Characteristics of dynamic assessments of word reading skills and their implications on 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

Abstract

Purpose: Dynamic assessments (DAs) of word reading skills (e.g., phonological awareness, decoding) demonstrate predictive validity with word reading outcomes but are characterized by substantial heterogeneity in terms of format, administration method, word, and symbol type used, factors which may affect their validity. This systematic review and meta-analysis examined whether the validity of DAs of word reading skills is affected by these characteristics. **Method:** Five electronic databases (Medline, Embase, PsycINFO, ERIC and CINAHL), 3 preprint repositories (MedRxiv, PsyArxiv and EdArxiv) and the gray literature were searched between March 2022 and March 2023, to identify studies with participants aged 4-10 that reported a Pearson's correlation coefficient between a DA of word reading and a word reading measure. A random effects meta-analysis and 4 subgroup analyses based on DA format, administration method, word and symbol type were conducted. **Results:** Thirty-two studies from 30 articles were identified. The overall effect size between DAs of word reading skills and word reading is large. There are no significant differences in mean effect sizes based on format (graduated prompt vs. train-test) or administration method (computer vs. in-person). However, DAs that use nonwords and those that use familiar letters or characters demonstrate significantly stronger correlations with word reading measures, than those that use real words and those that use novel symbols. **Conclusions:** Outcomes provide preliminary evidence to suggest that DAs of word reading skills that use nonwords and familiar letters in their test items are more strongly associated with later word reading ability than those that use real words or novel symbols. There were no significant differences between DAs administered in-person versus via computer. Results inform development of novel DAs of word reading, and clinical practice when it comes to selecting assessment tools.

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

Introduction

Literacy Assessment

Literacy, the ability to read and write, is a complex construct which requires integration of multiple skills but can simply be described as the product of the ability to decode words (or read words) and comprehend language (Hoover & Gough, 1990). In this review, we derive our definition of the construct of word reading from the subskills that comprise word recognition ability in the evidence-based model Scarborough's reading rope (2001). These subskills – phonological awareness, knowledge of the alphabetic principle (or sound-symbol knowledge) and word recognition (or decoding) ability, have been consistently found to be among the strongest and most accurate predictors of later reading ability for young children beginning to learn to read (e.g., Catts et al., 2005; Hogan et al., 2005; Scarborough, 1998;). In recent years, early assessment, and identification of difficulties with word reading skills has received increasing attention, in large part due to widespread global literacy challenges exacerbated by the COVID-19 pandemic (e.g., Annie E. Casey Foundation, 2014; OHRC (Ontario Human Rights Commission), 2022; The Conference Board of Canada, 2014; UNESCO (United Nations Educational, Scientific and Cultural Organization), 2013; UNESCO, 2021).

In the fields of speech-language pathology (SLP), psychology and education, many of the widely used traditional 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 measures an individual's learning product via a correct or incorrect binary grading system, and passively evaluates performance without provision of prompting or corrective feedback (Grigorenko & Sternberg, 1998). In this type of testing, children from, diverse linguistic background or those with limited literacy experiences are all prone to perform poorly (Bedore & Peña, 2008; Ginsborg, 2006; Sewell, 1987). When many children underperform on a test, it results in floor effects, which weaken the validity of a measure and render it difficult to discern those who are truly

at-risk from who have simply has not had enough linguistic or educational experience to perform test tasks. This can result in failure to identify word reading difficulties early and provide intervention to prevent the long-term negative effects associated with word reading difficulty (Catts et al., 2009).

Given the limitations associated with SAs, interest in alternative approaches to early word reading assessment have been increasing. Dynamic assessment (DA) is a potential solution. While SAs purport to quantify what a child is capable of at the time of testing, DAs endeavor to examine a child's current performance and their ability to learn a skill with support (Grigorenko & Sternberg, 1998). DA is characterized by provision of prompts, feedback, and interaction between the child and examiner within the assessment (Caffrey, 2006). This approach has been shown to reduce bias in testing and misidentification of difficulty because the impact of previous linguistic or educational experiences on test outcomes are minimized (Bedore & Peña, 2008; Petersen & Gillam, 2013). In the domain of word reading skills, DAs have been shown to predict unique variance beyond SAs in reading outcomes (Dixon et al., 2022a), to contribute to the accurate identification of reading difficulties (Dixon et al., 2022b), and to demonstrate strong concurrent validity with equivalent SAs and predictive validity with word reading outcome measures across typically developing, at-risk, bilingual, and monolingual children (Wood et al., 2023). However, these DAs are characterized by substantial heterogeneity in terms of their format, administration method, word, and symbol type. The impact of these characteristics on the strength of a DA measure's relationship with word reading measures has yet to be considered. The DA characteristics of interest are described below.

Dynamic Assessment Characteristics

Format

Broadly, there are two approaches to DA, interventionist, and interactionist (Lantolf & Poehner, 2004). Interactionist or contingent DA is typically unscripted and endeavors to modify cognitive or skill

ability. In this approach the examiner responds to the individual examinee and their capacities. Interventionist or non-contingent DA, however, more closely parallels traditional SA testing. The examiner provides pre-defined and increasingly explicit levels of support in response to student need. Its scripted nature requires less clinical skill and time to administer, and its standardization permits researchers to evaluate its validity (Poehner, 2008). In the field of word reading assessment, most studies have focused on developing and validating DAs that can be characterized as interventionist. The studies included in this review focus on two types of interventionist DA, which are utilized with similar frequency in DAs of word reading skills (Dixon et al., 2022b; Sternberg & Grigorenko, 2002).

The first is an approach pioneered by Milton Budoff (1987), that follows a (pre-test), train, re-test structure, referred to in this paper as the train/test (TT) format. This TT design consists of a static pre-test, followed by a dynamic teaching/training phase, and a final static post-test, though not all assessments incorporate the initial static pre-test. During the training phase, children receive feedback and instruction (e.g., encouragement to try again, a hint, or the correct solution, etc.). If a pre-test is conducted, outcomes of the post-test are compared to the child's initial score to assess the difference in their performance following the teaching session. If no pre-test is conducted, the post-test serves as a measure of how a child performs a skill after receiving explicit dynamic instruction in the task.

In contrast, the second format combines teaching and testing phases of the assessment within each item (Campione & Brown, 1987). In this approach, referred to as the graduated prompts (GP) format. Children are provided with feedback about whether they were correct or incorrect following their response. If incorrect, a series of increasingly explicit prompts are provided, until the child answers correctly, or all prompts are exhausted. Scoring is directly influenced by the individual's performance; the greater number of prompts required, the lower the score on an item (Brown & Ferrara, 1985; Campione et al., 1984). A meta-analysis found that interventionist DAs demonstrate stronger predictive validity than

interactionist DAs (Caffrey et al., 2008), and a recent systematic review documented that within interventionist DAs, the GP and TT formats are used with similar frequency in assessment of word reading (Dixon et al., 2022b), but to date there has been no consideration of whether these different formats result in increased strength of relationship between the DA and word reading outcomes.

Administration Method

Assessments, dynamic or otherwise, can be conducted in-person or via computer. Development of virtual or computer-based assessments of early literacy has become increasingly important, in the wake of the COVID-19 pandemic and the subsequent shift to distance/remote learning (Campbell & Goldstein, 2022; Tohidast et al., 2020). There is evidence to suggest that there are no significant differences between administering an SA online vs. in-person (e.g., Alfano et al., 2022; Nelson & Plante, 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 through a computer (e.g., Barker & Saunders, 2020), or in a computerized fashion where no examiner is required to carry out the assessment (e.g., Aravena et al., 2018). As previously noted, DA is characterized by increased interaction between examiner and examinee, and as a result may be impacted to a greater extent by computer administration. Post-pandemic, many clinicians and researchers continue to operate virtually or in a hybrid format and therefore the factor of administration method should be considered.

Word Type

Assessments of the word reading skills of phonological awareness and decoding can also be differentiated by the type of words used in their items (i.e., real or nonwords). Nonwords do not exist in the language of testing but that abide by its phonotactic and orthotactic constraints (e.g., “meeb” in English). Commercially available SAs that are in widespread clinical and research use, 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 reading and decoding, include subtests with both words and nonwords. Similarly, commercially developed DAs, like the CUBED dynamic test of decoding include a word reading and nonword decoding measure(Petersen et al., 2016). However, many DAs have fewer subtests and are at present, primarily used in research. These tests tend to employ either words or nonwords, not both. For example, Gellert and Elbro's phoneme identification task uses real words (2017b), but their decoding measure uses nonwords (2017a).

In young children, reading words vs. non-words are purported to tap into two different reading processes (Shapiro et al., 2013). There is evidence indicating that children may initially recognize some familiar, high frequency words by sight without activating their decoding skills (Ehri & Wilce, 1985). For example, children can often recognize their names or high frequency words like "the" in print without using knowledge of sound-symbol correspondences, phoneme blending and decoding skills. However, when it comes to reading nonwords, decoding skills are necessary to make sense of the written text because these words are categorically unfamiliar (Hoover & Tunmer, 1993). In this way, nonword decoding can be informative of word reading skills (e.g., sound-symbol knowledge, decoding etc.) and not just a measure of word familiarity and recognition. Nonword repetition tasks have been shown to reduce bias against culturally and linguistically diverse children in the domain of oral language assessment (Ortiz, 2021). Nonword reading tasks may similarly reduce bias against those with different or limited literacy experiences. Children enter kindergarten with a wide range of literacy abilities that can be attributed to factors like linguistic diversity, but also their home literacy environment, access to books and libraries, or exposure to literacy instruction in preschool or daycare (Ackerman & Barnett, 2005). Importantly, nonword reading tasks have not been found to disadvantage strong readers with advanced lexical knowledge (Castles, et al., 2018). They can account for significant unique variance in word

reading ability beyond word reading (e.g., Hogan et al., 2005). To date, no studies have considered the role of word type in validity of DA. It is possible that nonword tasks may be better suited to predict later reading ability in a DA paradigm, because DAs are designed to measure ability to learn, rather than acquired knowledge and this ability to learn can be more easily captured in a task with nonwords.

Symbol Type

Word reading assessments of sound-symbol knowledge and decoding can differ in the type of symbol used in their items (i.e., familiar letters and characters or novel symbols). Typically, SAs, which are developed and normed for a specific population, 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 English letters and sounds, and the phoneme decoding subtest (a decoding measure) evaluates their ability to read nonwords comprised of English graphemes. However, in recent years there has been increased interest in using novel symbols in place of familiar letters or characters in DAs. Using unfamiliar symbols allows researchers and clinicians to determine 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), all while minimizing the influence of previous linguistic and literacy exposure.

Measures that use novel symbols have been shown to differentiate between typical readers and those with dyslexia in adult populations (Elbro et al., 2012), and in children (Aravena et al., 2013, 2018). Additionally, DAs that use novel symbols have documented that these measures can explain unique variance in later reading ability beyond traditional measures for preliterate children (Horbach et al., 2015). Outcomes from two recent studies that examined the capacity of a nonword decoding measure administered in kindergarten to predict reading difficulty in grade 1 suggest that the measure that used

novel symbols (Gellert & Elbro, 2017a) had a superior diagnostic accuracy to a task that used familiar letters (Petersen et al., 2016). However, the 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 whether DAs that use novel symbols are comparably valid to those that use familiar letters or characters. It is possible that in DA, evaluating ability to learn SSK or decoding skills may be more easily achieved with novel symbols than with familiar letters.

Previous Research in Dynamic Assessment of Word Reading Skills

Reviews that evaluated the use of DAs reported promising findings on its utility and validity. Caffrey et al. (2008) found that DAs demonstrated greater predictive validity than SA across several domains (e.g., DAs of cognitive ability, literacy, and mathematics). A more recent review focussing exclusively on the domain of word reading skills, reported that DAs of phonological awareness and decoding demonstrate concurrent validity with SAs and predictive validity with word reading outcomes (Wood et al., 2023). In terms of format of DAs of word reading skills, no quantitative comparison has been conducted evaluating differences in associations with word reading outcomes, although findings suggest that there are no differences in classification accuracy of reading disorder for DAs that use a GP vs. TT format (Dixon et al., 2022b)). Regarding administration method, SAs of oral language skills are not affected by computer vs. in-person delivery (Alfano et al., 2022). A recent review reported that computerised DAs were used less frequently than in-person measures in research (Dixon et al., 2022b). However, the effect of administration method and its implications on the validity of DAs has not yet been considered. To our knowledge no prior reviews have qualitatively or quantitatively examined whether factors of word (real word vs. nonword) and symbol type (novel vs. familiar) are associated with stronger correlational relationships with later word reading outcomes. In summary, DAs of word reading skills vary based on several characteristics, which could have implications for the validity of these measures.

These factors should be considered to inform clinical decision-making and development of novel DA measures.

The Current Study

The current systematic review and correlational meta-analysis investigates whether the format, administration method, word, and symbol type affect DA's effectiveness. Like Caffrey et al., (2008) we will use Pearson's correlation coefficients as our effect size measure, given that these are the most observed type of effect size reported in studies investigating DA measures. We focus exclusively on DAs of word reading skills: phonological awareness, sound-symbol knowledge, and decoding – as these skills demonstrate concurrent validity with SA counterparts and predictive validity with later word reading outcomes (Wood et al., 2023). In our analyses, we stratify DAs by their format (graduated prompts vs. train/test), their administration method (in-person vs. computer vs. computerized); (also see Dixon et al., 2022a; 2022b), and include stratifications by word type (real word vs. nonword) and symbol type (familiar vs. novel). Unlike all previous reviews, we will conduct a comprehensive search of the grey literature and will include studies published in languages other than English. The outcomes of this review will inform which administration method, word and symbol types used in DAs of word reading skills are associated with the strongest correlations with word reading measures and will have both clinical and research implications. For clinicians, it is critical to understand which characteristics of DAs are associated with stronger correlational relationships with word reading outcomes, to make informed choices about which tools to use in their practice. For researchers, a quantitative examination how these factors affect validity of DAs of word reading skills can inform development of high quality, novel tools, or revisions of existing measures.

Method

The review objectives and meta-analytic approach were planned a priori and detailed in a registered protocol on the Open Science Framework. This protocol is available online at https://osf.io/bcghx/ (Wood & Molnar, 2022).

Research Questions

Do DAs of word reading skills (phonological awareness (PA) sound symbol knowledge (SSK) and decoding) demonstrate similarly strong correlations with word reading measures when stratified by:

A) Format (train-test (TT) vs. graduated prompts (GP))

B) Administration method (computer vs. in-person)

C) Word type (real word vs. nonword)

D) Symbol type (novel vs. familiar letters or characters)

Eligibility Criteria

Study inclusion criteria were determined a priori and outlined in the protocol on Open Science Framework (Wood & Molnar, 2022). All studies included in this review were:

(i) Primary research articles found in peer-reviewed journals, and unpublished grey literature found in preprint repositories and on Google Scholar. Systematic reviews, books or book chapters, case studies, commentaries, and editorials were excluded.

(ii) Studies that assessed children with a mean age between 4;0 and 10;0. Articles that included adults or children with 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 one of three word reading skills, and a static word reading measure, concurrently or longitudinally. This allowed for a comparison of

the relationships between DAs of different format and administration methods with word reading outcome measures.

(v) No limitation was placed on setting or geographical location, but only articles written in English, French, Spanish, or a different language with full text translations were included.

Search Strategy and Information Sources

The 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 Centre), using the terms “dynamic assessment” and “literacy” as well as their related keywords in titles and abstracts. Filters were not used in the search process. A complete list of search terms used in each database can be found in Table 1 and 2 of the supplemental files and online at https://osf.io/bcghx/. Equivalent terms “dynamic assessment” and “literacy” were searched in MedArxiv, EdArxiv and PsyArxiv preprint repositories. The first author and a research assistant started forward searching of included articles on Google Scholar upon the completion of the database and preprint repository search. Sources that cited each included article were located using the “cited by” function. To check whether any relevant articles were potentially missed during the database, preprint, and Google Scholar search, the first author and a research assistant reviewed the reference lists of the included articles and compared them with the list of included articles. Lastly, appeals for unpublished work were made on lab and researcher social media platforms and sent out twice to lists and labs across Canada, the United States and Europe that reported conducting research in field of literacy.

Data Collection

Data collection and extraction was managed in Covidence, a web-based software that facilitates completion of reviews (Covidence, 2023). A team of ten research assistants (RAs), trained by the first author, assisted in article screening and extraction. At the title/abstract stage, two independent team

members voted to include or exclude based on relevance. In the full text stage, two reviewers voted to include or exclude articles based on whether they met pre-defined characteristics. The same team of ten RAs extracted the data from included studies using a custom template in Covidence. This template is available at https://osf.io/bcghx/. In all stages, the first author resolved conflicts.

Data Items

The following information were extracted from the included papers.

General Information

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

Participants

The number of participants at the end of the study included in analyses (taking attrition into account for longitudinal studies), the percentage of males, the language(s) spoken by participants, as well as the mean age and grade level of the children at the outset of the study were noted.

Measures

Dynamic Assessment(s) In this review, DA is defined as an assessment that provides teaching, training, feedback on performance, or prompting during testing. The research team reported the word reading skills evaluated (either phonological awareness, sound-symbol knowledge, or decoding, or multiple), as well as the type of task used to assess the skill (e.g., phonological awareness can be assessed by syllable or phoneme blending). If multiple tasks were used to evaluate a skill, coders would list all tasks utilized. Reviewers also noted the format of the DA (i.e., graduated prompts (GP) or train/test (TT), which administration method was employed (i.e., in person or computer) and whether real or nonwords, and novel or familiar symbols were used.

Word Reading Measures (WRM)

For the purposes of this review, WRMs are assessments that measure ability to read single words using a standardized 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. Coders noted the name of the WRM, and the subtest used (e.g., the Woodcock Reading Mastery Tests –III, Word Identification subtest), which word reading skill(s) was evaluated and whether this task used words or nonwords (e.g., single word reading accuracy).

Effect Sizes

Pearson's correlation coefficients representing the relationship between DAs and WRMs were extracted. Coders initially extracted all correlation coefficients between a DA and an WRM listed in a study (e.g., a DA that used multiple PA tasks to assess PA skills and multiple measures to evaluate word reading ability). Following review of extracted data points, the first, second and last author developed a set of decision rules for selecting a single effect size from each study as to not violate the assumption of independence in the meta-analysis. This decision-making process was based on which measure was most frequently observed among the included studies; (i.e., every included study utilized word reading accuracy as an WRM, while word reading fluency was scarcely used). In cases where tasks were observed with equal frequency, the choice was informed by theory. For example, research suggests that phoneme level tasks demonstrate stronger predictive validity than syllable or onset-rime level tasks ones, and so a phoneme deletion task would be preferred over syllable deletion task. Effect sizes representing the relationship between DAs and WRMs are presented in Table 3 in the supplemental material. The excel table with extracted data and the R script can be found at https://osf.io/bcghx//.

Quality Appraisal Assessment

Each included study was evaluated independently by two RAs using an adapted and amalgamated version of two quality assessment tools for (i) cross sectional design and (ii) diagnostic accuracy studies from the Johanna Briggs Institute (Moola et al., 2020). Studies were assessed on the five following areas: (i) participant selection, (ii) index assessments (DAs) (iii) reference assessments (WRMs), (iv) flow and timing of the study, (v) statistical analysis.

First, coders rated whether the age, sex, and demographic characteristics of the participants were adequately described. The rating for the DA domain was informed by whether the tool was explained with adequate detail regarding the skills assessed, the format, the type of prompting and scoring used, and the method of administration. Coders also noted whether the word reading skill(s) employed were developmentally appropriate for the sample population. The developmental appropriateness of assessment tools for evaluating word reading skills were also evaluated when rating the standards of reference assessments (WRMs). Additionally, coders rated whether the studies indicated the psychometric properties of the reference measures. To evaluate flow and timing, coders evaluated whether the analyses included all participants and if not, whether the author(s) provided adequate 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 the review objectives. The first author reviewed all ratings and resolved any conflicts, and the quality of each study was ranked on the following spectrum: 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. Refer to Table 4 in the Supplemental Material for quality appraisal questions and ratings.

Analyses

A random effects meta-analysis was conducted to account for between-study variance. A single coefficient from each study was selected to ensure compliance with the assumption of independence. Coefficients were transformed into Z scores using Fisher Z transformation with the 'metacor' package in R studio (Laliberté, 2019; R Core Team, 2021). A weighted average of these scores was calculated, then transformed back to Pearson's correlation coefficients for interpretation. Heterogeneity statistics of Q, I2 and Tau2 were calculated and reported. The Sidik Jonkman estimator was used to calculate Tau2. A Baujat plot (Figure 1 in supplemental material) was generated to determine which studies contributed most to heterogeneity (Baujat et al., 2002). Significant between study variance was anticipated given differences in study design, participant factors, DA and WRM characteristics. To examine this heterogeneity, subgroup analyses by DA format (graduated prompts vs. train/test), administration method (computer vs in-person) word (real vs. nonword) and symbol type (novel vs. familiar letters/characters) were planned a priori. A funnel plot (Figure 2 in supplemental material) was generated, and Egger's regression test was conducted to examine risk of publication bias (Egger et al., 1997).

Results

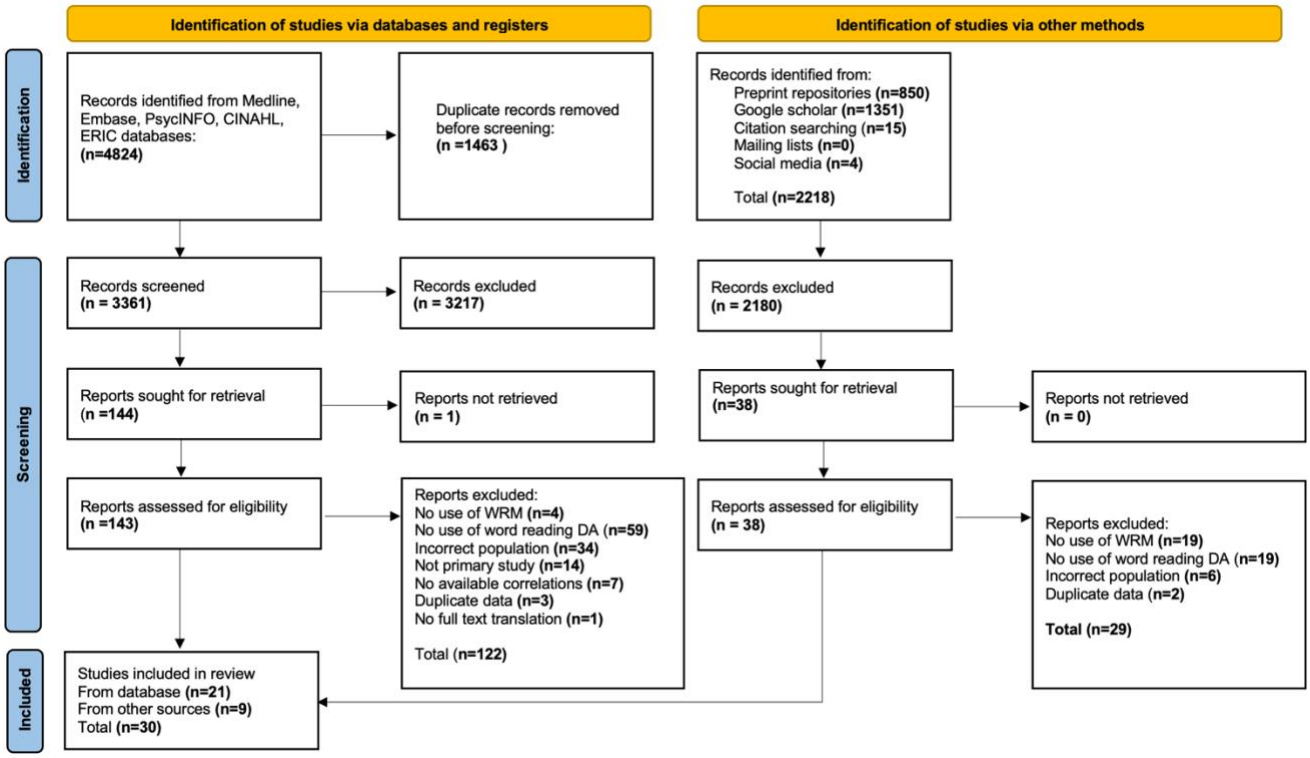
Study Selection

The database search yielded 4824 articles of which 21 were included. Three preprint repositories were searched of which 1 was included. The 22 included articles were then subjected to forward searching using the "cited by" function in Google Scholar which led to identification of an additional 7 studies. The reference lists of these 29 articles were reviewed to determine if there were relevant articles that had been missed. One additional study was identified. Callouts were made for unpublished studies or data to mailing lists, via post to social media and by directly contacting labs conducting literacy related research across Canada, the United State and Europe, but no relevant articles were identified via this process. In

summary, 30 articles including a total of 32 studies met the criteria for inclusion. The study identification process, including reasons for study exclusion such as incorrect population type (e.g.,den Ouden, 2019) is outlined in the PRISMA diagram below (Page et al., 2021).

Figure 1.

Preferred Reporting Items for Systematic Review and Meta-Analyses Flowchart



Note. N= Number of participants, DA= Dynamic Assessment, WRM = Word Reading Measure

Study Characteristics

Participant Characteristics

A total of 6225 participants were included across 32 studies from 30 articles. The overall mean age of participants was 5 years 8 months, and the overall average % of males was 49.27%. Table 1 below provides additional details regarding the mean age and % of males across subgroups stratified by DA characteristics.

Table 1.

Participant Characteristics

	N	Mean start age*	Mean % males**
Total (min-max)	6225 (9-1988)	5;8 (4;1 – 9;9)	49.27% (41%-66%)
Format			
Graduated Prompts	4017	5;2	50.75%
Train-test	1485	6;10	49.73%
Other	723	NR***	44.5%
Administration Method			
Computer	1434	7;6	46.17%
In-Person	4791	5;4	51.78%
Word type			
Real word	3988	5;5	48.85%
Nonword	1375	6;4	53.15%
N/A****	862	-	-
Symbol type			
Familiar	1908	6;2	48.45%
Novel	1553	6;9	50.96%
N/A*****	2784	-	-

Note. N= Number of participants

* 23/32 studies reported mean age of participants at study onset

**26/32 studies reported mean % males in their participant sample

***NR=Mean start age not reported

**** Tasks that did not use words (sound-symbol knowledge tasks)

***** Tasks that did not use print or symbols (phonological awareness tasks)

Study Location and Language

Most studies were conducted in the United States (n=17), followed by The Netherlands (n=2), Germany (n=2), Denmark (n=2), England (n=2), China (n=2), and Hong Kong, Finland, Belgium, Spain, and Singapore (n=1 each). The most common language profile across studies was monolingual English speakers (n=15), followed by bilingual English/Other speakers (n=6), and monolingual Danish (n=4), German (n=2), Mandarin (n=2), Finnish (n=1), Dutch (n=1) and Spanish (n=1) speakers. Refer to Table 2 for additional information about characteristics of participants in included studies.

Dynamic Assessments

Of the 32 included studies, 11 examined a DA of PA, 5 a DA of SSK and 16 a DA of decoding. Eleven studies were administered either via computer, either by a person or in a computerized program, while 21 were conducted in-person. Seventeen studies employed a GP format, 14 used a TT approach and 1 used a game-based method that was neither GP nor TT. Most studies used explicit verbal feedback (n=29) while only 3 used implicit feedback in the context of a game. Given their nature, DAs of SSK used neither words nor nonwords, only symbols and sounds or syllables. However, of the 27 studies that examined a DA of PA or decoding, 15 used real words and the remaining 12 used nonwords. Studies used either novel symbols or familiar letters and characters in their DAs of SSK and decoding. Given their auditory nature, PA tasks used neither symbols nor letters. Of the 21 DAs of SSK and decoding tasks, 12 used novel symbols, while 9 used familiar letters or characters. For DA details, refer to Table 2.

Word Reading Measures

WRMs used in included studies were characterized as either norm-referenced or researcher developed. Most studies (n=27) used norm-referenced measures, while fewer (n=5) used a researcher developed tool. The norm-referenced measures used included versions of the Woodcock Reading Mastery Test (either WRMT-R, WRMT-RNU, WRMT-III) (n=9), the Test Of Word Reading Efficiency (n=3), the Woodcock Johnson-III (n=2), the WRAT(n=2), the Woodcock Muñoz Language Survey-Revised (n=2), the Salzburger Lese und Rechtschreib Test-III (n=2), the One Minute Test (n=2), and the Test de Análisis de la Lecto-Escritura, and the British Ability Scales-2, Single Word Reading Test, Lukilasse, 3DM, San Diego Quick Assessment, (n=1 each). Most WRMs evaluated word reading accuracy (n=25), though several assessed both accuracy and speed (n=7). Finally, the majority of WRMs used real word reading tasks (n=28), apart from two studies that used a combination of word and nonword reading subtests, and two that used exclusively nonword reading tasks. Refer to Table 2 details regarding WRMs.

Table 2.

Country, Number of Participants, Mean Age, Grade, % Males, Language Status, Reading Status, Study Design, Type and Characteristics of DAs, SAs and WR Outcome Measures of Included Studies.

Study	Country	N	Mean age T1 (Year; month)	Grade T1	Sex % Males	Language Status	Reading Status	Study Timing	DA Skill(s)	DA Feedback	DA Format	DA Administration	WRM Skill/Subtest/Test
Aravena, S., Snellings, P., Tijms, J., & van der Molen, M. W. (2013)	The Netherlands	64 TD 62 Dx. 42 Dx. *	9;9 TD 9;9 Dx.	Not stated	48.4% TD 56.4%Dx.	ML Danish	TD & Dx	Cross-sectional	Decoding rate real words of novel symbols	Implicit	TT	Computerized	Word reading accuracy and speed/ One Minute Test
Aravena, S., Tijms, J. Snellings, P. & van der Molen, M.W. (2018)	The Netherlands	46 TD 72 Dx. 71 Dx. *	9;3 TD 9;2 Dx.	Not stated	47.8%TD 58.2%Dx.	ML Danish	TD & Dx	Cross-sectional	Decoding real words of novel symbols	Implicit	TT	Computerized	Word reading accuracy / 3DM test
Barker, R.M. & Saunders, K.J. (2020)	United States	27	4;11	Pre-K	51.8%	ML English	TD	Cross-sectional	SSK learn novel symbol-syllable	Explicit verbal	TT	Person via computer	Word reading / Word identification/ WRMT-III
Caffrey, E. (2006) Sample 1	United States	25	Not stated	K	52%	ML English	TD	Longitudinal	Decoding nonwords	Explicit verbal	GP	Person	Word reading / WRAT
Caffrey, E. (2006) Sample 2	United States	95	Not stated	G1	57%	ML English	TD	Longitudinal	Decoding nonwords	Explicit verbal	GP	Person	Word reading / WRAT
Cho, E., Compton, D., Fuchs, D., Fuchs, L.S. & Bouton, B. (2014)	United States	134	Not stated	G1	50%	ML English	AR	Cross-Sectional	Decoding nonwords	Explicit verbal	GP	Person	Nonword reading/ Word attack / WRMT-RNU
Cho, E. & Compton, D.L. (2015)	United States	112	6;8	G1	57.1%	ML English	TD	Cross-sectional	Decoding words of novel symbols	Explicit verbal	GP	Person	Word reading / Word identification/ WRMT-RNU
Cho, E., Compton, D.L., Gilbert, J.K., Steacy, L.M., Collins, A.A. & Lindström, E.R. (2017)	United States	54 TD 51 AR 105 total	6;7 TD 6;10 AR 6;8	G1	55%	ML English	TD & AR	Longitudinal	Decoding words of novel symbols	Explicit verbal	GP	Person	Word reading / Word identification/ WRMT-RNU
Chow, B. W-Y. (2014)	Hong Kong	121	7;0	G2	54.5%	BL Cantonese & English	TD	Cross-Sectional	SSK learn novel symbol syllables	Explicit verbal	TT	Person	Word reading/ Research developed tool
Clayton, F. et al., (2018)	England	97	9;2	Year 4 Year 5	50.5%	ML English	TD	Cross-Sectional	SSK learn novel symbol sounds	Explicit verbal	TT	Person via computer	Word read / SWRT

Compton, D. L., Fuchs, D., Fuchs, L. S., Bouton, B., Gilbert, J. K., Barquero, L. A., Cho, E., Crouch, R.C. (2010)	United States	140 TD 214 AR 354 total	Not stated	G1	53.45%	ML English	TD & AR	Longitudinal	Decoding nonwords	Explicit verbal	GP	Person	Composite: Untimed Word Identification, Word attack / Reading Comprehension/ WRMT-RNU, Timed Sight Word reading, Timed decoding / TOWRE
Coventry, W.L., Byrne, B., Olson, R.K., Corley, R. & Samuelsson, S. (2011)	Australia United States Norway Sweden	1988	4;11	Pre-K	48%MZ 53% DZ	ML English Norwegian Swedish BL not specified	TD	Cross-Sectional	PA phoneme identification real words	Explicit verbal	GP	Person	Composite: Sight word efficiency & Phonemic Decoding efficiency / TOWRE
Cunningham, A. & Carroll, J. (2011)	England	45	5;0 K 5.2 young G1 5;11 old G1	K G1	Not stated	ML English	TD	Longitudinal	PA phoneme segmentation nonwords	Not specified	GP	Person	Word reading / BAS-2
Edwards, A. (2020)	United States	312	Not stated	G1	Not stated	ML English	TD	Cross-sectional	Decoding nonwords	Explicit verbal	GP	Person	Word reading/ Word identification /WJ-III
Gan, Y., Zhang, J., Kharabi-Yamato, L., Su, Y., Zhang J., Jiang, Y., Hui, Y., & Li, H. (2022)	China	135	6; 8 males 7;2 females	G1	54%	ML Mandarin	TD	Longitudinal	Decoding compound ideographic characters	Explicit verbal	GP	Person	Character recognition / Research developed tool
Gellert, A.S. & Elbro, C. (2017a)	Denmark	89 TD 82 AR 171 total	6;11	K	Not stated	ML Danish-110 BL not specified 61	TD & AR	Longitudinal	Decoding nonwords of novel symbols	Explicit verbal	TT	Person	Word reading / researcher developed tool
Gellert, A.S. & Elbro, C. (2017b)	Denmark	84 TD 76 AR 160 total	6;4	K	48%	ML Danish -101 BL not specified 59	TD & AR	Longitudinal	PA phoneme identification real words	Explicit verbal	GP	Person	Word reading / researcher developed tool
Gillam, S.L., Fargo, J. Foley, B. & Olszewski, A. (2011)	United States	64	6;5 young 8;0 old	G1 G2 G3	47%	ML English	TD	Cross-Sectional	PA nonverbal phoneme deletion real words	Explicit verbal	GP	Person	Nonword decoding / Word attack / WRMT-R

Hautala, J., Heikkilä, R., Nieminen, L., Rantanen, V. Latvala, J-M. & Richardson, U. (2020)	Finland	723	Not stated	G1 G2 G3 G4	44.5%	ML Finnish	TD & AR	Cross-Sectional	Decoding words	Explicit verbal and auditory	Game-based	Computerized	Word reading / Graded word-level reading fluency test Lukilasse 2
Horbach, J., Scharke, W., Cröll, J., Heim, S. & Günther, T. (2015)	Germany	243	6;2	K	56.0%	ML German	TD & AR	Longitudinal	SSK learn novel symbol syllables	Train: Explicit verbal Test: None	TT	Person via computer	Word reading accuracy and speed / SLRT-II
Horbach, J. Weber, K. Opolony, F. Scharke, W. Radach, R. Heim, S. & Günther, T. (2018)	Germany	17	5;0	Pre-K	41%	ML German-6 BL not specified 11	TD	Longitudinal	Decoding nonwords of novel symbols	Train: Explicit verbal Test: None	TT	Person via computer	Word reading accuracy and speed / SLRT-II
Law, J.M., De Vos, A., Vanderauwera, J., Wouters, J. Ghesquiere, P. & Vandermosten, M. (2018)	Belgium	64 TD 20 Dx. 84 total	8;3	G3	Not stated	ML Dutch	TD & Dx.	Cross-Sectional	Decoding words of novel symbols	Implicit	TT	Computerized	Word reading accuracy and speed/ One Minute Test
Liu, C., Hoa Chung, K.K., Wang, L.C. & Liu, D. (2021)	China	203	5;0	K Year 2	51.23%	ML Mandarin	TD	Cross-Sectional	SSK learn novel symbol sounds	Train: Explicit verbal Test: None	TT	Person	Single and double character word reading/ Researcher developed tool
Loreti, B. (2015)	United States	10	4;10	Pre-K	60%	BL Spanish English	TD	Cross-Sectional	PA syllable & onset rime discrimination nonwords	Explicit verbal and visual	TT	Computerized	Word reading/ Letter and Word Identification / WMLS-R
Osa Fuentes, P.M. (2003)	Spain	164	5;6	Infantil (Pre-K)	45.7%	ML Spanish	TD	Longitudinal	PA syllable segmentation and matching words	Explicit verbal	GP	Person	Word reading accuracy / TALE
Petersen, D. B., & Gillam, R. B. (2015)	United States	63	5;4	K	53.9%	BL Spanish & English	AR	Longitudinal	Decoding nonwords	Explicit verbal	Pretest TT	Person	Word identification / WRMT-RNU
Petersen, D.B, Allen, M.M. & Spencer, T.D. (2016)	United States	280	Not stated	K	Not stated	ML English BL Spanish & English	TD & AR	Longitudinal	Decoding nonwords	Explicit verbal	Pretest TT	Person	Sight Word Efficiency / TOWRE
Sittner Bridges, M., & Catts, H.W. (2011) Sample 1	United States	90	Not stated	K	Not stated	English	TD	Longitudinal	PA syllable and phoneme deletion words	Explicit verbal	GP	Person	Word reading/ Word

													identification / WRMT-RNU
Sittner Bridges, M., & Catts, H.W. (2011) Sample 2	United States	96	Not stated	K	Not stated	English	TD & AR	Longitudinal	PA syllable and phoneme deletion words	Explicit verbal	GP	Person	Word reading/ Word identification / WRMT-RNU
Spector, J. (1992)	United States	38	5;11	K	Not stated	ML English	AR	Longitudinal	PA phoneme segmentation words	Explicit verbal	GP	Person	Word recognition / San Diego Quick Assessment List
Wyman Chin, K.R. (2018)	United States	9	Not stated	Not stated	66%	BL Spanish & English	TD	Cross-Sectional	PA syllable discrimination nonwords	Explicit verbal	TT	Computerized	Word reading/ Letter and Word Identification / WMLS-R
Yap, D.F-F. (2018)	Singapore	99	5;3	K	43.4%	BL English & Chinese (75%) -Malay (12%) -Other (13%)	TD	Longitudinal	PA phoneme segmentation words	Explicit verbal	GP	Person	Word reading / Letter word identification / WJ-III

Note. TD= typically developing, AR= at-risk, Dx. = diagnosed with reading difficulty, G=grade, K = kindergarten ML = monolingual, BL=bilingual, N= number of participants per study, DA=dynamic assessment, SA= static assessment, PA = phonological awareness, SSK = sound-symbol knowledge, WRM = word reading measure, BAS=British Ability Scales, SLRT-III=Salzburger Lese und Rechtschreibtest, TALE = Test de Análisis de la Lecto-escritura, TOWRE= Test of word reading efficiency, WJ-III = Woodcock Johnson-III, WRAT = Wide Range Achievement Test, WRMT(R/NU) =Woodcock Reading Mastery Tests-Revised, Normative Update, Woodcock-Muñoz Language Survey-Revised

* Provided separate correlation coefficients for 42 Dx. participants, this effect size used in analysis

** Provided separate correlation coefficients for 71 Dx. participants, this effect size used in analysis

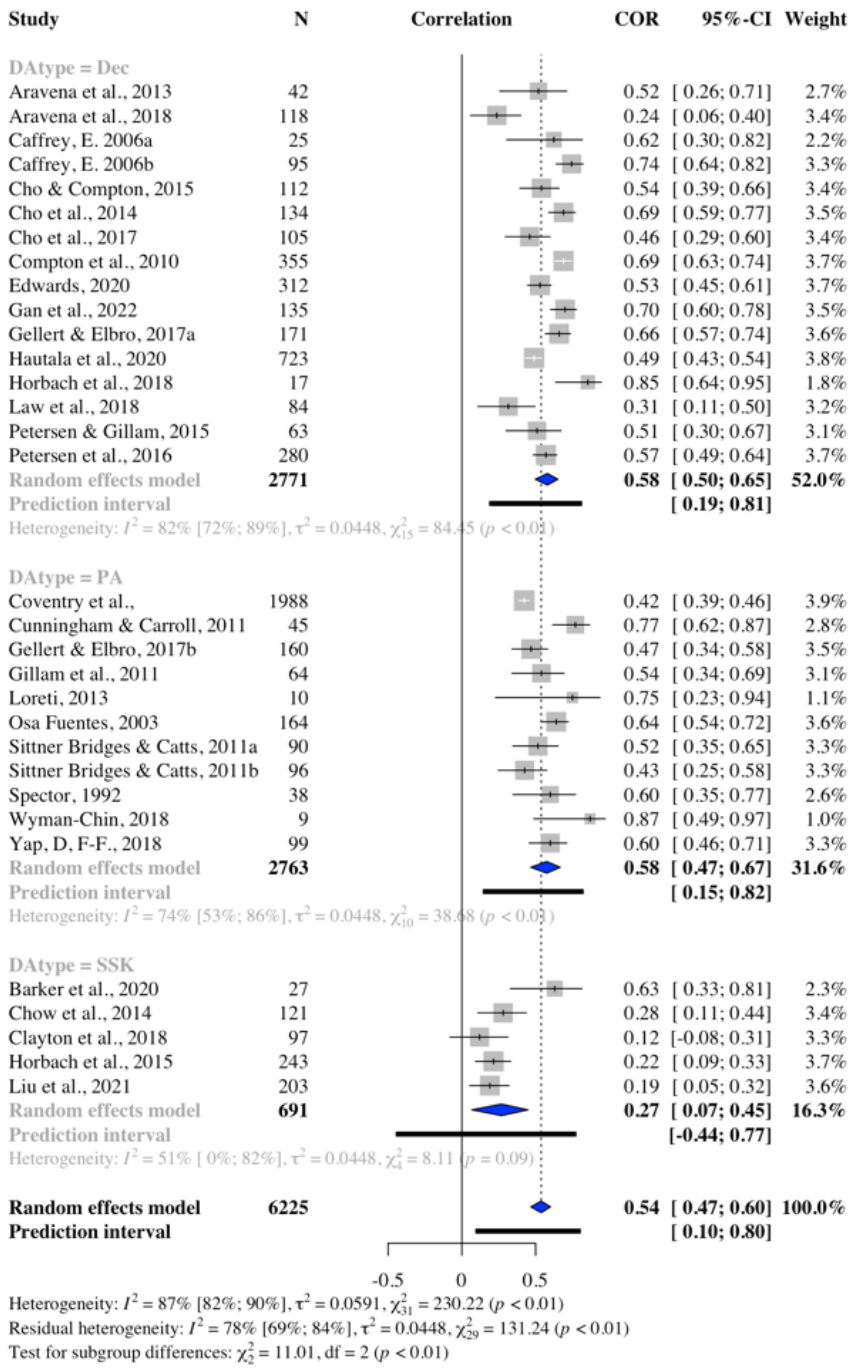
*** Data used from 2010 thesis

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

Thirty-two studies from 30 articles studies reported correlations between a DAs of word reading skills (phonological awareness, sound-symbol knowledge, or decoding) WRMs (See Table 3 in supplemental material). The effect sizes from these 32 studies were included in the correlational random effects meta-analysis examining the relationship between DAs word reading skills and WRMs, and results are displayed in Figure 2. The overall mean effect size is large (r=0.54, 95%CI = [0.47-0.60] suggesting that DAs of word reading skills are strongly correlated with WRMs. The prediction interval ranged from g=0.10-0.90 suggesting that future relevant studies would be likely to find a positive correlation. As expected, significant heterogeneity was detected (Q=230.22, p<0.01) indicating that a substantial amount of heterogeneity can be attributed to true between study variance rather than sampling error. The effect sizes are presented in subgroups according to DA type. Findings of a previous review examining validity of DAs by word reading skill type (Wood et al., 2023) are replicated here, with DAs of PA and decoding demonstrating narrower positive prediction intervals and significantly stronger correlations with WRMs than DAs of SSK (Q=11.01, df=2, p<0.01). As in previous analyses, even after subgroup analysis by DA type, significant residual heterogeneity was detected (Q=131.24, p<0.01). Additional subgroup analyses by DA format, administration method word and symbol type, were planned a priori and were conducted to further examine this heterogeneity.

Figure 2.

Forest plot of random effects meta-analysis examining the relationship between dynamic assessments word reading skills and word reading measures.



Note. Study names, sample size =N, effect sizes =COR, and 95% confidence intervals =CI (95%) are reported. The grey box associated with each study represents the weight allocated to each effect size, while the horizontal line that extends from either side of the box is a measure of the confidence interval (95%). The solid vertical line is the line of no effect while the dashed vertical line represents the significant overall mean effect size. The blue diamonds are an indication of the overall confidence interval, and the black bar represents the prediction interval. Figure drawn in R using `metacor` package (R Core Team, 2021; Laliberté, 2019).

Subgroup Analyses

Mixed effects models were used to examine whether there were significant differences in mean effect sizes for DAs based on their A) administration method, B) format C) word and D) symbol type. Results of these subgroups analyses are reported in Table 3 below.

Table 3.

Result of subgroup analyses

A) Administration method

	Included Studies	g	95%CI	I2	95%CI	p value heterogeneity	Prediction interval	p value subgroup
Admin method								0.21
Person	21	0.57	0.48-0.64	0.87	0.82-0.91	<0.01	0.12-0.80	
Computer	11	0.47	0.32-0.59	0.82	0.68-0.89	<0.01	-0.07-0.80	

B) Format

	Included Studies	g	95%CI	I2	95%CI	p value heterogeneity	Prediction interval	p value subgroup
Format								0.14
Graduated Prompts	17	0.59	0.51-0.67	0.85	0.78-0.90	<0.01	0.16-0.84	
Train/Test	14	0.46	0.33-0.56	0.86	0.79-0.91	<0.01	-0.05-0.77	
Other	1	0.49	0.43-0.54	-	-	-	-	

C) Word type

	Included Studies	g	95%CI	I2	95%CI	p value heterogeneity	Prediction interval	p value subgroup
Word type								<0.01
Words	15	0.50	0.43-0.57	0.73	0.55-0.84	<0.01	0.18-0.73	
Nonwords	12	0.67	0.60-0.73	0.65	0.35-0.81	<0.01	0.39-0.83	

D) Symbol type

	Included Studies	g	95%CI	I2	95%CI	p value heterogeneity	Prediction interval	p value subgroup
Symbol type								0.01
Familiar	9	0.62	0.51-0.71	0.82	0.67-0.90	<0.01	0.15-0.87	
Novel	12	0.42	0.29-0.53	0.85	0.76-0.91	<0.01	-0.10-0.76	

There is no significant difference in the strength of correlation between DAs of word reading skills and WRMs based on the DA format (p=0.14) or administration method (p=0.21). However, mean

effect sizes for DAs that used a TT format and those that were conducted via computer were moderate (r=.46, r=0.47 respectively) while effect sizes for DAs that used a GP format and that were conducted in-person were large (r=.59, r=.57 respectively). Furthermore, the prediction intervals for DAs that used a GP format and that were conducted in-person, did not cross 0, suggesting that future relevant studies may be more likely to document positive correlations for DAs with these characteristics. Significant differences were found in terms of strength of correlation between DAs and WRMs based on the type of word (p=<0.01) and symbol (p=0.01) used. Results suggest that DAs that use nonwords (r=.57) and those that use familiar symbols (r=.62) are more strongly correlated with WRMs than those that use real words (r=.50) or those that use novel symbols (r=.42).

Risk of Publication Bias

A funnel plot was generated to subjectively examine risk of publication bias in the meta-analysis of the relationship between DAs word reading skills and WRMs. This plot is presented in Figure 2 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, 2013; Wyman Chin, 2018). This suggests that there is a possibility that studies with negative outcomes were not completed, published, or submitted into the grey literature (Lee & Hotopf, 2012). To objectively investigate risk of publication bias, Egger's test was calculated. Despite the apparent visual asymmetry, this test was not found to be significant for presence of plot asymmetry (Intercept = 1.4922, 95%CI= [-0.14-3.12], p=.08). It is plausible that the word reading assessment constructs are reliably correlated with word reading outcome measures in young children, and therefore negative correlations between the two would not be anticipated. Ultimately, there is minimal risk of publication bias in this analysis.

Discussion

This review examined whether characteristics of dynamic assessments (DAs) of word reading skills (phonological awareness, sound-symbol knowledge, and decoding) affected the strength of the correlational relationship between the DA and performance on a word reading measure (WRM). Thirty-two studies from 30 articles met inclusion criteria of evaluating children between the ages of 4 and 10 and reporting a correlation coefficient between a DA of a word reading skill and a WRM. Results of the overall meta-analysis were consistent with previous findings and suggested that DAs of word reading are strongly correlated with WRMs (Wood et al., 2023).

The subgroup analyses evaluating DA *format* found no significant differences between the graduated prompts (GP) or train-test (TT) approaches. However, mean effect sizes were larger, and prediction intervals were narrower for DAs that employed a GP approach versus those that used the TT format. These results are consistent with findings from a previous review (Caffrey et al., 2008) which found that DAs that used contingent feedback demonstrated stronger predictive validity than those that used non-contingent feedback. GP DAs are typically very scripted and use highly 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 in the learning phase of their DA). However, the feedback in the training and teaching phase of TT DAs was characterized by greater variability. For example, in Petersen & 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 in may have contributed to the weaker relationship between TT DAs and WRMs.

A secondary subgroup 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 via computer. These findings are consistent with a previous review which reported no significant differences between in-person and computer administration methods for static assessments (SAs) across domains (Alfano et al., 2022). However, the mean effect size for DAs conducted in-person was larger and had a narrower prediction interval that did not cross zero. The weaker mean effect size and wider prediction interval associated with computer administration may be a result of two factors. First, all WRMs were conducted in-person which may have impacted the strength of the relationship between the two measures. It is promising that despite this difference in administration method, computer 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 computer and virtual use of SAs (e.g., Hodge et al., 2019; Wood et al., 2021). However, results from these studies and this review would indicate that much like SAs, computer-based or virtual DAs are a valid alternative to in-person administration.

Finally, two additional subgroup analyses examined the impact of *word* and *symbol type* in DAs of word reading skills. Results indicate that DAs that use nonwords and those that use familiar letters or characters demonstrate significantly greater mean effect sizes than those that use real words and those that employ novel symbols. These results differ from previous findings, 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 guess or recognize words in testing (Hoover

& Tunmer, 1993). The trend of unfamiliarity of test items leading to increased capacity to evaluate ability to learn was not reflected in the subgroup analysis by symbol type given that DAs that used familiar letters or characters were associated with stronger mean effect sizes than those that used novel symbols. We hypothesize 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.

Limitations

Correlation coefficients were selected as the measure of effect size because they were the most reported statistical analysis across studies. While this allowed for inclusion of a higher number of studies, it also means that only correlational inferences can be made about the results reported in this review. 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. While it has been suggested that some ‘paired associate learning’ (PAL) should be considered as DA measures (Dixon et al., 2022), our team elected not to include PA tasks, because most of these tasks are not dynamic in nature. Additionally, despite examination of four moderators via subgroup analysis, residual heterogeneity within groups was still significant. This suggests that other factors that were not examined in this review, may be contributing to the overall strength of relationship between DAs of word reading skills and word reading measures.

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 support the use of both graduated prompts and train/test type formats of DA, either conducted in-person or via computer. This is particularly relevant post-pandemic, as many professionals continue to evaluate children in a virtual context. Findings also suggest that clinicians may wish to favour DAs of word reading that use nonwords and familiar letters, over those that use real words or novel symbols, as these characteristics were associated with significantly stronger correlational relationships with word reading outcomes.

Future Research

Results of this study can inform development of novel DAs of word reading skills, or revisions of existing tools. It will be important for researchers to directly compare DAs with differing characteristics, using research designs and statistical analyses that permit a better understanding of the causal role these factors play. 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. Finally, studies should 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.

Author Notes

The authors do not declare any conflicts of interest at the time of publication.

Acknowledgements

This systematic review and meta-analysis is funded 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 and an Ontario Graduate Scholarship from the Ministry of Colleges and Universities, awarded to EW, by a University of Toronto Excellence Award, awarded to KB and by a Natural Sciences and Engineering Research Council of Canada grant awarded to MM (RGPIN-2019-06523).

Data Availability Statement

Additional supplemental material (full list of search terms, excel files with extracted correlation coefficients and R codes for correlational meta-analyses and subgroup analyses) are available on the Open Science Framework at https://osf.io/bcghx.

References

*****References with asterisks were those included in the systematic review and meta-analysis

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.

Annie E. Casey Foundation. (2014). *Early reading proficiency in the United States.* https://www.aecf.org/resources/early-reading-proficiency-in-the-united-states

**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

** Bridges, M. S. (2009). *The use of a dynamic screening of phonological awareness to predict reading risk for kindergarten students* (Doctoral dissertation, University of Kansas).

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. (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

Campione, J. C., Brown, A. L., Ferrara, R. A., & Bryant, N. R. (1984). The zone of proximal development: Implications for individual differences and learning. *New Directions for Child and Adolescent Development, 1984*(23), 77-91. https://doi.org/10.1002/cd.23219842308

Campione, J. C., & Brown, A. L. (1987). Linking dynamic assessment with school achievement. In C.S. Lidz (Ed.) *Dynamic assessment: An interactional approach to evaluating learning potential*. (pp.82-115). New York: Guildford Press.

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., Adlof, S. M., Hogan, T. P., & Weismer, S. E. (2005). Are specific language impairment and dyslexia distinct disorders*? Journal of Speech Language and Hearing Research*, *48*(6), 1378-1396.https://doi.org/ 10.1044/1092-4388(2005/096)

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., 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. (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., 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

**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 systematic review software (2023), Veritas Health Innovation, Melbourne, Australia. Available at www.covidence.org.

** Cunningham, A. J. (2010). *Age and schooling effects on the development of word reading and related skills* (Doctoral dissertation, University of Warwick).

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

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

den Ouden, M., Keuning, J., & Eggen, T. (2019). Fine-grained assessment of children's text comprehension skills. *Frontiers in Psychology.* doi: 10.3389/fpsyg.2019.01313

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

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

Elbro, C., Daugaard, H. T., & Gellert, A. S. (2012). Dyslexia in a second language? —a dynamic test of reading acquisition may provide a fair answer. *Annals of dyslexia, 62*, 172-185. https://doi-org. /10.1007/s11881-012-0071-7

Fuchs, L. S., Fuchs, D., & Compton, D. L. (2004). Monitoring early reading development in first grade: Word identification fluency versus nonsense word fluency. *Exceptional Children*, *71*(1), 7-21. https://doi-org./10.1177/001440290407100101

**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

** 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.

**Hautala, J., Heikkilä, R., Nieminen, L., Rantanen, V., Latvala, J. M., & Richardson, U. (2020). Identification of reading difficulties by a digital game-based assessment technology. *Journal of Educational Computing Research*, *58*(5), 1003-1028. https://doi-org./10.1177/0735633120905309

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

Hjetland, H. N., Brinchmann, E. I., Scherer, R., Hulme, C., & Melby-Lervåg, M. (2020). Preschool pathways to reading comprehension: A systematic meta-analytic review. *Educational Research Review*, *30*, 100323. https://doi.org/10.1016/j.edurev.2020.100323

Hodge, M. A., Sutherland, R., Jeng, K., Bale, G., Batta, P., Cambridge, A., Detheridge, J., Drevensek, S., Edwards, L., Everett, M., Ganesalingam, C., Geier, P., Kass, C., Mathieson, S., McCabe, M., Micallef, K., Molomby, K., Pfeiffer, S., Pope, S., Tait, F., … Silove, N. (2019). Literacy Assessment Via Telepractice Is Comparable to Face-to-Face Assessment in Children with Reading Difficulties Living in Rural Australia. *Telemedicine journal and e-health : the official journal of the American Telemedicine Association*, *25*(4), 279–287. https://doi-org.myaccess.library.utoronto.ca/10.1089/tmj.2018.0049

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

**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

Laliberté, E. (2019). metacor: Meta-Analysis of Correlation Coefficients in R. R package version 1.0-2.1. https://CRAN.R-project.org/package=metacor

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).

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

Nelson, N. W., & 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

Ontario Human Rights Commission. (2022). *Right to Read Inquiry Report.* https://www.ohrc.on.ca/en/right-to-read-inquiry-report

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

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., 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

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

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. (1998). Early identification of children at risk for reading disabilities: Phonological awareness and some other promising predictors. In B.K. Shapiro, P.J. Accardo & A.J. Capute (Eds.), *Specific reading disability: A view of the spectrum* (pp. 75-119). York Press.

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.

Sewell, T. E. (1987). Dynamic assessment as a non-discriminatory procedure. In C.S. Lidz (Ed.) *Dynamic assessment: An interactional approach to evaluating learning potential* (pp.426-443). 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

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

Sternberg, R. J., & Grigorenko, E. L. (2002). *Dynamic testing: The nature and measurement of learning potential*. Cambridge University Press.

The Conference Board of Canada. (2014). *Students with inadequate reading skills.* https://www.conferenceboard.ca/hcp/provincial/education/stu-lowread.aspx

Tohidast, S. A., Mansuri, B., Bagheri, R., & Azimi, H. (2020). Provision of speech-language pathology services for the treatment of speech and language disorders in children during the COVID-19 pandemic: Problems, concerns, and solutions. *International Journal of Pediatric Otorhinolaryngology*, *138*, 110262. https://doi.org/ 10.1016/j.ijporl.2020.110262

United Nations Educational, Scientific and Cultural Organization. (2013). *Literacy. https://en.unesco.org/themes/literacy*

United Nations Education, Scientific and Cultural Organization. (2021). *Supporting learning recovery one year into COVID-19: The Global Education Coalition in action.* https://unesdoc.unesco.org/ark:/48223/pf0000376061

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.

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).* https://doi-org. /10.1177/2050312121105051

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. https://www.osf.io/bcghx

Wood, E., Biggs, K., & Molnar, M. (2023). Concurrent and predictive validity of dynamic assessments of word reading in young children: A systematic review and meta-analysis. https://doi.org/10.1101/2022.09.19.22279942

**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)

Table 1.

Search terms for concept 1 – Dynamic assessment

Medline	**Embase**	**PsycINFO**	**CINAHL**	**ERIC**
No subject heading (dynamic OR mediat*) ADJ3 (Assess* OR test* OR screen* OR measur* OR tool*).tw,kf. OR respon* ADJ3 interven*.tw,kf. OR Modifiability ADJ3 Index.tw,kf. OR Learning Potential.tw,kf.	*Clinical assessment* OR *Clinical assessment tool* OR *Language test* dynamic OR mediat*) ADJ3 (Assess* OR test* OR screen* OR measur* OR tool*).tw,kw. OR respon* ADJ3 interven*.tw,kw. OR Modifiability ADJ3 Index.tw,kw. OR Learning Potential.tw,kw.	*Measurement* dynamic OR mediat*) ADJ3 (Assess* OR test* OR screen* OR measur* OR tool*).tw OR respon* ADJ3 interven*.tw OR Modifiability ADJ3 Index.tw OR Learning Potential.tw	*Speech and Language Assessment* TI (Dynamic N3 (Assess* OR test* OR screen* OR tool* OR task* OR measur*)) OR AB (Dynamic N3 (Assess* OR test* OR screen* OR tool* OR task* OR measur*)) OR TI ((Learning potential) N3 (assess* OR screen* OR test* OR tool* OR task* OR measur*)) OR AB ((Learning potential) ADJ3 (assess* OR screen* OR test* OR tool* OR task* OR measur*)) OR TI response to intervention OR AB response to intervention OR TI(comput* Adapt* test*) OR AB(comput* adapt* test*)	*Educational assessment* Dynamic) NEAR/3 (assess* OR test* OR screen* OR measur* OR tool*) OR (Mediat*) NEAR/3 (assess* OR test* or screen* or measur* OR tool*) OR (Respon*) NEAR/3(interven*) OR (Modifiability) NEAR/3 (Index) OR Learning potential

Table 2.

Search terms for concept 2 – Literacy

Medline	**Embase**	**PsycINFO**	**CINAHL**	**ERIC**
Literacy OR *Reading OR Writing*	*Literacy* OR *Reading*	*Literacy* OR *Reading* OR *Writing Skills* OR *Academic Writing*	*Literacy* OR *Reading*	*Emergent literacy* OR *Literacy* OR *Literacy education* OR *Literacy skills* OR *Assessment literacy* ...
phonem*.tw,kf.	phonem*.tw,kw.	phonem*.tw	TI phonem* OR AB phonem*	Phonem*
OR	OR	OR	OR	OR
phonolog*.tw,kf.	phonolog*.tw,kw.	phonolog*.tw	TI phonolog* OR AB phonolog*	Phonolog*
OR	OR	OR	OR	OR
phonic*.tw,kf.	phonic*.tw,kw.	phonic*.tw	TI phonic* OR AB phonic*	Phonic*
OR	OR	OR	OR	OR
(sound* ADJ3 (blend* OR segment* OR manipulat* OR substitut* OR delet*)).tw,kf.	(sound* ADJ3 (blend* OR segment* OR manipulat* OR substitut* OR delet*)).tw,kw.	(sound* ADJ3 (blend* OR segment* OR manipulat* OR substitut* OR delet*)).tw	TI (sound*)N3(blend* OR segment* OR manipulat* OR delet* OR substitut*)	(sound*) NEAR/3 (blend* OR segment* OR manipulat* OR substitut* OR delet*)
OR	OR	OR	OR	OR
((letter* OR alphabet*) ADJ3 knowledge)tw,kf.	((letter* OR alphabet*) ADJ3 knowledge)tw,kw.	((letter* OR alphabet*) ADJ3 knowledge)tw	AB sound*)N3(blend* OR segment* OR manipulat* OR delet* OR substitut*)	(letter*) NEAR/3 (knowledge)
(read OR reading).tw,kf.	(read OR reading).tw,kw.	(read OR reading).tw	OR	OR
OR	OR	OR	TI (letter* OR alphabet*) N3 (knowledge OR principle)	(alphabet*) NEAR/3 (knowledge OR principle)
(write OR writing).tw,kf.	(write OR writing).tw,kw.	(write OR writing).tw	OR	OR
OR	OR	OR		
(spell OR spelling).tw,kf.	(spell OR spelling).tw,kw.			

OR (decode OR decoding).tw,kf.	OR (decode OR decoding).tw,kw.	(spell OR spelling).tw OR (decode OR decoding).tw	AB (letter* OR alphabet*) N3 (knowledge OR principle) OR TI ((read OR reading)) OR AB ((read OR reading)) OR TI ((write OR writing)) OR AB ((write OR writing)) OR TI ((spell OR spelling)) OR AB ((spell OR spelling)) OR TI ((decode OR decoding)) OR AB ((decode OR decoding))	Write OR Writing OR Spell OR Spelling OR Decode OR decoding

Table 3.

Effect Sizes Representing the Relationship between Dynamic Assessments of Word Reading Skills (phonological Awareness, sound-symbol knowledge, and decoding) and Single Word Reading Measures

Study	N	DAPA – WRM	DASSK-WRNM	DA Dec - WRM
Aravena, et al., (2013)	42			0.52
Aravena, et al., (2018)	118			0.237
Barker & Saunders, (2020)	27		0.63	
Caffrey, (2006) Sample 1	120			0.624
Caffrey, (2006) Sample 2	95			0.745
Cho et al., (2014)	134			0.69
Cho & Compton (2015)	112			0.54
Cho et al., (2017)	105			0.46
Chow, (2014)	121		0.28	
Clayton et al., (2018)	97		0.12	
Compton et al., (2010)	355			0.69
Coventry et al., (2011)	1988	0.423*		
Cunningham & Carroll (2011) [a]	45	0.77*		
Edwards, (2020)	312			0.53
Gan et al., (2022)	135			0.7
Gellert & Elbro (2017a)	171			0.66
Gellert & Elbro (2017b)	160	0.47*		
Gillam et al., (2011)	64	0.54*		
Hautala et al., (2020)	723			0.49
Horbach et al., (2015)	243		0.215	
Horbach et al., (2018)	17			0.855

Law et al., (2018)	84			0.315
Liu et al., (2021)	203		0.19	
Loreti, (2015)	10	0.75**		
Osa Fuentes (2003)	164	0.64*		
Petersen & Gillam, (2015)	63			0.51
Petersen et al., (2016)	280			0.57
Sittner Bridges& Catts (2011) Sample 1 [b]	90	0.516*		
Sittner Bridges & Catts (2011) Sample 2 [c]	96	0.426*		
Spector (1992)	38	0.60*		
Wyman Chin (2018)	9	0.87**		
Yap (2018)	99	0.60*		

Note. DA= dynamic assessment, PA = phonological awareness, SSK= Sound-Symbol Knowledge, Dec= Decoding, WRM=word reading measure, N= number of participants per study

[a] Data used from 2010 Cunningham thesis

[b,c] Data used from 2009 Bridges thesis

*p<.05, **p<.01, *** p<.001

Table 4.

Quality Appraisal of Included Studies

Study	**Participants****	**Dynamic Assessment**		**Word Reading Measures**		**Flow and Timing**		**Statistical Analyses**	**Overall Score and %** **(/12)**	**Overall Appraisal** **Low – 0.33%** **Medium-34-66%** **High – 67-100%**
	Were participant characteristics adequately described? (1)	*Was the DA adequately described? (2)*	*Was the test appropriate for evaluating early literacy skills? (2)*	*Was the WRM adequately described? (2)*	*Was the WRM valid and reliable for evaluating reading skills* (2)	*Were all participants included in al analyses?* (1)	*If not, were reasons for exclusion or loss to follow up adequately described?* (1)	*Were appropriate statistical tests used?* (1)		
Aravena, S., Snellings, P., Tijms, J., & van der Molen, M.W. (2013)	1	2	2	2	2	0	1	1	11(92%)	High
Aravena, S., Tijms, J. Snellings, P. & van der Molen, M.W. (2018)	1	2	1	2	2	1	1	1	11(92%)	High
Barker, R.M. & Saunders, K.J. (2020)	1	1	2	2	2	0	1	1 ** only 27 participants	10(83%)	High
Caffrey, E. (2006)	1	2	2	2	2	0	1	1	11(92%)	High
Cho, E. & Compton, D.L. (2015)	1	2	2	2	2	1	1	1	12 (100%)	High
Cho, E., Compton, D., Fuchs, D., Fuchs, L.S. & Bouton, B. (2014)	1	2	2	2	2	0	1	1	11(92%)	High

Cho, E., Compton, D.L., Gilbert, J.K., Steacy, L.M., Collins, A.A. & Lindstrom, E.R. (2017)	1	2	2	2	2	0	1	1	11(92%)	High
Chow, B. W-Y. (2014)	1	2	2	2	2	2	1	1	12(100%)	High
Clayton, F. (2018)	0	2	2	1	2	0	1	1	9(75%)	High
Compton, D. L., Fuchs, D., Fuchs, L. S., Bouton, B., Gilbert, J. K., Barquero, L. A., Cho, E., Crouch, R.C. (2010)	1	2	2	2	2	0	1	1	11(92%)	High
Coventry, W.L., Byrne, B., Olson, R.K., Corley, R. & Samuelsson, S. (2011)	1	2	2	2	2	0	0	1	10(83%)	High
Cunningham, A. & Carroll, J. (2011)	0	2	2	2	2	1	1	1	11 (92%)	High
Edwards, A. (2020)	0	0	2	2	2	1	1	1	9(75%)	High
Gellert, A.S. & Elbro, C. (2017a)	1	2	2	2	2	0	1	1	11(92%)	High
Gellert, A.S. & Elbro, C. (2017b)	1	2	2	2	2	0	1	1	11(92) %	High
Gillam, S.L., Fargo, J. Foley, B. &	1	2	2	2	2	1	1	1	12(100%)	High

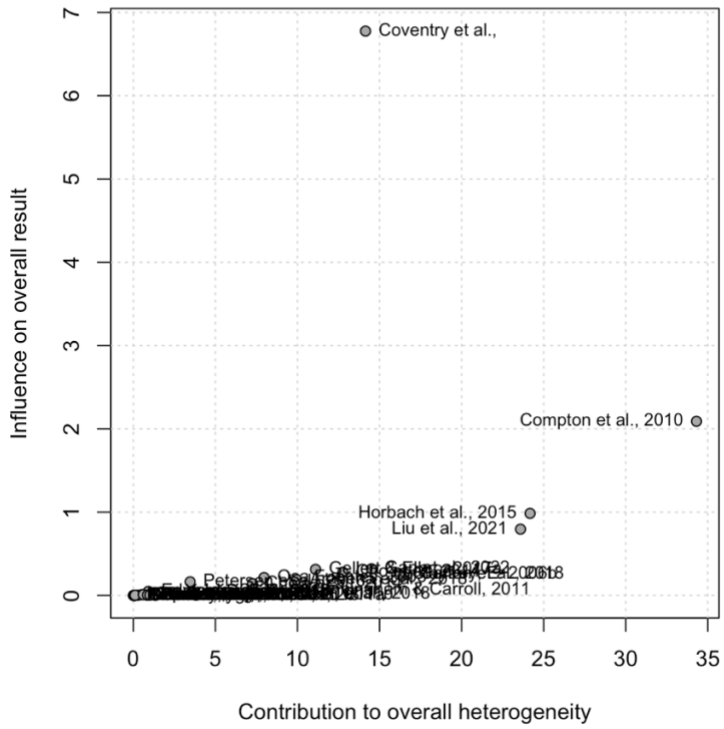
Olszewski A. (2011)										
Hautala, J., Heikkila, R., Nieminen, L., Rantanen, V. Latvala, J-M. & Richardson,U. (2020)	1	2	2	2	2	0	1	1	11(92%)	High
Horbach, J. Weber, K. Opolony, F. Scharke, W. Radach, R. Heim, S. & Günther, T. (2018)	1	2	2	2	2	0	1	0	10(83%)	High
Horbach, J., Shcarke, W., Cröll, J., Heim, S. & Günther, T. (2015)	1	2	2	2	2	0	0	1	10(83%)	High
Law, J.M., De Vos, A., Vanderauwera, J., Wouters, J. Ghesquiere, P. & Vandermosten, M. (2018)	1	2	2	2	2	1	1	1	12(100%)	High
Liu, C., Hoa Chung, L.C., Wang, L.C. & Liu, D. (2021)	1	1	2	2	0	1	1	1	9(75%)	High
Loreti, B. (2015)	1	0	1	2	2	1	1	1	9(75%)	High
Osa Fuentes, P.M. (2003)	0	2	2	2	0	1	1	1	9(75%)	High

Petersen, D. B., & Gillam, R. B. (2015)	1	2	2	2	2	1	1	1	12 (100%)	High
Petersen, D.B, Allen, M.M. & Spencer, T.D. (2016)	0	2	2	2	2	1	1	1	11(92%)	High
Sittner Bridges, M. & Catts, H.W. (2011)	0	2	2	2	2	1	1	1	11 (92%)	High
Spector, J. (1992)	1	2	2	2	2	0	1	1	11(92%)	High
Wyman Chin, K.R. (2018)	0	1	1	2	2	1	1	1	9(75%)	High
Yap, D.F-F. (2018)	1	2	2	2	2	0	1	1	11(92%)	High

Note. DA= Dynamic assessment, SA= Static assessment, WR = word reading, Low = 0-33%, Medium=34-66%, High=67-100%

Figure 1.

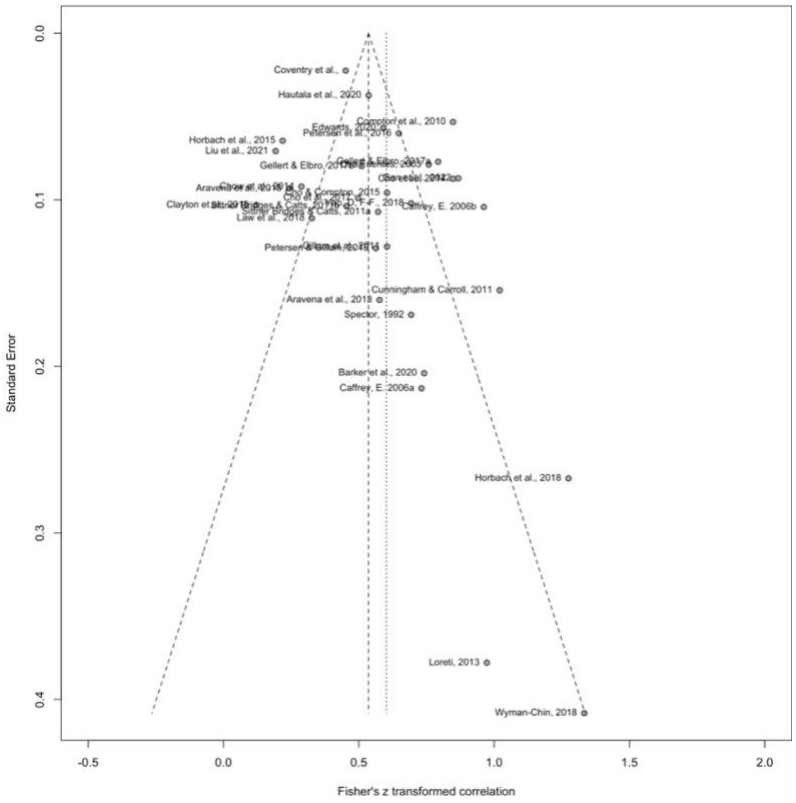
Baujat plot for studies included in the meta-analysis of the relationship between dynamic assessments of phonological awareness and word reading outcome measures



Note. In the Baujat plot, individual contribution to overall heterogeneity is represented on the horizontal axis, and the influence on overall result on the vertical axis. Studies with greatest influence are found in the top right quadrant of the figure. Drawn in R using the 'metacor' package (R Core Team, 2021; Laliberté, 2019).

Figure 2

Funnel plot of studies included in the meta-analyses of the relationships between dynamic assessments of word reading skills (phonological awareness, sound-symbol-knowledge, and decoding) and word reading measures



Note. In the funnel plots, individual Fisher z transformed effect sizes are presented on the horizontal axis, and the standard error on vertical axis. Studies with smaller standard errors (larger studies) are found closer to the top of the plot. Drawn in R using the ‘metacor’ package (R Core Team, 2021; Laliberté, 2019).