the score on an item (Brown & Ferrara, 1985). A previous review

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

suggested that there are no differences in classification accuracy of reading disorder for DAs that use a GP vs. TT format and that both formats are used with similar frequency in assessment of word reading (Dixon et al., 2022b).

Given that previous research has found that noncontingent DAs correlate more strongly with word reading outcomes than contingent ones (Caffrey et al., 2008), we anticipate that there DAs that use a GP format would be superior to those that use the TT format, because this approach is highly explicit and structured, while TT allows greater flexibility in response to the student in the training/teaching portion. However, to date no quantitative comparison of the validity of these two interactionist formats has been conducted.

Administration Method

Assessments, dynamic or otherwise, can be conducted in-person or via computer. Development of virtual or computer-based assessments has become increasingly important, in the wake of the COVID-19 pandemic and the subsequent shift to distance learning (Campbell & Goldstein, 2022). No significant differences have been found between administering an SA online vs. in-person (Alfano et al., 2022), but this factor has not been considered in the context of DAs, which can be administered in-person (e.g., Spector, 1992), virtually by an examiner (e.g., Barker & Saunders, 2020), or in a computer program where no examiner is required (e.g., Aravena et al., 2018). A recent review found that computerised DAs were used less frequently than in-person measures (Dixon et al., 2022b), but the implications of administration method on the validity of DAs have not yet been considered quantitatively. It is anticipated that because DA is characterized by increased interaction between examiner and examinee, it maybe be impacted to a greater extent by computer administration than static assessment. Importantly, post-pandemic, many clinicians and researchers continue to operate virtually, and therefore the factor of administration method and its implications on validity should be considered.

Word Type

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

Assessments of the word reading skills of phonological awareness and decoding use either real words or nonwords, which are made-up words that abide by the language's phonotactic and orthotactic constraints (e.g., "meeb" in English). Commercially developed SA tools such as the Comprehensive Test of Phonological Processing –2 (CTOPP-2, Wagner et al., 2013) which evaluates phonological awareness, or the Woodcock Reading Mastery Test – Third Edition (WRMT-III, Woodcock, 2011), which evaluates decoding, include subtests with words and nonwords. Some DAs like the CUBED-3 dynamic test of decoding include a word reading and nonword decoding measure (Petersen et al., 2016). However, many DAs have fewer subtests and tend to employ either words or nonwords. For example, Gellert and Elbro's dynamic phoneme identification task uses real words (2017b), but their dynamic decoding measure uses nonwords (2017a).

Reading words and non-words are purported to tap into different processes (Shapiro et al., 2013). Children may initially recognize familiar words by sight without activating their knowledge of sound-symbol correspondences, phoneme blending and decoding skills (Ehri & Wilce, 1985). For example, recognizing their name or a high frequency word like "the" in print. However, when reading nonwords, decoding skills are necessary because these words are unfamiliar (Hoover & Tunmer, 1993). Similarly, word and nonword phonological awareness tasks may activate different skills. When using real words, performance may be impacted by acquired vocabulary knowledge, while nonword tasks may be a purer form of phonological ability (Wagner et al., 2013).

In the domain of oral language assessment, nonword repetition tasks have been shown to reduce bias against culturally and linguistically diverse children (Ortiz, 2021). Nonword decoding and phonological awareness tasks may similarly reduce bias against those with different or limited literacy experiences in word reading skills assessments. Children enter kindergarten with a wide range of language and literacy abilities that can be attributed linguistic diversity, their home literacy environment, access to books and libraries, or exposure to literacy instruction in preschool (Ackerman & Barnett, 2005). Importantly, nonword tasks do not disadvantage strong readers with advanced lexical knowledge

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

(Castles et al., 2018). They can account for significant unique variance in word reading ability beyond real word reading (e.g., Hogan et al., 2005). To date, no studies have considered the role of word type in validity of DAs of word reading skills. We expect that nonword tasks will be better suited to predict later reading ability in a DA paradigm, because DAs are designed to measure ability to learn a skill. Because nonword items are unfamiliar, they require learning to complete the word reading skill task.

Symbol Type

Word reading assessments of sound-symbol knowledge (SSK) and decoding use either familiar or novel symbols. Typically, SAs use the letters or characters of the language for which they were created. For instance, in the PAT-2 (Robertson & Salter, 2017) the phoneme-grapheme subtest (a measure of sound-symbol knowledge) evaluates a child's acquired knowledge of the relationship between familiar English letters and sounds, and the phoneme decoding subtest evaluates ability to read nonwords comprised of English graphemes. Recently, there has been increased interest in using novel symbols in DAs as it permits evaluation of how well a child can learn new symbol-sound relationships (e.g., that the symbol ◊= sound /m/, Gellert & Elbro, 2017a), and apply this knowledge to decode symbol-based words (e.g., that the symbols ◊ ◘ = the nonword /ma/, Gellert & Elbro, 2017a), while minimizing the influence of previous linguistic and literacy exposure.

No prior reviews have examined whether symbol type (novel vs. familiar) affects DA's strength of correlational relationship with word reading ability. Primary studies suggest that DAs that use novel symbols can differentiate between typical readers and those with dyslexia (Aravena et al., 2013, 2018). These measures can explain unique variance in later reading ability beyond traditional measures for preliterate children (Horbach et al., 2015). When administered in kindergarten to predict ability in grade 1, a DA decoding measure that used novel symbols (Gellert & Elbro, 2017a) had a superior diagnostic accuracy to a one that used familiar letters (Petersen et al., 2016). Use of novel symbols is a recent development in the field of word reading assessment and there has not yet been a systematic quantitative examination of the relative validity of these two approaches. We hypothesize that in a DA paradigm,

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

evaluating ability to learn SSK or decoding skills with novel symbols may be associated with stronger criterion referenced validity again because it may permit greater capacity to evaluate ability to learn, given the novelty of the symbols used in the tasks.

The Current Study

The current study investigates whether these characteristics of word reading skill, format, administration method, word, and symbol type affect DA's validity as measure by association with performance on word reading measures. We examine criterion validity, as represented by the correlation between performance on a DA of word reading skills (phonological awareness, sound-symbol knowledge or decoding) and a word reading measure (single real or nonword accuracy or fluency). Like Caffrey et al., (2008) we will use Pearson's correlation coefficients as our effect size, given that these are the most observed type of effect size reported across studies. We focus exclusively on DAs of word reading skills as they are best suited to evaluate and predict reading ability in our target demographic of children are learning to read (Catts et al., 2005). We will examine overall validity and stratify DAs into subgroups by their format (graduated prompts vs. train/test), administration method (in-person vs. via computer), word type (real word vs. nonword) and symbol type (familiar vs. novel). We also conduct a comprehensive search of the gray literature and include studies published in French and Spanish. Outcomes of this review will inform which characteristics of DAs of word reading skills are associated with the strongest correlations with word reading measures. For clinicians, outcomes will provide insight into which DA measures are appropriate for use in their practice. For researchers, a quantitative examination how these factors affect validity of DAs of word reading skills can inform revisions of existing measures or development of new tools. This can be achieved by modifying or developing tests with characteristics shown to be most strongly associated with performance on word reading measures.

Research Question

Do the following variables have implications on the validity of dynamic assessments of word reading skills?

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

A) *Word reading skill type:* Phonological awareness (PA), sound-symbol knowledge (SSK) vs. Decoding

B) *Format*: (Test)-teach-retest (TT) vs. graduated prompts (GP)

C) *Administration method*: Computer vs. in-person

D) *Word type*: Real word vs. nonwords

E) *Symbol type*: Novel vs. Familiar

Method

The review objectives and meta-analytic approach were planned a priori and detailed in a registered protocol on the Open Science Framework (Wood & Molnar, 2022).

Eligibility Criteria

Study inclusion criteria were determined a priori and outlined in the review protocol (Wood & Molnar, 2022). Included studies are:

(i) Primary research articles found in peer-reviewed journals, or unpublished grey literature such as Masters or Doctoral theses found in preprint repositories and on Google Scholar.

(ii) Studies that assessed children with a mean age between 4;0 and 10;0 who were monolingual or bi/multilingual, typically developing, at-risk for reading or diagnosed with a reading difficulty. Articles that included adults or children with other developmental challenges, such as hearing impairment, developmental language disorder, or autism spectrum disorder were excluded.

(iv) Articles that reported a correlation coefficient between a DA of a word reading skill, and a word reading measure, concurrently or longitudinally.

(v) No limitation was placed on setting or location, but only articles written in English, French, Spanish, or a different language with full text translation to one of these languages were included.

Search Strategy and Information Sources

An initial search was carried out in 5 databases, MEDLINE, Embase, CINAHL (Cumulative Index to Nursing and Allied Health Literature), PsycINFO and ERIC (Education Resources Information

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

Centre), using the terms "dynamic assessment" and "literacy" as well as their related keywords in titles and abstracts. The search strategy was developed via consultation with a University of Toronto librarian. A complete list of search terms used in each database can be found in Tables 1 and 2 of the supplemental files. No filters were used. Equivalent terms "dynamic assessment" and "literacy" were searched in MedArxiv, EdArxiv and PsyArxiv. Forward searching was then completed on Google Scholar using the "cited by" function with included articles. To check whether any relevant articles were missed during the database, preprint, and Google Scholar search, the reference lists of the included articles were reviewed and compared to the list of included articles. Finally, appeals for unpublished work were made via social media callouts, posts to mailing lists, and direct emails to labs across Canada, the United States and Europe that reported conducting research in field of literacy.

Study Selection and Screening Reliability

Study selection and data extraction were managed in Covidence, a web-based software that facilitates completion of reviews (Covidence, 2023). A team of ten research assistants (RAs) assisted in title/abstract screening and full text review. At the title/abstract stage, RAs received a one-hour training session covering key concepts and relevant terms (e.g., defining dynamic assessment and each word reading skill) and subsequently completed 100 practice title/abstract screenings on a mock review prior to screening in earnest. At this stage, two independent team members voted to include or exclude based on whether the title and abstract indicated that the paper evaluated a word reading skill DA. Across all pairs of reviewers, the weighted mean average agreement was 94% and the Cohen's Kappa coefficient was 0.40 which is characterized as fair (McHugh, 2012). In the full text stage, RAs again received a one-hour training session lead by the first author detailing specific eligibility criteria (e.g., reviewing whether a word reading measure was included, whether the age group was correct, etc.). Each RA completed a practice full text review of a paper with feedback from the first author. Two independent reviewers then voted to include or exclude full texts based on whether they met the pre-defined eligibility criteria. At

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

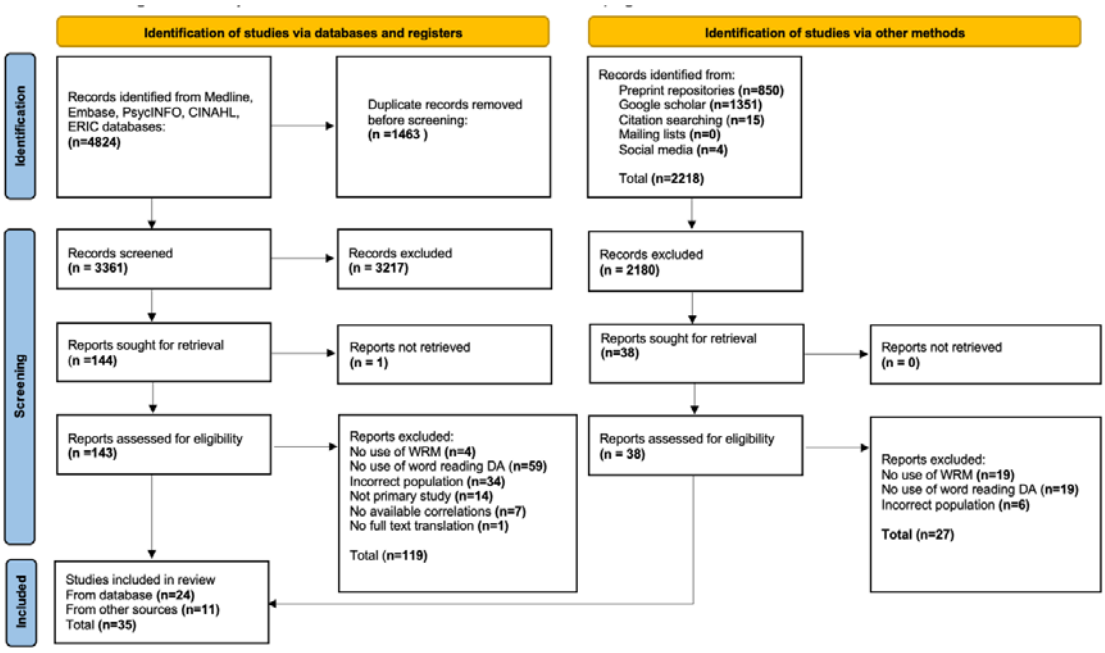
It is made available under a CC-BY-ND 4.0 International license .

this stage, interrater agreement across pairs was 85% and the weighted mean Cohen's Kappa coefficient was 0.66, which is considered substantial (McHugh, 2012).

As demonstrated in Figure 1, 24 articles of the 4824 records identified via the database search were relevant and included. A search of three preprint repositories yielded 850 articles of which one was included. Forward searching of these 25 included articles via Google Scholar led to identification of an additional 9 studies. The reference lists of the 34 articles were reviewed to determine if there were relevant articles that had been missed. One additional study was identified through this process. Finally, callouts were made for unpublished studies or data to mailing lists, via posts to social media and by directly contacting labs conducting literacy related research across Canada, the United State and Europe, but this did not lead to identification of any additional relevant articles. In summary, 35 articles met the criteria for inclusion. The study identification process, including reasons for study exclusion (e.g., no use of a dynamic assessment of one of three word reading skills as in Navarro et al., 2018) is outlined in the PRISMA diagram below (Page et al., 2021).

Figure 1

Preferred Reporting Items for Systematic Review and Meta-Analyses Flowchart



Note. DA = N=number, Dynamic assessment, WRM=Word reading measure

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

Coding Data Items

Data from relevant articles was extracted using a custom template on Covidence, which is available on the Open Science Framework protocol site (Wood & Molnar, 2022). The first and second author both extracted data from all articles and compared and consolidated findings. Any disagreements were resolved through discussion between all three authors. The following data points were extracted for each included study:

General Information

The study title, journal name, date of publication, DOI, author name(s), institutional affiliation(s), funding, any potential conflicts of interest, and the country in which the study took place.

Participant Characteristics

The number of participants included in analyses, the percentage of males, as well as the mean age of the children at the outset of the study was noted. A total of 6683 participants were included. The overall mean age was 5 years 6 months, and the overall average percentage of males was 51%. Mean age was not reported for 14 studies and the percentage of males was not reported for 9. We were able to ensure that studies that did not report mean age still met inclusion criteria, as all minimally reported the grade of participants (e.g., indicated that participants are in grade 1 and therefore between age 4-10). Authors also extracted information regarding participant reading status (typically developing vs. at-risk), language status (monolingual vs. bilingual; and age (4-5, 6-7 vs. 8-9). These factors are examined in a separate paper evaluating the validity of DAs of word reading skills across diverse populations (Wood et al., 2023). Table 1 provides additional details regarding the mean age and % of males of included studies.

Effect Sizes

Pearson's correlation coefficients representing the relationship between DAs and WRMs were extracted. All relevant correlation coefficients listed were noted (e.g., a DA that used two PA tasks and their correlations with WRMs). In some instances, a lower score on a DA translated to better

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

performance. In these situations, negative correlations were transformed to positive ones for analysis. From the 35 studies (*k*), a total of 192 effect sizes (*m*) were extracted, with a mean of 5.6 effect sizes per study and a range of 1-30. To account for dependencies between samples, we noted whether two studies conducted analyses on the same set of participants, and if so, nested these effect sizes together under the same study group. For example, Petersen et al., (2018), conducted a follow-up study using the same participant set as in Petersen et al., (2016), and so these effect sizes were grouped together under the umbrella of one study, using robust variance estimation (see statistical analyses below for further details). Study design (e.g., concurrent or longitudinal) was noted. Effect sizes representing the relationship between DAs and WRMs are presented in Table 1.

Measures

Dynamic Assessments (DAs)

In this review, DA is defined as an assessment that provides teaching, training, feedback on performance, and/or prompting in testing. In some instances, these measures were not reported as "DAs" but were described as paired associate learning tasks. This was typically for measures that evaluated SSK skills (e.g., Liu & Chung, 2021). We reported the word reading skills evaluated (phonological awareness, sound-symbol knowledge, and/or decoding), and the task used to assess the skill (e.g., phonological awareness/phoneme blending). If multiple tasks were used to evaluate a skill, authors listed all tasks utilized. A DA was characterized as SSK if the task involved learning the relationship between a visual referent (symbol or letter) and a syllable or phoneme. DAs were considered to evaluate PA skills if they assessed one or more of the auditory skills of rhyming, blending, segmenting, manipulating, deleting, substituting phonemes, syllables, words or onset and rimes. Tasks that required a child to recognize more than one symbol-sound relationship and blend these sounds together to "read" multi-symbol words (e.g., CV, VC, CVC etc.) were labelled as decoding tasks. Authors also noted the format of the DA (i.e., graduated prompts (GP) or train/test (TT). DAs were considered GP if a series of prompts was used after a participant response to an individual test item and were characterized as TT if they minimally

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

incorporated a teaching/training phase, followed by a separate static post-test. In terms of administration method (i.e., in person or computer) DAs that were conducted virtually by a clinician over the computer, or those that were computerized (i.e., no clinician) were considered as computer-based administration, while all others were characterized as in-person. For word type, DAs that used words that existed in the language of testing's lexicon were considered real words (e.g., cat in English) while those that used invented words (e.g., meeb in English), or words from other languages were considered nonwords (e.g., "copa" a Spanish word in an English task). Finally, authors indicated whether novel or familiar symbols were used. DAs that used letters or characters that belonged to the orthography of the language of testing were said to use familiar symbols (e.g., using alphabet letters in an English DA), while those that used invented symbols or letters from orthographies distinct from the language of testing were characterized as using novel symbols (e.g., using Hebrew letters in an English DA measure).

Of the 192 effect sizes from the 35 included studies that examined use of a DA, most evaluated decoding (k= 20, m=123), some assessed sound-symbol knowledge (k=12, m=45) and fewest examined phonological awareness (k=12, m=24). In some studies, the DA evaluated more than one word reading skill, such as in the case of Cho and Compton (2015) where one subtest evaluated SSK and another decoding. In terms of format, 20 studies used a graduated prompts approach (m=98), while 15 used a (test)-teach-retest approach (m=94). Most DAs were conducted in person, (k=24, m=156) and fewer via computer (k=11, m=36). DAs of phonological awareness and decoding tasks used either real or nonwords in their items. Of the studies that evaluated these skills, most used nonwords (k=*17*, m=96) while fewer used real words (k=15, m=51). The studies that evaluated sound-symbol knowledge did not use words given the nature of the task (k=12, m=45). Decoding and sound-symbol knowledge tasks employed either novel or familiar symbols. Of the studies that evaluated these word reading skills, most used novel symbols (k=15, m=93) while fewer used familiar symbols (k=10, m=75). Phonological awareness tasks are auditory in nature and so studies that evaluated this skill did not use symbols (k=12, m=24). For details about DA characteristics, see Table 1.

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

Word Reading Measures (WRMs)

For the purposes of this review, WRMs are assessments that evaluate word reading ability of single real or nonwords using a correct/incorrect grading system and without provision of feedback, prompting or teaching. WRMs were conducted concurrently with the DA or longitudinally at a later timepoint. Of the 35 studies, most effect sizes represented longitudinal relationships (*k*=20, *m*=116) while fewer are associated with concurrent correlations (*k*=, 21, *m*=76). Some studies provided both concurrent and longitudinal correlation coefficients between measures. In extraction, authors noted the name of the WRM, (e.g., the Woodcock Reading Mastery Tests –III), what type of word reading ability was evaluated and whether this was achieved using words or nonwords (e.g., real word accuracy vs. nonword rate). The most used type of WRM was norm-referenced tests (*k*=22, *m*=127), followed by researcher developed tools (*k*=9, *m*=35), criterion-referenced measures (*k*=3, *m*=29) and finally inventories (*k*=1, *m*=1). The Test of Word Reading Efficiency (TOWRE), a norm-referenced test, was the most commonly used tool (*k*=9, *m*=49), along with the Dynamic Indicators of Basic Early Literacy Skills (DIBELS), (*k*=3, *m*=29) and versions of the Woodcock Reading Mastery Tests (WRMT-R/NU), (*k*=10, *m*=28). Other measures like the Woodcock-Johnson-III (WJ-III), were used less frequently (*k*=2, *m*=6). There was variety across task type, with some effect sizes representing correlations between DAs and nonword accuracy measures (*m*=25), real word accuracy measures (*m*=52), nonword rate tasks (*m*=40), real word rate tasks (*m*=49), or a combination or composite of several word reading abilities (*m*=26). Characteristics of WRMs used are detailed in Table 1.

Quality Appraisal Ratings

Following extraction, included studies were evaluated independently by the first and second author using an adapted version of two quality assessment tools from the Johanna Briggs Institute (Moola et al., 2020). This tool is available on the Open Science Framework protocol page (Wood & Molnar, 2022). Most studies (*k*=26) were peer-reviewed and those that were not (*k*=9) were Masters or Doctoral theses. Studies were assessed on: (i) participant selection, (ii) index assessments (DAs) (iii)

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

reference assessments (WRMs), (iv) flow and timing of the study, (v) statistical analyses. First, coders rated whether the age, sex, and demographic characteristics of the participants were adequately described. The DA domain rating was informed by whether the tool was explained with adequate detail regarding the skills assessed, and the characteristics of the measure. Coders also noted whether the word reading skill(s) employed were developmentally appropriate for the sample population. The reference assessments (WRMs) were also evaluated on their appropriateness of use for the population, and their psychometric properties. To assess flow and timing, coders evaluated whether the analyses included all participants and if not, whether the author(s) provided reasoning for attrition. Lastly, coders considered whether appropriate statistical analyses were conducted.

Overall, the quality appraisal consisted of 8 items to be rated over 5 domains. Items regarding participants, flow and timing, and statistical analyses were assigned one point, while items concerned with the index test (DA) and the reference tests (WRMs) were worth two points due to their greater significance in achieving review objectives. Conflicts were resolved through discussion between all three authors. Quality scores were ranked as either low quality (0-33%), medium quality (34-66%) or high quality (67-100%). Only medium and high-quality studies were included in the analyses. No studies were excluded based on their score. The overall quality appraisal rating for each study is included in Table 1 below. Please refer to Table 3 in the supplemental material for individual ratings for each question for each study.

Table 1.

Number of Participants, Effect Size, Mean Age, Grade, % Males, Study Design, Skills Evaluated and Characteristics of DAs and Word Reading Measures of Included Studies.

Study	*N*	*r*	Age	% Male	Study Design	DA Admin Method	DA Format	DA Skill	DA Word Type	DA Symbol Type	WRM Skill	WRM Type	WRM Name	PR	QAR /12	
Aravena et al., 2013	42	0.52	9;9	56.4	C	Computer	TT	Dec	words	novel	Real Word Rate	NR	One Minute Test	Y	11	
Aravena et al., 2018	71	0.015	9;2	58.2	C	Computer	TT	SSK	NOT	novel	Real Word Accuracy	NR	3DM	Y	11	
Aravena et al., 2018	71	0.147	9;2	58.2	C	Computer	TT	SSK	NOT	novel	Real Word Rate	NR	3DM	Y		
Aravena et al., 2018	71	0.075	9;2	58.2	C	Computer	TT	SSK	NOT	novel	Real Word Accuracy	NR	3DM	Y		
Aravena et al., 2018	71	0.214	9;2	58.2	C	Computer	TT	SSK	NOT	novel	Real Word Rate	NR	3DM	Y		
Aravena et al., 2018	71	0.237	9;2	58.2	C	Computer	TT	Dec	words	novel	Real Word Accuracy	NR	3DM	Y		
Aravena et al., 2018	71	0.392	9;2	58.2	C	Computer	TT	Dec	words	novel	Real Word Rate	NR	3DM	Y		
Barker & Saunders, 2020	27	0.47	4;11	51.8	C	Computer	TT	SSK	NOT	novel	Nonword Accuracy	NR	WRMT-III	Y	10	
Barker & Saunders, 2020	27	0.63	4;11	51.8	C	Computer	GP	SSK	NOT	novel	Real Word Accuracy	NR	WRMT-III	Y		
Caffrey, 2006a	25	0.624	NA	52	L	Person	GP	Dec	nonwords	familiar	Real Word Accuracy	NR	WRAT	N	10	
Caffrey, 2006a	25	0.706	NA	52	L	Person	GP	Dec	nonwords	familiar	Nonword Accuracy	NR	WRAT	N		

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license.

Caffrey, 2006a	25	0.585	NA	52	L	Person	GP	Dec	nonwords	familiar	Nonword Rate	NR	WRAT	N		
Caffrey, 2006b	95	0.745	NA	57	L	Person	GP	Dec	nonwords	familiar	Real Word Accuracy	NR	WRAT	N	10	
Caffrey, 2006b	95	0.765	NA	57	L	Person	GP	Dec	nonwords	familiar	Nonword Accuracy	NR	WRAT	N		
Caffrey, 2006b	95	0.613	NA	57	L	Person	GP	Dec	nonwords	familiar	Nonword Rate	NR	WRAT	N		
Cho & Compton, 2015	112	0.24	6;8	57.1	C	Person	GP	SSK	NOT	novel	Nonword Accuracy	NR	WRMT-R/NU	Y	12	
Cho & Compton, 2015	112	0.24	6;8	57.1	C	Person	GP	SSK	NOT	novel	Nonword Rate	NR	TOWRE	Y		
Cho & Compton, 2015	112	0.23	6;8	57.1	C	Person	GP	SSK	NOT	novel	Real Word Accuracy	NR	WRMT-R/NU	Y		
Cho & Compton, 2015	112	0.14	6;8	57.1	C	Person	GP	SSK	NOT	novel	Real Word Rate	NR	TOWRE	Y		
Cho & Compton, 2015	112	0.5	6;8	57.1	C	Person	GP	Dec	words	novel	Nonword Accuracy	NR	WRMT-R/NU	Y		
Cho & Compton, 2015	112	0.4	6;8	57.1	C	Person	GP	Dec	words	novel	Nonword Rate	NR	TOWRE	Y		
Cho & Compton, 2015	112	0.43	6;8	57.1	C	Person	GP	Dec	words	novel	Real Word Accuracy	NR	WRMT-R/NU	Y		
Cho & Compton, 2015	112	0.32	6;8	57.1	C	Person	GP	Dec	words	novel	Real Word Rate	NR	TOWRE	Y		
Cho & Compton, 2015	112	0.64	6;8	57.1	C	Person	GP	Dec	words	novel	Nonword Accuracy	NR	WRMT-R/NU	Y		
Cho & Compton, 2015	112	0.51	6;8	57.1	C	Person	GP	Dec	words	novel	Nonword Rate	NR	TOWRE	Y		
Cho & Compton, 2015	112	0.54	6;8	57.1	C	Person	GP	Dec	words	novel	Real Word Accuracy	NR	WRMT-R/NU	Y		

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

Cho & Compton, 2015	112	0.41	6;8	57.1	C	Person	GP	Dec	words	novel	Real Word Rate	NR	TOWRE	Y		
Cho et al., 2014	134	0.69	NA	50	C	Person	GP	Dec	nonwords	familiar	Nonword Accuracy	NR	WRMT-R/NU	Y	10	
Cho et al., 2014	134	0.58	NA	50	C	Person	GP	Dec	nonwords	familiar	Nonword Rate	NR	TOWRE	Y		
Cho et al., 2017	105	0.21	6;8	55	C	Person	GP	SSK	NOT	novel	Real Word Accuracy	NR	WRMT-R/NU	Y	11	
Cho et al., 2017	105	0.1	6;8	55	C	Person	GP	SSK	NOT	novel	Real Word Rate	NR	TOWRE	Y		
Cho et al., 2017	105	0.22	6;8	55	C	Person	GP	SSK	NOT	novel	Nonword Accuracy	NR	WRMT-R/NU	Y		
Cho et al., 2017	105	0.24	6;8	55	C	Person	GP	SSK	NOT	novel	Nonword Rate	NR	TOWRE	Y		
Cho et al., 2017	105	0.24	6;8	55	C	Person	GP	Dec	words	novel	Real Word Accuracy	NR	WRMT-R/NU	Y		
Cho et al., 2017	105	0.24	6;8	55	C	Person	GP	Dec	words	novel	Real Word Rate	NR	TOWRE	Y		
Cho et al., 2017	105	0.43	6;8	55	C	Person	GP	Dec	words	novel	Nonword Accuracy	NR	WRMT-R/NU	Y		
Cho et al., 2017	105	0.31	6;8	55	C	Person	GP	Dec	words	novel	Nonword Rate	NR	TOWRE	Y		
Cho et al., 2017	105	0.43	6;8	55	C	Person	GP	Dec	words	novel	Real Word Accuracy	NR	WRMT-R/NU	Y		
Cho et al., 2017	105	0.27	6;8	55	C	Person	GP	Dec	words	novel	Real Word Rate	NR	TOWRE	Y		
Cho et al., 2017	105	0.55	6;8	55	C	Person	GP	Dec	words	novel	Nonword Accuracy	NR	WRMT-R/NU	Y		
Cho et al., 2017	105	0.38	6;8	55	C	Person	GP	Dec	words	novel	Nonword Rate	NR	TOWRE	Y		
Cho et al., 2017	105	0.22	6;8	55	L	Person	GP	SSK	NOT	novel	Real Word Accuracy	NR	WRMT-R/NU	Y		
Cho et al., 2017	105	0.11	6;8	55	L	Person	GP	SSK	NOT	novel	Real Word Rate	NR	TOWRE	Y		
Cho et al., 2017	105	0.22	6;8	55	L	Person	GP	SSK	NOT	novel	Nonword Accuracy	NR	WRMT-	Y		

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

													R/NU			
Cho et al., 2017	105	0.19	6;8	55	L	Person	GP	SSK	NOT	novel	Nonword Rate	NR	TOWRE	Y		
Cho et al., 2017	105	0.45	6;8	55	L	Person	GP	Dec	words	novel	Real Word Accuracy	NR	WRMT-R/NU	Y		
Cho et al., 2017	105	0.34	6;8	55	L	Person	GP	Dec	words	novel	Real Word Rate	NR	TOWRE	Y		
Cho et al., 2017	105	0.42	6;8	55	L	Person	GP	Dec	words	novel	Nonword Accuracy	NR	WRMT-R/NU	Y		
Cho et al., 2017	105	0.39	6;8	55	L	Person	GP	Dec	words	novel	Nonword Rate	NR	TOWRE	Y		
Cho et al., 2017	105	0.46	6;8	55	L	Person	GP	Dec	words	novel	Real Word Accuracy	NR	WRMT-R/NU	Y		
Cho et al., 2017	105	0.32	6;8	55	L	Person	GP	Dec	words	novel	Real Word Rate	NR	TOWRE	Y		
Cho et al., 2017	105	0.47	6;8	55	L	Person	GP	Dec	words	novel	Nonword Accuracy	NR	WRMT-R/NU	Y		
Cho et al., 2017	105	0.43	6;8	55	L	Person	GP	Dec	words	novel	Nonword Rate	NR	TOWRE	Y		
Chow, 2014	121	0.22	7;0	54.5	C	Person	TT	SSK	NOT	novel	Real Word Accuracy	NR	HKT-SpLD	Y	12	
Chow, 2014	121	0.28	7;0	54.5	C	Person	TT	SSK	NOT	novel	Real Word Accuracy	RD	NA	Y		
Clayton et al., 2018	97	0.12	9;2	50.5	C	Computer	TT	SSK	NOT	novel	Real Word Accuracy	NR	SWRT6-16	Y	10	
Clayton et al., 2018	97	0.1	9;2	50.5	C	Computer	TT	SSK	NOT	novel	Real Word Rate	NR	TOWRE	Y		
Clayton et al., 2018	97	0.05	9;2	50.5	C	Computer	TT	SSK	NOT	novel	Nonword Rate	NR	TOWRE	Y		
Clayton et al., 2018	97	0.29	9;2	50.5	C	Computer	TT	SSK	NOT	novel	Real Word Accuracy	NR	SWRT6-16	Y		
Clayton et al., 2018	97	0.26	9;2	50.5	C	Computer	TT	SSK	NOT	novel	Real Word Rate	NR	TOWRE	Y		
Clayton et al., 2018	97	0.22	9;2	50.5	C	Computer	TT	SSK	NOT	novel	Nonword Rate	NR	TOWRE	Y		
Compton et al.,	355	0.69	NA	53.45	L	Person	GP	Dec	nonwords	familiar	Composite-Real Word	NR	TOWRE,	Y	10	

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

2010											Accuracy, Real Word Rate, Nonword Accuracy, Nonword Rate, Reading Comprehension		WMRT-R/NU			
Coventry et al., 2011	1988	0.423	4'11	50.5	C	Person	GP	PA	words	NOT	Real Word Accuracy, Real Word Rate	NR	TOWRE	Y	10	
Cunningham & Carroll, 2011	45	0.67	NA	NA	C	Person	GP	PA	nonwords	NOT	Real Word Accuracy	NR	BAS-2	Y	11	
Cunningham & Carroll, 2011	45	0.77	NA	NA	L	Person	GP	PA	nonwords	NOT	Real Word Accuracy	NR	BAS-2	Y		
Edwards, 2020	312	0.47	NA	NA	C	Person	GP	Dec	nonwords	familiar	Real Word Rate	NR	TOWRE	N	9	
Edwards, 2020	312	0.46	NA	NA	C	Person	GP	Dec	nonwords	familiar	Nonword Rate	NR	TOWRE	N		
Edwards, 2020	312	0.53	NA	NA	C	Person	GP	Dec	nonwords	familiar	Real Word Accuracy	NR	WJ-III	N		
Edwards, 2020	312	0.5	NA	NA	C	Person	GP	Dec	nonwords	familiar	Nonword Accuracy	NR	WJ-III	N		
Edwards, 2020	312	0.55	NA	NA	C	Person	GP	Dec	nonwords	familiar	Real Word Rate	NR	TOWRE	N		
Edwards, 2020	312	0.6	NA	NA	C	Person	GP	Dec	nonwords	familiar	Nonword Rate	NR	TOWRE	N		
Edwards, 2020	312	0.61	NA	NA	C	Person	GP	Dec	nonwords	familiar	Real Word Accuracy	NR	WJ-III	N		
Edwards, 2020	312	0.64	NA	NA	C	Person	GP	Dec	nonwords	familiar	Nonword Accuracy	NR	WJ-III	N		
Fuchs et al., 2011	318	0.72	NA	49.4	L	Person	GP	Dec	nonwords	familiar	Real Word Accuracy	NR	WRMT-R/NU	Y	10	
Fuchs et al., 2011	318	0.61	NA	49.4	L	Person	GP	Dec	nonwords	familiar	Real Word Rate	NR	TOWRE	Y		
Gan et al., 2022	135	0.68	NA	54	C	Person	GP	Dec	words	familiar	Real Word Accuracy	RD	Character Recognition Task	Y	11	

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

Gan et al., 2022	135	0.7	NA	54	L	Person	GP	Dec	words	familiar	Real Word Accuracy	RD	Character Recognition Task	Y		
Gan et al., 2022	135	0.66	NA	54	L	Person	GP	Dec	words	familiar	Real Word Accuracy	RD	Character Recognition Task	Y		
Gan et al., 2022	135	0.64	NA	54	L	Person	GP	Dec	words	familiar	Real Word Accuracy	RD	Character Recognition Task	Y		
Gellert & Elbro, 2017a	171	0.09	6;11	NA	C	Person	TT	SSK	NOT	novel	Real Word Accuracy	RD	NA	Y	10	
Gellert & Elbro, 2017a	171	0.65	6;11	NA	C	Person	TT	Dec	nonwords	novel	Real Word Accuracy	RD	NA	Y		
Gellert & Elbro, 2017a	171	0.21	6;11	NA	L	Person	TT	SSK	NOT	novel	Real Word Accuracy, Nonword Accuracy	RD	NA	Y		
Gellert & Elbro, 2017a	171	0.66	6;11	NA	L	Person	TT	Dec	nonwords	novel	Real Word Accuracy, Nonword Accuracy	RD	NA	Y	11	
Gellert & Elbro, 2017b	160	0.46	6;4	48	L	Person	GP	PA	words	NOT	Real Word Accuracy	RD	NA	Y		
Gellert & Elbro, 2017b	160	0.47	6;4	48	L	Person	GP	PA	words	NOT	Real Word Accuracy, Nonword Accuracy	RD	NA	Y		
Gellert & Elbro, 2017b	160	0.36	6;4	48	L	Person	GP	PA	words	NOT	Real Word Accuracy, Nonword Accuracy	RD	NA	Y		
Gellert & Elbro, 2017b	160	0.65	6;4	48	L	Person	GP	Dec	nonwords	novel	Real Word Accuracy, Nonword Accuracy	RD	NA	Y		
Gellert & Elbro, 2017b	160	0.76	6;4	48	L	Person	GP	Dec	nonwords	novel	Real Word Accuracy, Nonword Accuracy	RD	NA	Y		

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

Gellert & Elbro, 2017b	160	0.63	6;4	48	L	Person	GP	Dec	nonwords	novel	Real Word Accuracy, Nonword Accuracy	RD	NA	Y		
Gellert & Elbro, 2018	158	0.2	6;11	NA	C	Person	GP	SSK	NOT	novel	Real Word Accuracy	RD	NA	Y	10	
Gellert & Elbro, 2018	158	0.47	6;11	NA	C	Person	GP	SSK	NOT	novel	Real Word Accuracy	RD	NA	Y		
Gellert & Elbro, 2018	158	0.76	6;11	NA	C	Person	GP	Dec	nonwords	novel	Real Word Accuracy	RD	NA	Y		
Gellert & Elbro, 2018	158	0.84	6;11	NA	C	Person	GP	Dec	nonwords	novel	Real Word Accuracy, Nonword Accuracy	RD	NA	Y		
Gellert & Elbro, 2018	158	0.25	6;11	NA	L	Person	GP	SSK	NOT	novel	Real Word Accuracy, Nonword Accuracy	RD	NA	Y		
Gellert & Elbro, 2018	158	0.25	6;11	NA	L	Person	GP	SSK	NOT	novel	Real Word Accuracy, Nonword Accuracy	RD	NA	Y		
Gellert & Elbro, 2018	158	0.19	6;11	NA	L	Person	GP	SSK	NOT	novel	Real Word Rate, Nonword Rate	RD	NA	Y		
Gellert & Elbro, 2018	158	0.42	6;11	NA	L	Person	GP	SSK	NOT	novel	Real Word Accuracy, Nonword Accuracy	RD	NA	Y		
Gellert & Elbro, 2018	158	0.35	6;11	NA	L	Person	GP	SSK	NOT	novel	Real Word Accuracy, Nonword Accuracy	RD	NA	Y		
Gellert & Elbro, 2018	158	0.8	6;11	NA	L	Person	GP	Dec	nonwords	novel	Real Word Accuracy, Nonword Accuracy	RD	NA	Y		
Gellert & Elbro, 2018	158	0.65	6;11	NA	L	Person	GP	Dec	nonwords	novel	Real Word Accuracy, Nonword Accuracy	RD	NA	Y		
Gellert & Elbro, 2018	158	0.56	6;11	NA	L	Person	GP	Dec	nonwords	novel	Real Word Rate, Nonword Rate	RD	NA	Y		
Gellert & Elbro,	158	0.69	6;11	NA	L	Person	GP	Dec	nonwords	novel	Real Word Accuracy,	RD	NA	Y		

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

2018											Nonword Accuracy					
Gellert & Elbro, 2018	158	0.61	6;11	NA	L	Person	GP	Dec	nonwords	novel	Real Word Rate, Nonword Rate	RD	NA	Y		
Gillam et al., 2011	64	0.54	NA	47	C	Computer	GP	PA	words	NOT	Nonword Accuracy	NR	WRMT-R/NU	Y	11	
Horbach et al., 2015	75	0.283	6;2	56	C	Computer	TT	SSK	NOT	novel	Nonword Accuracy	RD	NA	Y	10	
Horbach et al., 2015	75	0.384	6;2	56	C	Computer	TT	Dec	nonwords	novel	Nonword Accuracy	RD	NA	Y		
Horbach et al., 2015	75	0.212	6;2	56	C	Computer	TT	Dec	nonwords	novel	Nonword Accuracy	RD	NA	Y		
Horbach et al., 2015	75	0.215	6;2	56	L	Computer	TT	SSK	NOT	novel	Composite-Real Word Accuracy, Real Word Rate, Nonword Accuracy, Nonword Rate, Reading Comprehension	NR	SLRT-II	Y		
Horbach et al., 2015	75	0.372	6;2	56	L	Computer	TT	Dec	nonwords	novel	Composite-Real Word Accuracy, Real Word Rate, Nonword Accuracy, Nonword Rate, Reading Comprehension	NR	SLRT-II	Y		
Horbach et al., 2015	75	0.258	6;2	56	L	Computer	TT	Dec	nonwords	novel	Composite-Real Word Accuracy, Real Word Rate, Nonword Accuracy, Nonword	NR	SLRT-II	Y		

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

											Rate, Reading Comprehension					
Horbach et al., 2015	193	0.169	6;2	56	L	Computer	TT	SSK	NOT	novel	Composite-Real Word Accuracy, Real Word Rate, Nonword Accuracy, Nonword Rate, Reading Comprehension	NR	SLRT-II	Y		
Horbach et al., 2015	193	0.279	6;2	56	L	Computer	TT	Dec	nonwords	novel	Composite-Real Word Accuracy, Real Word Rate, Nonword Accuracy, Nonword Rate, Reading Comprehension	NR	SLRT-II	Y		
Horbach et al., 2015	193	0.362	6;2	56	L	Computer	TT	Dec	nonwords	novel	Composite-Real Word Accuracy, Real Word Rate, Nonword Accuracy, Nonword Rate, Reading Comprehension	NR	SLRT-II	Y		
Horbach et al., 2018	17	0.855	5;0	41%	L	Computer	TT	Dec	nonwords	novel	Real Word Rate	NR	SLRT-II	Y	10	
Horbach et al., 2018	17	0.807	5;0	41%	L	Computer	TT	SSK	nonwords	novel	Nonword Rate	NR	SLRT-II	Y		
Law et al., 2018	84	0.086	8;3	NA	C	Computer	TT	SSK	NOT	novel	Real Word Accuracy	NR	EMT	Y	11	
Law et al., 2018	84	0.044	8;3	NA	C	Computer	TT	SSK	NOT	novel	Nonword Accuracy	NR	Klepel	Y		

Law et al., 2018	84	0.027	8;3	NA	C	Computer	TT	Dec	words	novel	Real Word Accuracy	NR	EMT	Y		
Law et al., 2018	84	0.022	8;3	NA	C	Computer	TT	SSK	NOT	novel	Nonword Accuracy	NR	Klepel	Y		
Law et al., 2018	84	0.315	8;3	NA	C	Computer	TT	SSK	NOT	novel	Real Word Accuracy	NR	EMT	Y		
Law et al., 2018	84	0.279	8;3	NA	C	Computer	TT	Dec	words	novel	Nonword Accuracy	NR	Klepel	Y		
Liu & Chung, 2021	203	0.19	5;0	51.23	C	Person	TT	SSK	NOT	novel	Real Word Accuracy	RD	NA	Y	9	
Liu & Chung, 2021	203	0.18	5;0	51.23	C	Person	TT	SSK	NOT	novel	Real Word Accuracy	RD	NA	Y		
Loreti, 2013	10	0.75	4;10	60%	C	Computer	TT	PA	nonwords	NOT	Real Word Accuracy	NR	WMLSR	N	9	
Osa Fuentes, 2003	164	0.44	5;6	45.7	L	Person	GP	PA	words	NOT	Real Word Accuracy	NR	K-ABC	N	9	
Osa Fuentes, 2003	164	0.28	5;6	45.7	L	Person	GP	PA	words	NOT	Real Word Rate	NR	TALE	N		
Osa Fuentes, 2003	164	0.44	5;6	45.7	L	Person	GP	PA	words	NOT	Real Word Accuracy	NR	K-ABC	N		
Osa Fuentes, 2003	164	0.33	5;6	45.7	L	Person	GP	PA	words	NOT	Real Word Rate	NR	TALE	N		
Osa Fuentes, 2003	164	0.64	5;6	45.7	L	Person	GP	PA	words	NOT	Real Word Accuracy	NR	K-ABC	N		
Osa Fuentes, 2003	164	0.36	5;6	45.7	L	Person	GP	PA	words	NOT	Real Word Rate	NR	TALE	N		
Petersen & Gillam, 2015	63	0.24	5;4	53.9	L	Person	TT	Dec	nonwords	familiar	Nonword Rate	CR	DIBELS	Y	12	
Petersen & Gillam, 2015	63	0.28	5;4	53.9	L	Person	TT	Dec	nonwords	familiar	Real Word Accuracy	NR	WRMT-R	Y		
Petersen & Gillam, 2015	63	0.33	5;4	53.9	L	Person	TT	Dec	nonwords	familiar	Nonword Rate	CR	DIBELS	Y		

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

Petersen & Gillam, 2015	63	0.39	5;4	53.9	L	Person	TT	Dec	nonwords	familiar	Real Word Accuracy	NR	WRMT-R	Y		
Petersen & Gillam, 2015	63	0.46	5;4	53.9	L	Person	TT	Dec	nonwords	familiar	Nonword Rate	CR	DIBELS	Y		
Petersen & Gillam, 2015	63	0.51	5;4	53.9	L	Person	TT	Dec	nonwords	familiar	Real Word Accuracy	NR	WRMT-R	Y	11	
Petersen et al., 2016	320	0.32	NA	NA	L	Person	TT	Dec	nonwords	familiar	Real Word Rate	NR	TOWRE	Y		
Petersen et al., 2016	320	0.32	NA	NA	L	Person	TT	Dec	nonwords	familiar	Nonword Rate	NR	TOWRE	Y		
Petersen et al., 2016	320	0.26	NA	NA	L	Person	TT	Dec	nonwords	familiar	Nonword Rate	CR	DIBELS	Y		
Petersen et al., 2016	320	0.21	NA	NA	L	Person	TT	Dec	nonwords	familiar	Real Word Rate	NR	TOWRE	Y		
Petersen et al., 2016	320	0.2	NA	NA	L	Person	TT	Dec	nonwords	familiar	Nonword Rate	NR	TOWRE	Y		
Petersen et al., 2016	320	0.26	NA	NA	L	Person	TT	Dec	nonwords	familiar	Nonword Rate	CR	DIBELS	Y		
Petersen et al., 2016	320	0.39	NA	NA	L	Person	TT	Dec	nonwords	familiar	Real Word Rate	NR	TOWRE	Y		
Petersen et al., 2016	320	0.37	NA	NA	L	Person	TT	Dec	nonwords	familiar	Nonword Rate	NR	TOWRE	Y		
Petersen et al., 2016	320	0.34	NA	NA	L	Person	TT	Dec	nonwords	familiar	Nonword Rate	CR	DIBELS	Y		
Petersen et al., 2016	320	0.52	NA	NA	L	Person	TT	Dec	nonwords	familiar	Real Word Rate	NR	TOWRE	Y		
Petersen et al.,	320	0.49	NA	NA	L	Person	TT	Dec	nonwords	familiar	Nonword Rate	NR	TOWRE	Y		

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

2016																
Petersen et al., 2016	320	0.37	NA	NA	L	Person	TT	Dec	nonwords	familiar	Nonword Rate	CR	DIBELS	Y		
Petersen et al., 2016	320	0.54	NA	NA	L	Person	TT	Dec	nonwords	familiar	Real Word Rate	NR	TOWRE	Y		
Petersen et al., 2016	320	0.5	NA	NA	L	Person	TT	Dec	nonwords	familiar	Nonword Rate	NR	TOWRE	Y		
Petersen et al., 2016	320	0.4	NA	NA	L	Person	TT	Dec	nonwords	familiar	Nonword Rate	CR	DIBELS	Y		
Petersen et al., 2016	280	0.34	NA	NA	L	Person	TT	Dec	nonwords	familiar	Real Word Rate	NR	TOWRE	Y		
Petersen et al., 2016	280	0.35	NA	NA	L	Person	TT	Dec	nonwords	familiar	Nonword Rate	NR	TOWRE	Y		
Petersen et al., 2016	280	0.32	NA	NA	L	Person	TT	Dec	nonwords	familiar	Nonword Rate	CR	DIBELS	Y		
Petersen et al., 2016	280	0.27	NA	NA	L	Person	TT	Dec	nonwords	familiar	Real Word Rate	NR	TOWRE	Y		
Petersen et al., 2016	280	0.32	NA	NA	L	Person	TT	Dec	nonwords	familiar	Nonword Rate	NR	TOWRE	Y		
Petersen et al., 2016	280	0.44	NA	NA	L	Person	TT	Dec	nonwords	familiar	Nonword Rate	CR	DIBELS	Y		
Petersen et al., 2016	280	0.48	NA	NA	L	Person	TT	Dec	nonwords	familiar	Real Word Rate	NR	TOWRE	Y		
Petersen et al., 2016	280	0.5	NA	NA	L	Person	TT	Dec	nonwords	familiar	Nonword Rate	NR	TOWRE	Y		
Petersen et al., 2016	280	0.41	NA	NA	L	Person	TT	Dec	nonwords	familiar	Nonword Rate	CR	DIBELS	Y		

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

Petersen et al., 2016	280	0.55	NA	NA	L	Person	TT	Dec	nonwords	familiar	Real Word Rate	NR	TOWRE	Y		
Petersen et al., 2016	280	0.52	NA	NA	L	Person	TT	Dec	nonwords	familiar	Nonword Rate	NR	TOWRE	Y		
Petersen et al., 2016	280	0.53	NA	NA	L	Person	TT	Dec	nonwords	familiar	Nonword Rate	CR	DIBELS	Y		
Petersen et al., 2016	280	0.57	NA	NA	L	Person	TT	Dec	nonwords	familiar	Real Word Rate	NR	TOWRE	Y		
Petersen et al., 2016	280	0.56	NA	NA	L	Person	TT	Dec	nonwords	familiar	Nonword Rate	NR	TOWRE	Y		
Petersen et al., 2016	280	0.54	NA	NA	L	Person	TT	Dec	nonwords	familiar	Nonword Rate	CR	DIBELS	Y		
Petersen et al., 2018	370	0.46	NA	50.4	L	Person	TT	Dec	nonwords	familiar	Real Word Rate	CR	DIBELS	Y	10	
Petersen et al., 2018	370	0.45	NA	50.4	L	Person	TT	Dec	nonwords	familiar	Real Word Rate	CR	DIBELS	Y		
Petersen et al., 2018	370	0.44	NA	50.4	L	Person	TT	Dec	nonwords	familiar	Real Word Rate	CR	DIBELS	Y		
Petersen et al., 2018	370	0.42	NA	50.4	L	Person	TT	Dec	nonwords	familiar	Real Word Rate	CR	DIBELS	Y		
Petersen et al., 2018	370	0.48	NA	50.4	L	Person	TT	Dec	nonwords	familiar	Real Word Rate	CR	DIBELS	Y		
Petersen et al., 2018	370	0.46	NA	50.4	L	Person	TT	Dec	nonwords	familiar	Real Word Rate	CR	DIBELS	Y		
Petersen et al., 2018	370	0.45	NA	50.4	L	Person	TT	Dec	nonwords	familiar	Real Word Rate	CR	DIBELS	Y		
Petersen et al.,	370	0.42	NA	50.4	L	Person	TT	Dec	nonwords	familiar	Real Word Rate	CR	DIBELS	Y		

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

2018																
Petersen et al., 2018	370	0.41	NA	50.4	L	Person	TT	Dec	nonwords	familiar	Real Word Rate	CR	DIBELS	Y		
Petersen et al., 2018	370	0.4	NA	50.4	L	Person	TT	Dec	nonwords	familiar	Real Word Rate	CR	DIBELS	Y		
Petersen et al., 2018	370	0.38	NA	50.4	L	Person	TT	Dec	nonwords	familiar	Real Word Rate	CR	DIBELS	Y		
Petersen et al., 2018	370	0.38	NA	50.4	L	Person	TT	Dec	nonwords	familiar	Real Word Rate	CR	DIBELS	Y		
Petersen et al., 2018	370	0.5	NA	50.4	L	Person	TT	Dec	nonwords	familiar	Real Word Rate	CR	DIBELS	Y		
Petersen et al., 2018	370	0.49	NA	50.4	L	Person	TT	Dec	nonwords	familiar	Real Word Rate	CR	DIBELS	Y		
Petersen et al., 2018	370	0.48	NA	50.4	L	Person	TT	Dec	nonwords	familiar	Real Word Rate	CR	DIBELS	Y		
Petersen et al., 2018	370	0.44	NA	50.4	L	Person	TT	Dec	nonwords	familiar	Real Word Rate	CR	DIBELS	Y		
Sittner Bridges & Catts, 2011a	90	0.477	NA	NA	L	Person	GP	PA	words	NOT	Real Word Accuracy	NR	WRMT-R	Y	11	
Sittner Bridges & Catts, 2011a	90	0.516	NA	NA	L	Person	GP	PA	words	NOT	Nonword Accuracy	NR	WRMT-R/NU	Y		
Sittner Bridges & Catts, 2011b	96	0.485	NA	NA	L	Person	GP	PA	words	NOT	Real Word Accuracy	NR	WRMT-R	Y	11	
Sittner Bridges & Catts, 2011b	96	0.426	NA	NA	L	Person	GP	PA	words	NOT	Nonword Accuracy	NR	WRMT-R/NU	Y		
Spector, 1992	38	0.6	5;11	NA	L	Person	GP	PA	words	NOT	Real Word Accuracy	Inventory	San Diego Quick	Y	10	

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

													Assessment List			
Wyman-Chin, 2018	9	0.44	NA	66	C	Computer	TT	PA	nonwords	NOT	Real Word Accuracy	NR	WMLSR	N	9	
Wyman-Chin, 2018	9	0.87	NA	66	C	Computer	TT	PA	nonwords	NOT	Real Word Accuracy	NR	WMLSR	N		
Yap, 2018	99	0.6	5;3	43.4	L	Person	GP	PA	words	NOT	Real Word Accuracy	NR	WJ-III	N	11	
Yap, 2018	99	0.56	5;3	43.4	L	Person	GP	PA	words	NOT	Nonword Accuracy	NR	WJ-III	N		
Zumeta, 2010	37	0.59	6;2	49.4	C	Person	GP	PA	words	NOT	Real Word Rate	RD	Dolch word list	N	12	

Note. N= number of participants per study, r= Pearson's correlation coefficient / effect size, DA=dynamic assessment, WRM = Word reading measure, PR=Peer-reviewed, QAR=Quality Appraisal Rating, C = Cross-sectional, L=Longitudinal, TT=(Test)-teach-retest, GP=Graduated prompts, PA = phonological awareness, SSK = sound-symbol knowledge, Dec=Decoding, NR=Norm-referenced, CR=Criterion-referenced, RD=Researcher developed, NA=Not applicable/available, TOWRE=Test of word reading efficiency, DIBELS = Dynamic indicators of basic early literacy skills , WJ-III = Woodcock Johnson-III, WRAT = Wide Range Achievement Test, WRMT(R/NU)=Woodcock Reading Mastery Tests-Revised/Normative Update, BAS=British Ability Scales, SLRT-III=Salzburger lese und rechtschreibtest, SWRT=Single word reading test, TALE = Test de Análisis de la Lecto-escritura, HKT-SpLd=Hong Kong test of specific learning difficulties, EMT=One minute test, Y= Yes, N= No,

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

Analyses

All statistical analyses were conducted in R using the metafor package (Viechtbauer, 2010; R Core Team, 2021). First, a random effects meta-analysis with robust variance estimation (RVE) was conducted to examine the overall mean effect representing the association between DAs of word reading skills and WRMs. Random effects models assume that variability can stem from multiple sources, both individual sampling error and heterogeneity between studies, This is appropriate in this instance given that studies evaluated populations of different ages, from distinct locations using varied DAs and WRMs (Borenstein et al., 2010). We elected to use RVE because it permits inclusion of multiple effect sizes from a single study, while accounting for dependence between samples (Pustejovsky & Tipton, 2022). Prior to analysis, the 192 effect sizes were transformed into Z scores using Fisher Z transformation (Corey et al., 1998). Effect sizes are nested within study cluster, with the assumption that they intercorrelate both within and across clusters. A weighted average of these effect sizes was calculated, then transformed back to a Pearson's correlation coefficient for interpretation as an overall effect (Hedge's g). We then conducted five subgroup analyses to examine whether the strength of the association between DAs and WRMs differed across the test characteristics of skill type, administration method, format, word and symbol type. These subgroup analyses were planned a priori. To account for multiple comparisons, we used a Bonferroni adjustment to set a new value for significance (p=.006). An additional three subgroup analyses were conducted using the same data based on participant characteristics in a separate forthcoming study for a total of 8 subgroup analyses (Wood et al., 2023). The mixed effect model used in subgroup analysis allows us to determine if the mean effect sizes differ significantly based on DA characteristics.

Results

Research Question: Do DAs of word reading skills demonstrate consistent validity with word reading measures when stratified by word reading skill type, administration method, format, word, or symbol type?

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

As demonstrated in Table 1, 35 studies reported 192 correlations between a DA of word reading skills (phonological awareness, sound-symbol knowledge, or decoding) and a WRM. The effect sizes from these 35 studies were included in the random effects meta-analysis with robust variance estimation examining the relationship between DAs word reading skills and WRMs. The forest plot of these 192 effect sizes can be found in Figure 1 of the supplemental material. As anticipated, the overall mean effect size is large (r=.49, 95%CI = [0.42-0.55]) suggesting that DAs of word reading skills are strongly correlated with WRMs. This result was expected as measures of word reading skills like decoding, phonological awareness and sound symbol knowledge are known to correlate strongly with and be consistent predictors of later word reading ability. The RVE method used in this analysis assumes a consistent correlation (r) across study clusters (Pustejovsky & Tipton, 2022). We confirmed that this result was consistent across all values of r (.1 - 1.0). Results reported here are for r=.8. This three-level model considered variance from sampling (level 1), within study cluster (level 2) and between study cluster (level 3) heterogeneity. The total I^2 was 91.99%. Level 1 sampling error variance was negligible (8.01%), within cluster variance was moderate (30.15%) and between cluster variance was substantial (61.83%), suggesting that the largest proportion of variance is attributed to differences between study clusters. Between study heterogeneity was anticipated given the variation in dynamic assessment characteristics (e.g., administration method, format etc.). This heterogeneity was examined via subgroup analyses.

The contribution of individual effect sizes to heterogeneity was also examined via a Baujat plot, (Figure 2 in the Supplemental Material) (Baujat, 2002). Two effect sizes from Gellert and Elbro (2017a) were identified in the upper right quadrant of the plot, suggesting they contributed significantly to heterogeneity. No plausible reasons for this were identified based on the characteristics of the study. We reran the meta-analysis with these effect sizes removed, but this did not change magnitude of the overall effect (g=.49, 95% CI: 0.41 - 0.55, p=<0.0001) or the degree of heterogeneity (I^2=91.99%), and so we

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

elected to retain them in our analysis given that it is not advisable to remove studies without specific reason (Deeks et al., 2022).

Risk of Publication Bias

A funnel plot was generated to subjectively examine risk of publication bias (See Figure 3 in the Supplemental material). Visual inspection of the funnel plot suggests potential asymmetry. Several studies with small sample sizes and positive findings were identified and included, compared to studies with small sample sizes and negative findings (e.g., Horbach et al., 2018; Loreti, 2015) and are clustered at the bottom right of the funnel. There is a possibility that studies with negative outcomes were not completed, published, or submitted into the grey literature (Lee & Hotopf, 2012). However this may also be simply because the skills evaluated in the DAs included (phonological awareness, sound-symbol knowledge and decoding) are known to correlate with word reading performance, and so negative effects are not anticipated.. Furthermore, despite the apparent visual asymmetry, Egger's test was calculated and not found to be significant for presence of plot asymmetry (z=1.42, p=0.16) (Egger et al., 1997). Ultimately, there is minimal risk of publication bias in this analysis.

Subgroup Analyses

Subgroup analyses by DA word reading skill type, format, administration method word and symbol type, were planned a priori and were conducted to examine whether these characteristics have implications for the criterion reference validity of DAs with word reading measures. Mixed effects models were used to examine whether there were significant differences in mean effect sizes for DAs based on these factors. Results of these subgroup analyses are reported in Table 2.

Table 2.

Results of Subgroup Analysis by Skill Type, Format, Administration Method, Word and Symbol Type

	Effect size *(g)*	SE	95%CI	*p* value
Skill Type				**Subgroup *p*= <.0001***
Decoding	.56	.04	.51 - .62	<.0001
Phonological awareness	.47	.06	.30 - .61	.05

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

Sound-Symbol Knowledge	.32	.03	.18 - .43	<.0001
Format				**Subgroup *p*=<.0001***
Graduated Prompts	.54	.04	.49 - .59	<.0001
(Test) Teach – Retest	.36	.06	.19 - .56	<.0001
Administration Method				**Subgroup *p*=0.09**
In-Person	.51	.09	.23 - .91	.09
Computer	.39	.08	.25 - .52	<.0001
Word Type				**Subgroup *p*=<.0001***
Nonwords	.63	.05	. 56 – .69	<.0001
Words	.45	.06	.25 - .61	<.0001
Symbol Type				**Subgroup *p*=0.007**
Familiar	.59	.08	.48 -.68	<.0001
Novel	.38	.10	.04 - .64	.009

Note. SE=Standard Error, CI= Confidence Interval,

*= significant result

Adjusted *p* value = .006.

The adjusted significance value was set to *p*=.006 following our Bonferroni adjustment for multiple comparisons. Findings indicate that there are no significant differences in the strength of correlation between DAs of word reading skills and WRMs based on administration method (in-person vs via computer) (*p*=.09) or symbol type (novel vs familiar) (*p*=.007), though the latter approached significance. However, DAs that employed a graduated prompts format were significantly more strongly correlated with WRMs (*g*=.54) than those that used a (test)-teach-retest format (*g*=.36; *p*<.0001). Significant differences were also observed in terms of strength of correlation between DAs and WRMs based on the type of word reading skill used. Though multiple comparisons for each of the three subgroups were not completed, the mean effect sizes for DAs of decoding (*g*=.56) and phonological awareness (*g*=.47) were larger than DAs of sound-symbol knowledge (*g*=.32; *p*<.0001). Finally, outcomes reveal that DAs that use nonwords (*g*=.63) demonstrated significantly larger mean effect sizes than those that used real words (*g*=.45; *p*<.0001).

Discussion

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

This review examined whether characteristics of dynamic assessments (DAs) of word reading skills (phonological awareness, sound-symbol knowledge, and decoding) affected the criterion reference validity of DAs of word reading skills as measured by strength of correlational relationship with performance on a word reading measure (WRM). Thirty-five articles met inclusion criteria of evaluating children with a mean age between 4;0 and 10;0 and that reported a Pearson's correlation coefficient between a DA of a word reading skill and a WRM. As expected, results of the overall meta-analysis suggest that DAs of word reading are strongly correlated with WRMs.

Analysis of *word reading skill type* indicates that performance on DAs of decoding and phonological awareness (PA) is strongly correlated with performance on word reading measures, while performance on DAs of sound-symbol knowledge (SSK) is only moderately correlated. These results are in line with Dixon et al., (2022a), who in their systematic review documented that DAs of PA and decoding generally accounted for greater unique variance beyond static word reading accuracy performance (between 4-21%, and 1-17% respectively), and in nonword reading fluency (between 1-9% and 1-17% respectively), than SSK tasks, which accounted for only 2-6% unique variance in later word reading fluency and accuracy. These findings suggest that in DA, PA and decoding tasks may be better suited than SSK tasks to evaluate current and predict later word reading ability. This contrasts with SA, where letter-sound tasks have been found to demonstrate stronger correlations with word reading performance than PA tasks (e.g., National Early Literacy Panel, 2008). We speculate that the differences in relative strength of correlation between word reading skills of PA and SSK and word reading across static and dynamic domains may be due to the nature of the assessments themselves. In static assessments, children are evaluated on acquired knowledge. PA tasks may be less familiar, and therefore more challenging than SSK tasks. A static SSK task merely requires a child to recognize a letter and name the sound, while a PA task could range in complexity, from something simple like asking the child to segment a word into syllables, to something more challenging like manipulating phonemes in words. Since static PA tasks are often unfamiliar and complex, and children receive no prompting, teaching or

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

feedback in the assessment, many perform poorly on these measures at the earliest stages of learning to read. This results in floor effects, which weaken the predictive value of a measure. Indeed, previous research has documented that static PA tasks demonstrate floor effects for a longer period of time relative to letter naming or letter-sound tasks (Catts et al., 2009). However, in DA, the opposite may be true, given that clinicians and researchers are attempting to not only evaluate what the child already knows, but also how they learn. When presented with an unfamiliar complex task, such as a PA task, there is ample opportunity to evaluate *ability to learn*. However, a simple, more familiar SSK task may not create the opportunities necessary to demonstrate this ability. These results highlight the differences between static and dynamic assessments of early word reading skills. In SA, tasks must be simple enough to ensure that not all children will struggle given their limited acquired knowledge. However, in DA, tasks must be complex enough that children have the opportunity to learn in the context of the assessment. more complex

The analyses evaluating DA *format* found significant differences between the graduated prompts (GP) or train-test (TT) approaches in favour of the GP format. These results are consistent with findings from a previous review which found that DAs that used contingent feedback demonstrated stronger predictive validity than those that used non-contingent feedback (Caffrey et al., 2008). GP DAs are scripted and use contingent feedback, employing a series of pre-defined, increasingly explicit prompts following an examinee's response (e.g., Spector's (1992) use of 6 pre-defined prompts). Many of the TT DAs also used non-contingent feedback (e.g., Horbach et al., used non-contingent scripted verbal feedback). However, the feedback in the training and teaching phase of TT DAs was characterized by greater variability. For example, in Petersen and Gillam's (2015) study examining a DA of nonword decoding, examiners used noncontingent feedback to teach children how to read the nonwords if they were unsuccessful in the initial pre-test phase. This increased variability may have contributed to the weaker relationship between TT DAs and WRMs as the potential for individual differences in teaching

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

styles across examiners and researchers in TT formats is more likely to have had an influence on DA performance.

The analysis examining the role of *administration method* in DAs of word reading skills found that there were no significant differences between those administered in-person, vs. those conducted by a person via computer/computerized. These findings are consistent with a previous review which reported no significant differences between in-person and computer administration methods for static assessments (SAs) in pediatric language assessment (Alfano et al., 2022). Though the difference was not significant, a larger mean effect size was found for DAs administered in person. This could be a result of several factors. First, all WRMs were conducted in-person. It is promising that despite this difference in administration method, computer administered DAs still demonstrated strong mean effect sizes with in-person WRMs. Second, as posited earlier, it is possible that because DAs are characterized by increased interaction between examiner and examinee relative to SAs, administering them via computer may result in a reduced ability to engage in meaningful interaction or provision of accurate feedback. Similar challenges (e.g., technical issues disruption assessment, the need for caregiver support in evaluation and difficulties associated with providing feedback and maintaining child engagement) have been documented in the literature examining virtual use of SAs (e.g., Hodge et al., 2019, Wood et al., 2021). However, results from these studies and this review indicate that much like SAs, computer-based or virtual DAs are a valid alternative to in-person administration.

Regarding *word type* in DAs of word reading skills, results indicate that DAs that use nonwords are associated with significantly greater mean effect sizes than those that use real words. These results differ from previous findings in the SA literature, which suggested that nonwords were too distant from real word reading to be valid and were impractical for beginning readers who lack the necessary skills to decode (Wagner et al., 1997). However, as previously stated, it is possible that use of nonwords in DAs permits evaluation of a child's ability to learn, since all children are unfamiliar with them, and cannot use previous knowledge or experiences to recognize words in testing (Hoover & Tunmer, 1993). The

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

issue of lacking the necessary skills is also resolved, given that DAs incorporate teaching, feedback and prompting that parallels what a child might receive in a classroom context. This finding again highlights the differences between static and dynamic approaches. In a static approach, familiar material is required to evaluate acquired knowledge, but in DA, unfamiliar material, such as nonwords, may be better suited to evaluate learning potential.

This trend of unfamiliarity leading to increased capacity to evaluate ability to learn in DA was not reflected in the subgroup analysis by *symbol type*. There were no significant differences between DAs that used novel vs. familiar symbols. DAs that used familiar letters or characters were associated with larger mean effect sizes than those that used novel symbols, and this difference approached significance. We speculate that this may be a result of the types of symbols used in these DAs. For instance, some used real letters and characters from a different language in their test items (e.g., Aravena et al., 2013 used Hebrew characters in evaluating Dutch children). Others however, used symbols or that did not resemble any letter or character in an existing script (e.g., Horbach et al., used dots and dashes to represent the syllable-sound correspondences in their DA measure). True letters and characters, whether familiar or unfamiliar, exhibit features and characteristics that allow them to be differentiated from scribbles or symbols (Dehaene, 2009; Heimann, et al., 2013). It is possible that unfamiliar symbols that minimally resemble real characters or letters may be better suited to predict word reading ability.

Despite this, these results still have important implications for the validity of DAs of word reading skills. The use of nonwords and novel symbols has the potential to further reduce bias against children with diverse linguistic experiences. When words and symbols are unfamiliar to all, lack of experience with the oral and print system of the language of evaluation has less impact on performance. Results of this review support the use and development of nonword novel-symbol based DA measures. These types of tools could be used to equitably evaluate children with and without experience in the language of testing. This is in contrast with static tools, which are developed for and can only be used in a valid, unbiased manner with a single language group, typically a monolingual population. A DA of

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

word reading skills that uses nonwords and novel-symbols evaluates ability to learn word reading skills, while minimizing impact of previous linguistic experience. This type of tool and could act as an equitable alternative to language-specific SAs for culturally and linguistically diverse children for whom there are limited assessment tools.

Limitations

First, while we endeavored to examine relevant DA characteristics and their implications on validity, it is possible that there are other factors that may be contributing to the overall strength of relationship between DAs of word reading skills and word reading measures. One such factor is word reading measure type. In the Caffrey et al., review, (2008), authors examined whether type of outcome measure had implications for validity of DAs and reported that researcher developed tools demonstrated the largest mean effect size relative to norm and criterion-referenced measures, or teacher/clinician ratings. Regrettably, we were not able to replicate the subgroup analysis of word reading measure type fairly, since a large majority (156/192) of effect sizes represented correlations between DAs and a norm-referend or criterion-referenced WRM, while significantly fewer (35/192) used a researcher developed tool, and none used a teacher or clinician rating. Second, correlation coefficients were selected as the measure of effect size because they are consistently reported. While this allowed for inclusion of additional studies, as a result, only correlational inferences can be made about results. Finally, it is possible that relevant studies may not have been identified because they were published in a language that our review team was not able to read (e.g., many studies in Korean and Hebrew were excluded in the title and abstract screening phase), or because they used key terms not captured by our search strategy.

Clinical Implications

The results of this systematic review and meta-analysis have implications for clinicians like speech-language pathologists and psychologists and educators who routinely evaluate word reading skills. Outcomes suggest that, when possible, clinicians should favour DAs of phonological awareness and decoding skills, that are structured in a graduated prompts format, and that use nonwords comprised

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

of familiar or novel symbols. Results indicate that these measures can be conducted in-person or virtually, which is particularly relevant post-pandemic, as many professionals continue to evaluate children in a virtual context. An example to consider that meets most of these criteria and that is available for free online is the CUBED-3 dynamic decoding measure (DDM). The DDM, a measure developed based on the included studies conducted by Petersen et al., (2016) and Petersen & Gillam (2015), uses a test teach retest approach rather than a graduated prompts approach but evaluates both phonological awareness and decoding, uses nonwords comprised of familiar letters and is administered in person. This measure can be used as a criterion-referenced screening tool, to set intervention targets or to monitor progress (CUBED-3 Dynamic Decoding Measure (DDM), Petersen & Spencer, 2023).

Tool Development and Suggested Future Research

Results of this study also inform development of novel DAs of word reading skills, or revisions of existing tools. Notably, findings support the virtual administration of DAs of word reading skills. This is an important consideration given that there has been a significant increase in the provision of virtual care and tele-assessment since the COVID-19 pandemic (Campbell & Goldstein, 2021). Researchers should consider developing virtual versions of DAs to ensure that assessment of these critical early reading skills can be made available in any future disruptions to in-person education or clinical services, or simply to ensure that children living in rural or remote areas have access to high quality, equitable dynamic assessments. Study outcomes also support the use of dynamic assessments of word reading skills that use nonwords and novel symbols. Not only does unfamiliarity of words and symbols lend itself well to a DA task where the goal is to evaluate ability to learn, this type of measure may also minimize linguistic and cultural bias associated with traditional static word reading skill tasks. Developing new DAs or revising existing measures to include nonword, novel-symbol based versions is critical for monolingual and bilingual children for whom there are no language specific tests of word reading skills available.

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

Beyond this, future studies can also directly compare DAs with differing characteristics, using research designs and statistical analyses that permit a better understanding of the causal role these factors play in predicting reading ability and identifying reading disorder at various timepoints in a child's journey of learning to read. This can be achieved through longitudinal studies comparing the relative predictive validity of DAs that differ in their format, administration method, word and symbol type, or other relevant factors via regression or structural equation modelling. Studies should also explicitly examine whether specific characteristics of DAs of word reading skills have a greater capacity to limit floor effects associated with traditional static measures or result in improved diagnostic accuracy. Ideally, these studies should include populations for whom DA is purported to be most useful, particularly bilingual children and those with limited previous literacy experiences.

Acknowledgements

This study was supported by a Canada Graduate Scholarship-Master's grant from the Social Sciences and Humanities Research Council of Canada, at the Rehabilitation Sciences Institute at the University of Toronto, an Ontario Graduate Scholarship from the Ministry of Colleges and Universities awarded to E. Wood, a University of Toronto Excellence Award awarded to K. Biggs, and a Natural Sciences and Engineering Research Council of Canada grant awarded to Dr. M. Molnar.

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

References

Ackerman, D. J., & Barnett, W.S. (2005). *Prepared for Kindergarten: What does "readiness" mean?* National Institute for Early Education Research. https://nieer.org/policy-issue/policy-report-prepared-for-kindergarten-what-does-readiness-mean

Alfano, A. R., Concepcion, I., Espinosa, A. & Menendez, F. (2022). Pediatric language assessments via telehealth: A systematic review. *Journal or Telemedicine and Telecare.* https://doi.org/1357633X221124998.

*Aravena, S., Snellings, P., Tijms, J., & van der Molen, M. W. (2013). A lab-controlled simulation of a letter–speech sound binding deficit in dyslexia. *Journal of Experimental Child Psychology, 115*(4), 691-707. https://doi.org/10.1016/j.jecp.2013.03.009

*Aravena, S., Tijms, J., Snellings, P., & van der Molen, M. W. (2018). Predicting individual differences in reading and spelling skill with artificial script–based letter–speech sound training. *Journal of Learning Disabilities, 51*(6), 552-564. https://doi-org. /10.1177/0022219417715407

*Barker, R. M., & Saunders, K. J. (2020). Validity of a nonspeech dynamic assessment of the alphabetic principle in preschool and school-aged children. *Augmentative and Alternative Communication, 36*(1), 54-62. https://doi-org./10.1080/07434618.2020.1737965

Baujat, B., Mahé, C., Pignon, J. P., & Hill, C. (2002). A graphical method for exploring heterogeneity in meta-analyses: Application to a meta-analysis of 65 trials. *Statistics in Medicine 21*(18):2641-52. https://doi-org./10.1002/sim.1221

Bedore, L. M., & Peña, E. D. (2008). Assessment of bilingual children for identification of language impairment: Current findings and implications for practice. *International Journal of Bilingual Education and Bilingualism, 11*(1), 1-29. https://doi-org. /10.2167/beb392.0

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research synthesis methods, 1*(2), 97–111. https://doi-org./10.1002/jrsm.12

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

Brown, A. L., & Ferrara, R. A. (1985). Diagnosing zones of proximal development. In J. Wertsch (Ed.) *Culture, communication, and cognition: Vygotskian perspectives*, (pp.273-305). Cambridge University Press.

Budoff, M. (1987). The validity of learning potential assessment. In C.S. Lidz (Ed.) *Dynamic assessment: An Interactional approach to evaluating learning potential.* (pp.173-195). New York: Guildford Press.

*Caffrey, E. (2006). A comparison of dynamic assessment and progress monitoring in the prediction of reading achievement for students in kindergarten and first grade (Doctoral dissertation, Vanderbilt University).

Caffrey, E., Fuchs, D., & Fuchs, L. S. (2008). The predictive validity of dynamic assessment: A review. *The Journal of Special Education, 41*(4), 254-270. https://doi-org./10.1177/0022466907310366

Campbell, D.R., & Goldstein, H. (2021). Genesis of a new generation of telepractitioners: The COVID-19 pandemic and pediatric speech-language pathology services. *American Journal of Speech-Language Pathology,* 30(5), 2143-2154. https://doi-org.10.1044/2021_AJSLP-21-00013

Campbell, D. R., & Goldstein, H. (2022). Evolution of telehealth technology, evaluations, and therapy: Effects of the COVID-19 pandemic on pediatric speech-language pathology services. *American Journal of Speech-Language Pathology*, *31*(1), 271-286. https://doi.org/ 10.1044/2021_AJSLP-21-00069

Castles, A., Polito, V., Pritchard, S., Anandakumar, T., & Coltheart, M. (2018). Do nonword reading tests for children measure what we want them to? An analysis of year 2 error responses. *Australian Journal of Learning Difficulties*, *23*(2), 153-165. https://doi.org/10.1080/19404158.2018.1549088

Catts, H. W., Hogan, T.P., & Adlof, S.M. (2005). Developmental changes in reading and reading disabilities. In H. Catts & A. Kamhi (Eds.) *Connections between language and reading disabilities.* (pp.25-40). Mahwah, NJ. Erlbaum.

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

Catts, H. W., Petscher, Y., Schatschneider, C., Sittner Bridges, M., & Mendoza, K. (2009). Floor effects associated with universal screening and their impact on the early identification of reading disabilities. *Journal of Learning Disabilities, 42*(2), 163–176. https://doi-org./10.1177/0022219408326219

*Cho, E., & Compton, D. L. (2015). Construct and incremental validity of dynamic assessment of decoding within and across domains. *Learning and Individual Differences, 37*, 183-196. https://doi.org/10.1016/j.lindif.2014.10.004

*Cho, E., Compton, D. L., Fuchs, D., Fuchs, L. S., & Bouton, B. (2014). Examining the predictive validity of a dynamic assessment of decoding to forecast response to tier 2 intervention. *Journal of Learning Disabilities, 47*(5), 409-423. https://doi-org. /10.1177/0022219412466

*Cho, E., Compton, D. L., Gilbert, J. K., Steacy, L. M., Collins, A. A., & Lindström, E. R. (2017). Development of first-graders' word reading skills: For whom can dynamic assessment tell us more? *Journal of Learning Disabilities, 50*(1), 95-112. https://doi-org.\ /10.1177/0022219415599343

*Chow, B. W. Y. (2014). The differential roles of paired associate learning in Chinese and English word reading abilities in bilingual children. *Reading and Writing, 27*(9), 1657-1672. https://doi.org/10.1007/s11145-014-9514-3

*Clayton, F. J., Sears, C., Davis, A., & Hulme, C. (2018). Verbal task demands are key in explaining the relationship between paired-associate learning and reading ability. *Journal of Experimental Child Psychology, 171*, 46-54. https://doi.org/10.1016/j.jecp.2018.01.004

*Compton, D. L., Fuchs, D., Fuchs, L. S., Bouton, B., Gilbert, J. K., Barquero, L. A., ... & Crouch, R. C. (2010). Selecting at-risk first-grade readers for early intervention: Eliminating false positives and exploring the promise of a two-stage gated screening process. *Journal of Educational Psychology, 102*(2), 327. https://doi-org. /10.1037/a0018448

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

Corey, D. M., Dunlap, W. P., & Burke, M. J. (1998). Averaging correlations: Expected values and bias in combined Pearson Rs and Fisher's z transformations. *The Journal of general psychology*, *125*(3), 245-261. https://doi-org. /10.1080/00221309809595548

*Coventry, W. L., Byrne, B., Olson, R. K., Corley, R., & Samuelsson, S. (2011). Dynamic and static assessment of phonological awareness in preschool: A behavior-genetic study. *Journal of Learning Disabilities*, *44*(4), 322-329. https://doi-org. 10.1177/0022219411407862

Covidence (2023). Veritas Health Innovation, Melbourne, Australia. Available at www.covidence.org.

*Cunningham, A., & Carroll, J. (2011). Age and schooling effects on early literacy and phoneme awareness. *Journal of Experimental Child Psychology*, *109*(2), 248-255. https://doi.org/10.1016/j.jecp.2010.12.005

Deeks, J. J., Higgins, J. P. T., & Altman, D. G. (2023). Chapter 10: Analysing data and undertaking meta-analyses. In: P.T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page & V. A. Welch. (Eds). *Cochrane Handbook for Systematic Reviews of Interventions* version 6.4. Cochrane. Available from www.training.cochrane.org/handbook.

Dehaene, S. (2009). *Reading in the brain. The new science of how we read.* Penguin Books.

Dixon, C., Oxley, E., Gellert, A. S., & Nash, H. (2022a). Dynamic assessment as a predictor of reading development: a systematic review. *Reading and Writing*, 1-26. https://doi-org. /10.1007/s11145-022-10312-3

Dixon, C., Oxley, E., Nash, H., & Gellert, A. S. (2022b). Does dynamic assessment offer an alternative approach to identifying reading disorder? A systematic review. *Journal of Learning Disabilities*. https://doi-org./10.1177/00222194221117510

*Edwards, A. (2020). Predictor Importance in Future and Concurrent Predictions of Oral Reading Fluency. (Thesis, Florida State University). https://doi.org/10.31234/osf.io/4apbu

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

Egger, M., Smith, G. S., Schneider, M. & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal 315*(7109):629-634. https://doi-org. /10.1136/bmj.315.7109.629

Ehri, L. C., & Wilce, L.S. (1985). Movement into reading: Is the first stage of printed word learning visual or phonetic? *Reading Research Quarterly, 20*(2), 163-179.https://doi.org/10.2307/747753

Ehri, L. C. (1998). Grapheme–phoneme knowledge is essential for learning to read words in English. In *Word recognition in beginning literacy* (pp. 3-40). Routledge.

Elbro, C., & Scarborough, H. (2003). Early identification. In T. Nunes & P. Bryant (Eds.), *Handbook of Children's Literacy*. (pp.339-360). Kluwer Academic Publishers.

*Fuchs, D., Compton, D. L., Fuchs, L. S., Bouton, B., & Caffrey, E. (2011). The construct and predictive validity of a dynamic assessment of young children learning to read: Implications for RTI frameworks. *Journal of learning disabilities*, *44*(4), 339-347.

*Gan, Y., Zhang, J., Kahrabi-Yamato, L., Su, Y., Zhang, J., Jiang, Y., Hui, Y., & Li, H. (2022). The unique predictive value of dynamic assessment of character decoding in reading development of Chinese children from grades 1-2. *Scientific Studies of Reading*, 1-17. https://doi.org/10.1080/10888438.2022.2143271

*Gellert, A. S., & Elbro, C. (2017a). Does a dynamic test of phonological awareness predict early reading difficulties? A longitudinal study from kindergarten through grade 1. *Journal of learning disabilities*, *50*(3), 227-237. https://doi-org. /10.1177/002221941560918

*Gellert, A. S., & Elbro, C. (2017b). Try a little bit of teaching: A dynamic assessment of word decoding as a kindergarten predictor of word reading difficulties at the end of grade 1. *Scientific Studies of Reading*, *21*(4), 277-291. https://doi-org. /10.1080/10888438.2017.1287187

*Gellert, A. S., & Elbro, C. (2018). Predicting reading disabilities using dynamic assessment of decoding before and after the onset of reading instruction: a longitudinal study from kindergarten through grade 2. *Annals of Dyslexia*, *68*, 126-144.

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

* Gillam, S. L., Fargo, J., Foley, B., & Olszewski, A. (2011). A nonverbal phoneme deletion task administered in a dynamic assessment format. *Journal of Communication Disorders*, *44*(2), 236-245. https://doi.org/10.1016/j.jcomdis.2010.11.003

Ginsborg, J. (2006). The effects of socio-economic status on children's language acquisition and use. In J. Clegg & J. Ginsborg (Eds.) *Language and social disadvantage: Theory into practice*. (pp.9-27). Hoboken, NJ: Wiley Press.

Grigorenko, E. L., & Sternberg, R. J. (1998). Dynamic testing. *Psychological Bulletin*, *124*(1), 75-111.

Gruhn, S., Segers, E., Keuning, J., & Verhoeven, L. (2020). Profiling children's reading comprehension: A dynamic approach. *Learning and individual differences*, *82*, 101923.

Heimann, K., Umilta, M. A., & Gallese, V. (2013). How the motor-cortex distinguishes among letter, unknown symbol, and scribbles. A high-density EEG study. *Neuropsychologia*, *51*(13), 2833-2840. https://doi.org/10.1016/j.neuropsychologia.2013.07.014

Hodge, M. A., Sutherland, R., Jeng, K., Bale, G., Batta, P., Cambridge, A., ... & Silove, N. (2019). Literacy assessment via telepractice is comparable to face-to-face assessment in children with reading difficulties living in rural Australia. *Telemedicine and e-Health*, *25*(4), 279-287.

Hogan, T. P., Catts, H. W., & Little, T. D. (2005). The Relationship Between Phonological Awareness and Reading: Implications for the Assessment of Phonological Awareness. *Language, Speech & Hearing Services in Schools*, *36*, 4. https://doi-org. /10.1044/0161-1461(2005/029)

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing*, *2*(2), 127-160. http://dx.doi.org/10.1007/BF00401799

Hoover, W. A., & Tunmer, W. E. (1993). The components of reading. In G.B. Thompson, W.E. Tunmer, & T. Nicholson (Eds.) *Reading acquisition processes* (pp.1-19). Multilingual Matters.

*Horbach, J., Scharke, W., Cröll, J., Heim, S., & Günther, T. (2015). Kindergarteners' performance in a sound–symbol paradigm predicts early reading. *Journal of Experimental Child Psychology*, 139, 256-264. https://doi.org/10.1016/j.jecp.2015.06.007

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

*Horbach, J., Weber, K., Opolony, F., Scharke, W., Radach, R., Heim, S., & Günther, T. (2018). Performance in sound-symbol learning predicts reading performance 3 years later. *Frontiers in psychology*, 9. https://doi-org. /10.3389/fpsyg.2018.01716

Lantolf, J. P. & Poehner, M.E. (2004). Dynamic assessment: bringing the past into the future. *Journal of Applied Linguistics*, 1, 49–74. https://doi.org/ 10.1558/japl.1.1.49.55872

*Law, J. M., De Vos, A., Vanderauwera, J., Wouters, J., Ghesquière, P., & Vandermosten, M. (2018). Grapheme-phoneme learning in an unknown orthography: a study in typical reading and dyslexic children. *Frontiers in psychology*, 9. https://doi-org. /10.3389/fpsyg.2018.01393

Lee, W., & Hotopf, M. (2012). 10-Critical appraisal: Reviewing scientific evidence and reading academic papers. In P. Wright, J. Stern, & M. Phelan (Eds.) *Core Psychiatry (Third Edition)* pp. 131-142. Elsevier Ltd. https://doi.org/ 10.1016/B978-0-7020-3397-1.00010-0

*Liu, C., Chung, K. K. H., Wang, L. C., & Liu, D. (2021). The relationship between paired associate learning and Chinese word reading in kindergarten children. *Journal of Research in Reading, 44*(2), 264-283. https://doi-org. /10.1111/1467-9817.12333

*Loreti, B. (2015). Validity of a Spanish nonspeech dynamic assessment of phonological awareness in children from Spanish-speaking backgrounds. (Master's dissertation, University of South Florida).

McHugh M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, *22*(3), 276–282.

Melby-Lervåg, M., Lyster, S. A. H., & Hulme, C. (2012). Phonological skills and their role in learning to read: a meta-analytic review. *Psychological bulletin*, *138*(2), 322.

Moola, S., Munn, Z., Tufanaru, C., Aromataris, E., Sears, K., Sfetcu, R., Currie, M., Lisy, K., Qureshi, R., Mattis, P., & Mu, P. (2020). Chapter 7: Systematic reviews of etiology and risk. In: Aromataris E and Munn, Z. (Eds.) *JBI Manual for Evidence Synthesis*. JBI. https://doi.org/ DOI: 10.46658/JBIMES-20-08

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

National Early Literacy Panel. (2008). Developing Early Literacy: A Scientific Synthesis of Early Literacy Development and Implications for Intervention.

Navarro, J. J., Mourgues-Codern, C., Guzmán, E., Rodríguez-Ortiz, I. R., Conejo, R., Sánchez-Gutiérrez, C., ... & Saracostti, M. (2018). Integrating curriculum-based dynamic assessment in computerized adaptive testing: Development and predictive validity of the EDPL-BAI battery on reading competence. *Frontiers in Psychology*, *9*, 1492.

Ortiz, J. A. (2021). Using nonword repetition to identify language impairment in bilingual children: A meta-analysis of diagnostic accuracy. *American Journal of Speech-Language Pathology, 30*(5), 2275-2295. https://doi.org/10.1044/2021_AJSLP-20-00237

*Osa Fuentes, P. M. D. L. (2003). Evaluación dinámica del procesamiento fonológico en el inicio lector. (Doctoral Dissertation, Universidad de Granada).

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C.D., (2021) The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *The British Medical Journal*. https://doi.org/10.1136/bmj.n71

Peña, E., Gillam R.B., & Bedore, L.M. (2014). Dynamic assessment of narrative ability in English accurately identifies language impairment in English language learners. *Journal of Speech, Language and Hearing Research, 57*(6), 2208-2220. https://doi.org/10.1044/2014_JSLHR-L-13-0151

Peña, E., Iglesias, A., & Lidz, C. S. (2001). Reducing test bias through dynamic assessment of children's word learning ability. *American Journal of Speech-Language Pathology, 10*(2), 138-154.

*Petersen, D. B., Allen, M. M., & Spencer, T. D. (2016). Predicting reading difficulty in first grade using dynamic assessment of decoding in early kindergarten: A large-scale longitudinal study. *Journal of Learning Disabilities, 49*(2), 200-215. https://doi.org/10.1177/0022219414538518

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

Petersen, D. B., & Gillam, R. B. (2013). Accurately predicting future reading difficulty for bilingual Latino children at risk for language impairment. *Learning Disabilities Research & Practice, 28*(3), 113-128. https://doi.org/ 0.1111/ldrp.12014

*Petersen, D. B., & Gillam, R. B. (2015). Predicting reading ability for bilingual Latino children using dynamic assessment. *Journal of Learning Disabilities, 48*(1), 3-21. https://doi.org/10.1177/0022219413486930

*Petersen, D. B., Gragg, S. L., & Spencer, T. D. (2018). Predicting reading problems 6 years into the future: Dynamic assessment reduces bias and increases classification accuracy. *Language, Speech, and Hearing Services in Schools*, *49*(4), 875-888.

Petersen, D.B. & Spencer, T.D. (2023). *CUBED-3 Dynamic Decoding Measure (DDM).* Language Dynamics Group.

Poehner, M. E. (2008). *Dynamic assessment: A Vygotskian approach to understanding and promoting L2 development* (Vol. 9). Springer Science & Business Media.

Pustejovsky, J. E., & Tipton, E. (2022). Meta-analysis with robust variance estimation: Expanding the range of working models. *Prevention Science*, *23*(3), 425-438.

R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Robertson, C., & Salter, W. (2017). *Phonological awareness test, second edition: Normative update (PAT-2: NU).* PAR Inc.

Scarborough, H. S. (2001). Connecting early language and literacy to later reading (dis)abilities: Evidence, theory, and practice. In S. Neuman & D. Dickinson (Eds.), *Handbook for research in early literacy* (pp. 97–110). Guilford Press.

Shapiro, L. R., Carroll, J. M., & Solity, J. E. (2013). Separating the influences of prereading skills on early word and nonword reading. *Journal of experimental child psychology*, *116*(2), 278-295. https://doi.org/ 10.1016/j.jecp.2013.05.011

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

*Sittner Bridges, M., & Catts, H. W. (2011). The use of a dynamic screening of phonological awareness to predict risk for reading disabilities in kindergarten children. *Journal of Learning Disabilities, 44*(4), 330-338. ttps://doi-org./ 10.1177/0022219411407863

*Spector, J. E. (1992). Predicting progress in beginning reading: Dynamic assessment of phonemic awareness. *Journal of Educational Psychology, 84*(3), 353-364. https://doi.org/10.1037/0022-0663.84.3.353

Swanson, H. L. (1994). The role of working memory and dynamic assessment in the classification of children with learning disabilities. *Learning Disabilities Research & Practice, 9*(4), 190-202.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1-48. doi:10.18637/jss.v036.i03.

Wagner, R. K., Torgesen, J. K., Rashotte, C. A., Hecht, S. A., Barker, T. A., Burgess, S. R., ... & Garon, T. (1997). Changing relations between phonological processing abilities and word-level reading as children develop from beginning to skilled readers: a 5-year longitudinal study. *Developmental psychology*, *33*(3), 468-479. https://doi.org/ 10.1037//0012-1649.33.3.468

Wagner, R. K., Torgesen, J. K., Rashotte, C. A., & Pearson, N. A. (2013). *CTOPP-2: Comprehensive test of phonological processing-2*. Austin: Pro-ed.

Wolf Nelson, N. & Plante, E. (2022). Evaluating the equivalence of telepractice and traditional administration of the Test of Integrated Language and Literacy Skills. *Language Speech and Hearing Services in Schools*, *53*(2), 376-390.https://doi-org.10.1044/2022_LSHSS-21-00056

Woodcock, R. (2011). *The Woodcock Reading Mastery Tests – Third Edition (WRMT-III).* Bloomington: Pearson.

Wood, E., Bhalloo, I., McCaig, B., Feraru, C., & Molnar, M. (2021). Towards development of guidelines for virtual administration of paediatric standardized language and literacy assessments: Considerations for clinicians and researchers. *SAGE Open Medicine, 9*, 20503121211050510.

medRxiv preprint doi: https://doi.org/10.1101/2023.03.20.23287486; this version posted November 27, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

Wood, E., Biggs, K., & Molnar, M. (2023). Dynamic assessment of word reading skills with diverse school-aged children: A systematic review and meta-analysis. https://doi.org/10.1101/2022.09.19.22279942

Wood, E., & Molnar, M. (2022, March 4). Screening Protocol for a Systematic Review and Meta-Analysis of Dynamic Assessment of Early Literacy Skills in Children: Concurrent and Predictive Validity.

*Wyman Chin, K. R. (2018). Validity of a dynamic assessment of phonological awareness in emergent bilingual children. (Master's Dissertation, University of South Florida).

*Yap, D. F. F. (2018). The Utility of Dynamic Assessment of Phonological Awareness for Bilingual Children in Singapore. (Doctoral Dissertation, San Francisco State University & University of California, Berkeley).

*Zumeta, R. O. R. (2010). Enhancing the accuracy of kindergarten screening. (Doctoral Dissertation, Vanderbilt University.