medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Concurrent and predictive validity of dynamic assessments of word reading in young children: 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

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

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

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

Email: e.wood@utoronto.ca

Author Note

The authors declare no conflicts of interest.

Funding

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 E. Wood, by a University of Toronto Excellence Award, awarded to K. Biggs, and by a Natural Sciences and Engineering Research Council of Canada grant awarded to Dr. M. Molnar (RGPIN-2019-06523).

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

Abstract

Early evaluation of word reading skills is an important step in understanding and predicting children's future literacy abilities. Traditionally, word reading evaluations are conducted using 'static' assessments (SA), which measure a child's acquired knowledge and are prone to floor effects. Additionally, many of these tools are developed exclusively for English monolinguals, and therefore cannot be used equitably to evaluate the abilities of bilingual children. Dynamic assessment (DA), which evaluates the ability to learn a skill, is a potentially more equitable alternative. To establish that use of DAs is a valid alternative to traditional SAs, their concurrent agreement with gold standard SA measures and their predictive agreement with later word reading outcomes should be considered. In line with this, the primary objective of this systematic review and meta-analysis is to examine the concurrent and predictive validity of DAs of word reading skills. Two secondary objectives are (i) to address which types of word reading DAs (phonological awareness, sound-symbol knowledge, or decoding) demonstrate the strongest relationships with equivalent concurrent static measures and later word reading outcomes, and (ii) to consider for which populations, defined by language status (monolingual vs. bilingual vs. mixed) and reading status (typically developing vs. at-risk vs. mixed) these DAs are valid. Thirty-four studies from 32 papers were identified through searching 5 databases, and the grey literature. Included studies provided a correlation between a DA and concurrent SA, or a DA and a later word reading outcome measure. Regarding concurrent validity, we observed a strong relationship between DAs and SAs in general (r=.60); however, subgroup analyses indicate that DAs of decoding (r=.54) and phonological awareness (r=.73) measures demonstrate greater strength of correlation with their static counterparts, compared to DAs of sound-symbol knowledge (r=.34). In terms of predictive validity, we observed a similarly strong relationship between DAs and word reading outcome measures (r=.57), independently of the type of measure. Subgroup analyses conducted based on participant

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

language status suggested that there are significant differences between mean effect sizes for monolingual, bilingual and mixed language groups in terms of DAs' concurrent validity with SAs, but no significant differences for predictive validity with word reading outcome measures. There were also no significant differences between mean effect sizes for at-risk, typically developing, or mixed groups in terms of DAs concurrent validity with SAs or predictive validity with word reading outcome measures. Results provide preliminary evidence to suggest that DAs of phonological awareness and decoding skills are a valid alternative to SAs of equivalent constructs and are valid for the future prediction of word reading outcomes across population groups regardless of their language or reading status.

Keywords: Dynamic assessment, Static assessment, Concurrent validity, Predictive Validity, Early literacy, Reading, Decoding, Phonological awareness, Alphabetic principle, Bilingual, At-risk

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

Introduction

Literacy

Literacy is the ability to understand, interpret, create, and communicate with information in print, is necessary for social, academic, professional, and personal success and is considered a fundamental human right (Montoya, 2018; Moretti & Frandell, 2013). In 2017, the United Nations Educational, Scientific and Cultural Organization (UNESCO) reported that more than 50% of children worldwide had literacy difficulties, and problem that has been exacerbated by the COVID-19 pandemic (UNESCO, 2021; Aurini & Davies, 2021). Inadequate literacy skills are associated with negative life outcomes, such as poor physical (Wolf et al., 2005) and mental health (Daniel et al., 2006), reduced academic attainment (Ritchie & Bates, 2013), restricted socioeconomic mobility (Barwick & Siegel, 1996), and increased rates of poverty, homelessness, and incarceration (Shelley-Tremblay et al., 2007).

Research has established that early identification is key in mitigating future literacy difficulties and their associated adverse effects (Lundberg, 1994; Ontario Human Rights Commission, 2022). The construct of word reading skills in this review was informed by the subskills that comprise word recognition ability in the evidence-based reading model - Scarborough's Reading Rope (2001). This model identifies three subskills (i) phonological awareness (PA)– the ability to identify and manipulate parts of speech, (ii) knowledge of the alphabetic principle, or sound-symbol knowledge (SSK)- the ability to recognize the systematic relationship between symbol(s) (letter) and the sound(s) they represent in print and (iii) sight recognition – the ability to apply PA and SSK skills to rapidly and automatically read or decode words, as essential for success in early word recognition and word reading (Scarborough, 2001). While these early word reading skills are foundational, they alone are not sufficient for developing reading ability, as children must also comprehend what they decode. However, word recognition skills contribute more greatly to initial reading success than comprehension skills (Scarborough, 1998). Numerous studies demonstrated that word recognition skills play a significant role

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

in early reading development and prediction of future reading outcomes (e.g., Catts et al., 2005; Hogan et al., 2005). Given that this review is focused on assessments used to evaluate children who are still in the stages of learning to read and decode (i.e., ages 4-9) our goal is to examine the validity of DAs of word reading ability – specifically, phonological awareness, sound-symbol knowledge, and decoding. Examining early reading comprehension is beyond the scope of the current review.

Traditional Static Word Reading Assessment

A range of traditional standardized assessment tools (e.g., The Comprehensive Test of Phonological Processing-2 (CTOPP-2), Wagner et al., 2013; Phonological Awareness Test-2: Normative Update (PAT-2:NU), Robertson & Salter, 2017; Test of Auditory Processing Skills-4 (TAPS-4), Martin et al., 2018) are commonly used by clinicians, educators, and researchers, to assess early word reading skills. These so-called "static assessments" (SAs) quantify a child's acquired knowledge and either compare their performance to same-aged peers through norm-referenced scores or determine if they can demonstrate an expected skill by evaluating whether they can complete a specific criterion-referenced task (Grigorenko & Sternberg, 1998). Two potential difficulties are associated with these traditional SAs. First, many of these assessments have been developed exclusively for use with English monolinguals. This English, monolingual-centric focus in test development is at odds with the global population as about half of individuals speak at least two languages (Grosjean, 2010). From both a research and clinical perspective, it is inequitable to use English monolingual assessments to evaluate the skills of a bilingual child. Not only might these tools be linguistically and culturally inappropriate, but testers also cannot be sure if performance differences are due to lack of ability, or language effects (bilingual vs. monolingual) that mask a child's capacity to perform the skill. A bilingual kindergarten student who speaks Tamil at home and begins learning English in school is likely to perform much more poorly than a child who has grown up speaking exclusively English on a static English test of letter-

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

sound knowledge not because they lack the ability to learn letter sound relationships, but either because of limited exposure to English letters and sounds relative to English monolingual peers; or because of an inability to understand English instructions or a lack of familiarity with Eurocentric or Western assessment practices and ways of measuring knowledge. These cultural and language effects associated with SAs often result in misidentification of reading difficulties in bilingual populations (Bedore & Peña, 2008; Petersen & Gillam, 2015).

The second potential issue associated with SAs is floor effects (Catts et al., 2009). Floor effects are commonly observed in traditional assessment of word reading because these tools attempt to quantify a child's acquired ability in an area with which they have limited to no experience. Many kindergarten-aged children, even English monolinguals for whom the tests are developed, perform poorly on SAs of word reading ability at the start of the school year, simply because they have had limited experience with these types of tasks. When most children who take a test perform poorly, examiners are unable to differentiate those who truly are at risk for future reading challenges from those who are so-called false positives; students with limited previous experience who will quickly catch up, or those whose linguistic and cultural experiences did not permit them to demonstrate their capabilities on the test. Traditional tests may underestimate the capabilities of a child with minimal literacy experiences, by suggesting that their current lack of knowledge is predictive of their future ability.

Dynamic Assessment as an Alternative

Dynamic assessment (DA) is an alternative to the traditional SA paradigm. Unlike SAs which evaluate acquired knowledge DAs attempt to measure the ability to learn. This is achieved through interactive testing which can incorporate teaching, scaffolding, prompting and feedback, resulting in an assessment that more closely resembles real world learning experiences (Grigorenko & Sternberg, 1998; Poehner, 2008). Interest in DAs of word reading abilities has been steadily growing in the research

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

community, (e.g., Cho et al., 2017; Gellert & Elbro, 2017a; Petersen et al., 2016), but their clinical use remains limited, with professionals like Speech-Language Pathologists (SLPs) continuing to favour use of SAs (e.g., Arias & Friberg, 2017; D'Souza et al., 2012).

A potential reason for this lack of uptake in use of DA could be that these tools, which are typically less structured than SAs and often unstandardized, are not viewed as a psychometrically valid alternative to SAs. Examining DAs concurrent validity with traditional SAs and predictive validity with word reading outcome measures will contribute to understanding DAs potential as an alternative method for evaluating and predicting literacy abilities. This is the overarching goal of the current systematic review and meta-analyses.

Previous Reviews of Dynamic Assessment

Several prior reviews have evaluated the use and validity of DAs. Here, we focus on three that address DAs of literacy. Caffrey et al.'s 2008 systematic review and correlational meta-analysis offered support for the predictive validity of DAs, but their analyses did not differentiate between the various included domains (e.g., math, cognition and reading). Furthermore, at-risk and bilingual children were merged in their analyses, despite bilingualism not being an inherent risk factor for reading disorders or any disorders in general. In two subsequent systematic review papers, Dixon et al. (2022a, 2022b) investigated whether DAs can uniquely predict variance in the growth of a child's reading development beyond SAs, and whether DAs can act as a viable alternative to diagnosing reading disorders in children. In these reviews, the assessment of specific skills that typically predict future reading success were considered, such as phonological awareness, decoding, and morphological awareness. Their findings suggest that DAs of phonological awareness, decoding, and morphological awareness can account for variance in the growth of different outcome measures, ranging from 1-33% (Dixon et al., 2022a); and that DAs can account for unique variance in predicting later reading disorder, particularly when the test

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

construct is similar to reading (e.g., a DA of decoding better predicts word reading outcomes than a DA of working memory) and when predicting abilities proximally vs distally (e.g., in early vs later school years; Dixon et al., 2022b).

While findings from these three previous reviews are promising, there are gaps that remain unaddressed. First, no prior review has examined the concurrent validity of DAs of word reading skills, qualitatively or quantitatively. Evaluating a tool's concurrent agreement with gold standard measures is a key component of establishing its criterion related validity and determining whether it can serve as a legitimate alternative to established assessments. Furthermore, while the Caffrey et al., (2008) and Dixon et al., (2022a, 2022b) studies considered predictive validity of DAs, neither conducted quantitative analyses which systematically examined (i) the predictive validity of distinct types of word reading word reading DAs or (ii) the validity of word reading DAs for use with populations based on their language (monolingual vs. bilingual) and reading status (at-risk vs. typically developing). All of which are important for understanding the validity of DAs across various populations and contexts.

The current study

This systematic review and correlational meta-analysis will address these gaps. First, we will quantitatively evaluate the concurrent validity of DAs of word reading skills (phonological awareness, sound-symbol knowledge, and decoding), with their equivalent SAs, as well as the predictive validity of these same word reading DAs with later word reading outcomes, defined in this review as single word or nonword reading. Secondly, we will investigate whether these DAs of word reading skills demonstrate consistent concurrent and predictive validity across population groups, defined by language status (monolingual vs. bilingual), and reading status (at-risk vs. typically developing). Finally, unlike all previous reviews of DA, we will conduct an extensive grey literature search, to reduce potential publication bias, and will include studies published in languages other than English (Spanish, French).

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

Method

Methods and analyses were planned a priori and outlined in a systematic review and meta-analysis protocol. This protocol was preregistered on Open Science Framework and is available at online at https://osf.io/bcghx/ (Wood & Molnar, 2022).

Research Questions

This systematic review and meta-analyses was designed to address the following questions:

1.A) Do dynamic assessments of word reading skills (phonological awareness, sound-symbol knowledge, and decoding), demonstrate ***concurrent validity*** with static assessments of word reading skills (PA, SSK, decoding) across all populations?

1. B) Do dynamic assessments of word reading skills demonstrate ***predictive validity*** with reading outcome measures (single word reading) across all populations?

2. A) Do dynamic assessments of word reading skills demonstrate concurrent validity with static assessments of word reading skills and predictive validity with word and nonword reading outcomes within population groups defined by their ***language status*** (monolingual vs. Bilingual)?

2. B) Do dynamic assessments of word reading skills demonstrate concurrent validity with static assessments of word reading skills and predictive validity with word reading outcome measures within population groups defined by their ***reading status*** (at-risk vs. typically developing)?

Eligibility Criteria

Study inclusion criteria were decided upon in advance and outlined in the systematic review and meta-analysis screening protocol (Wood & Molnar, 2022):

(i) Only primary research articles were included. Case reports, commentaries and editorials were excluded, as well as systematic reviews and books or book chapters. Articles published in peer-reviewed

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

journals, and unpublished grey literature from preprint repositories, and reports or dissertations were included.

(ii) Given that the focus of the current paper is early word reading skills, only studies that evaluated children with a mean age of 4;0 – 10;0, whose participants were either typically-developing, at-risk for reading difficulty, diagnosed with a reading disorder and monolingual or bi/multilingual were included. Studies conducted with adults or with children with developmental difficulties (e.g., developmental language disorder, autism spectrum disorder, hearing difficulty) were excluded.

(iii) All included studies used a DA of word reading skills and (i) a SA of word reading skills at the same time point, AND/OR (ii) a word reading outcome measure at a later timepoint. Included studies reported correlation coefficients to quantify relationships between DAs and SAs and/or DAs and word reading outcome measures.

(iv) Articles published in English, French or Spanish, or those written in another language but with full text translations were included. No exclusions were made based on setting.

Search Strategy and Information Sources

An initial search was carried out in March 2022 on the following 5 databases: MEDLINE, Embase, CINAHL (Cumulative Index to Nursing and Allied Health Literature), PsycINFO and ERIC (Education Resources Information Center), using two concepts “dynamic assessment” and “literacy” and their associated keywords in titles and abstracts. Synonyms for dynamic assessment included (dynamic test* OR screen* OR tool* OR task* OR measur*) OR (learning potential assess* OR screen* OR test* OR tool* OR task OR measur*), OR (response to intervention). Associated keywords for literacy included phonem* OR phonolog* OR phonic* OR (sound* blend* OR segment* OR manipulat* OR substitut* OR delet*), OR (letter* OR alphabet* knowledge or principle) OR read OR reading OR write OR writing OR spell OR spelling OR decode OR decoding. No filters were used in the search process.

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

This search strategy was developed with support from the University of Toronto librarian. Following the initial database search, in June 2022 an additional synonym for "dynamic assessment" was added – computerized adaptive testing. This search term (comput* adapt* test*) was rerun on the same databases with all synonyms for the key word "literacy" and results of this search were screened. Computerized adaptive tests (CAT) were not identified as a method/synonym of DA in the initial preliminary searches but were determined to be necessary to search following identification of a study that utilized a dynamic CAT approach to reading assessment. For a list of search terms used in each database see Figure 1 and Figure 2 in Appendix 1.

Next, a search was performed in 3 preprint repositories, MedRxiv, EdArxiv and PsyArxiv. The same two concepts "dynamic assessment" and "literacy" were used to conduct this search. Following the database and preprint repository search, the first author and second author began forward searching of included articles on Google Scholar. The "cited by" function was used to identify articles that had cited the included/relevant studies identified from the database and preprint search. Subsequently, an ancestral search of the included articles was then conducted. The reference lists of the included articles were reviewed by one of ten coders or the first author and crosschecked with the included article list to determine if there were any articles of interest that had not been identified in the database, preprint, or Google Scholar search. Finally, requests for unpublished data or studies were posted to lab and researcher social media accounts and sent out on two occasions to relevant mailing lists, and to labs across Canada, the United States and Europe conducting early literacy research. An updated search of databases and preprint repositories was conducted in December 2022 to identify new relevant articles.

At each stage of screening, articles were rated by two independent reviewers, either the first author or one of ten trained research assistants (RAs). In the title/abstract stage reviewers voted whether an article was relevant or irrelevant, in the full-text eligibility screening, reviewers voted to include or

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

exclude a study based on its' characteristics. In all phases, disagreements were resolved by the first author. Given that 11 reviewers participated in the article identification process, there were many unique pairings of raters (i.e., 66 pairings at the title/abstract stage). Consequently, calculation and reporting of Cohen's Kappa coefficients for all reviewer pairings was not meaningful. Rather, the average Kappa coefficient for all rater pairs was calculated to be 0.29 (90% proportionate agreement) for the title abstract screening and 0.39 (76% proportionate agreement for the full text review screening, which can both be characterized as fair (Cohen, 1960; McHugh, 2012).

Data Collection Process

Following identification of relevant articles, data was extracted using a template generated on Covidence. The same team who completed the title/abstract and full text screening performed the extraction. All coders received a training session led by the first author prior to extraction. All articles were extracted by two reviewers. Conflicts and consensus were completed by the first author. For each study, the following data points were extracted:

Data Items

General Information

Study title, journal, year of publication, the DOI, author names and institutional affiliations, the country in which the study was conducted, and whether the project received any funding or reported any conflicts of interest.

Study Type

Studies were coded as **c**ross-sectional or longitudinal. Longitudinal studies that also included a cross-sectional correlation between SA and DA measures at a single timepoint and a correlation between DA and a reading OM across two timepoints were counted as both cross-sectional and longitudinal.

Participants

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

For each study, the total number of participants, the percentage of males and the mean age and grade at the study outset were coded. Age and grade are not consistent across countries so both data points were required. The reading status (typically developing, at-risk or diagnosed with a reading difficulty- or any combination of these groups), and the language status (monolingual, bilingual, or a combination of both) of the study participants was coded, along with which language(s) were reported to be spoken by the participants.

Measures

Dynamic Assessment(s) DAs in this review are defined as any assessment which provided explicit or implicit teaching, training, feedback on performance or prompting in the context of the assessment. Coders extracted the name of the DA if one was provided, the word reading skill(s) that the DA evaluated (either PA, SSK or decoding, or a combination of two or three of these skills) and a brief description of the specific task used to evaluate the literacy skill (e.g., PA-phoneme segmentation, or SSK-letter-sound knowledge of the English alphabet). If more than one task was used to evaluate a skill, as was often the case, coders listed all tasks. These data points permit comparison of the concurrent and predictive validity of each construct of word reading DA.[1]

Static Assessment(s) and Word Reading Outcome Measures (OMs) SAs and OMs in this review are any assessments which evaluate a skill using a binary correct/incorrect response scoring system, and which are characterized by the absence of feedback, prompting or training and teaching components in the assessment. SAs in this review are tests that are conducted at the same time point

[1] Coders also reported whether the DA was administered in the traditional "in-person" format, or via computer, which type of DA was used, the Train/Test or Graduated Prompt format, and whether the DA used real words or nonwords, familiar letters/characters of novel symbols. Results of the comparisons between in-person and computerized DA, train/test vs. Graduated prompts DA, word/nonword and familiar/novel symbol will be reported in a separate forthcoming study regarding the relationship between the different characteristics of DAs of word reading skills and word reading measures.

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

(concurrently) as the DA, while OMs are tests that are conducted at a later time point and used to investigate predictive validity of DA. Both SAs and OMs in this review can be norm-referenced tests (e.g., CTOPP), criterion-referenced tests (e.g., the Dynamic Indicators of Basic Literacy Skills) or researcher developed tools. When extracting information related to SAs, coders indicated the name and type of any assessments used (e.g., CTOPP, norm-referenced), the word reading skills evaluated (PA, SSK, Decoding or a combination) and the specific tasks used to evaluate these skills (e.g., PA-phoneme blending, SSK- novel symbol-sound knowledge). Regarding outcome measures, coders identified the name, if any, of the outcome measure used (e.g., WRMT-R NU) and the specific subtests of the measure (e.g., Word Attack) and the skill evaluated (e.g., nonword reading accuracy).

Effect Sizes

The correlation coefficients representing the relationships between the DAs and SAs, and/or the DAs and OMs were extracted. If a study reported multiple correlations between a DA and an SA (e.g., a DA of PA that utilized multiple PA tasks, and a SA evaluating PA that also employed multiple PA tasks), or a DA and an OM, coders were instructed to extract all relevant correlations coefficients at this stage. Following review of all extracted coefficients, the authors created a set of decision rules for choosing a single correlation coefficient to represent the relationship between the DA and SA, or the DA and the OM for each analysis, to ensure that the synthesis did not violate the assumption of independence. These decision rules for selecting a single effect size were made based on which measure was most consistently used across studies. For example, word reading accuracy was reported in the majority of included study as an outcome measure, and as such it was identified as the primary outcome measure for estimating predictive validity. Effect sizes between DA and WR accuracy were selected over those that were less commonly observed, like non-word reading, passage reading, or word reading fluency. Similarly, because most studies examined Kindergarten aged students at time point 1, and

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

Grade 1 students at time point 2, effect sizes representing the association between a DA in kindergarten and an OM in grade 1 were favoured over those representing less commonly observed time points such as preschool and grade 2.

In instances where when one measure was not more common than all others, decision rules grounded in literacy theory were used. For example, there was not a single most used PA task, and so, when possible complex PA tasks (e.g., manipulation) were preferred over simple phonemic awareness tasks (e.g., blending), and smaller unit tasks (e.g., phoneme level) were preferred over larger unit phonological awareness tasks (syllable level tasks) as these complex, smaller grain, tasks have consistently been linked to later decoding success (Høien et al., 1995). In terms of SSK tasks- those which required a child to name a make a connection between a symbol and a sound were preferred over those that required mere naming of the symbol. This is because this skill more closely approximates the construct of the alphabetic principle, which research has determined is an excellent predictor of later reading ability (Ehri, 1998). Finally, when choosing decoding tasks – single real word, untimed, decoding tasks were prioritized over timed, nonword or passage level decoding tasks as this best represents the construct of decoding as it is defined in this review. The coefficients representing the effect sizes for concurrent and predictive validity are presented in Tables 5 and 6, respectively.

Quality Appraisal Assessment

Two trained independent reviewers assessed the included studies using a modified and combined 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 evaluated on the following five domains (i) participant selection, (ii) index/dynamic assessment, (iii) reference/static assessments and/or outcome measures, (iv) flow and timing of the study, (v) statistical analysis. Please refer to Table 3 in Appendix 2 for the full list of questions and ratings for each study.

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

Regarding participants, reviewers rated whether the participant sample was adequately described in terms of age, sex breakdown, language and reading status and demographic characteristics. Reviewers evaluated the DA domain by rating whether it was described with sufficient detail in terms of the skills evaluated, the format of the test, the prompting and scoring used, and administration process. They also recorded whether the task used to evaluate the word reading skill(s) in the DA was developmentally appropriate for the population. When assessing the domain related to the reference standards (SAs and OMs) reviewers evaluated whether the studies employed developmentally appropriate tools for evaluating word reading skills or word reading outcomes. They also rated whether psychometric properties of the reference measures used were reported. In evaluating flow and timing, reviewers noted whether all participants were included in the analyses and whether authors explained and accounted for reasons for loss to follow up and attrition if necessary. Finally, coders assessed whether appropriate statistical analyses were used to draw conclusions about study findings.

In summary, this yielded 8 items across the 5 domains to be rated. Items relating to participants, flow and timing and statistical analyses were weighted one point, while items pertaining to the index test (DA) and the reference tests (SA and OMs) were weighted two points given their relative importance in addressing the review objectives. Following ratings by two reviewers, conflicts were resolved by the first author, and studies were ranked as either low quality (0-33%), medium quality (34-67%) or high quality (68-100%). Only studies rated as medium and high quality were included in the analyses. Please refer to Appendix 2 Table 3 for the questions and study ratings.

Analyses

We used a random effects model for our meta-analyses (Borenstein et al., 2010), and included subgroup analyses by DA type (PA, SSK and decoding), language status (monolingual, bilingual and mixed) and reading status (typically developing, at-risk/diagnosed with reading difficulty and mixed).

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

The 'metacor' package (Laliberté, 2019) in R studio (R Core Team, 2021) was used to conduct a Fisher Z transformation of Pearson's correlation coefficients into Z scale scores (Corey et al., 1998; Silver & Dunlap, 1987). Following the transformation, a weighted average of these values was then calculated and transformed back to Pearson r correlation coefficients with accompanying p values for interpretation. To provide a robust picture of the degree heterogeneity between studies, Q, I^2 and τ^2 heterogeneity statistics were calculated (Borenstein et al., 2017; Higgins & Thompson, 2002). To evaluate which studies contributed most to overall heterogeneity, Baujat plots were generated and are presented in Figure 12 and 13 in Appendix 3 (Baujat et al., 2002). To examine the potential risk of publication bias, funnel plots for both the concurrent and predictive validity analyses were generated in R studio (R Core Team, 2021). Egger's regression test, a test of significance, was conducted to determine objectively whether funnel plot asymmetry was present (Egger et al., 1997).

Results

Study Selection

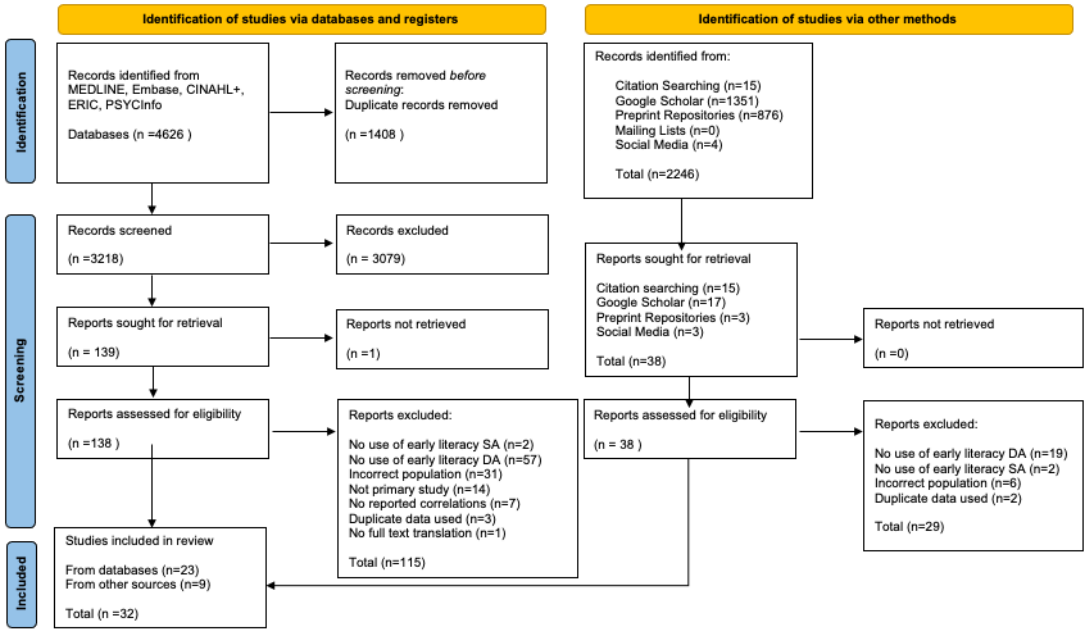
The process of study identification is visualized in Figure 3 in the PRISMA flowchart below (Page et al., 2021). The database searches produced 4,626 articles which were uploaded to Covidence. The software automatically detected and removed 1408 duplicate articles leaving 3218 titles and abstracts for review. Of these articles, 138 were reviewed at the level of full text, and 23 articles were identified for inclusion. Next, the 876 articles identified from the preprint repositories searches were uploaded. There were 0 duplicates. Following title/abstract screening, 3 articles were reviewed as full texts, and 1 article was included for analysis. An additional 1,351 articles were identified via forward Google Scholar searching of the 24 already included articles. Over three rounds of iterative searching, 17 relevant articles were identified, of which 11 were excluded and 7 included. Subsequently, an ancestral search of the 31 identified articles was completed. This yielded 15 potential articles reviewed

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

at the full text level. One additional study was deemed relevant for inclusion. Finally, the callout to social media, mailing lists and labs was made. Two authors contacted the first author to share 4 papers. 3 papers were reviewed at the full text level and 0 were included. In summary, 32 papers including a total of 34 studies were identified for inclusion in the systematic review and meta-analysis. The process of study identification is visualized in the PRISMA flowchart below (Page et al., 2021). Reasons for exclusion are also reported.

Figure 3.

PRISMA Flowchart of Literature Search



Study Characteristics

Age and Sex

A total of 6791 participants were included across 34 studies in 32 included articles. The mean age across participants was 5;8. The most common grade tested at a single time point (for cross-sectional studies) or at timepoint 1 (for longitudinal studies) was Kindergarten, and the most common follow-up grade was approximately one year later in grade 1 (n=9) while the second most common follow-up age was later in the kindergarten year (n=6). Across studies, the mean % of males were

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

50.92%. Mean age, % of males included and grade at start was not reported in all included studies. Table 1 below lists the breakdown of mean age and % males across subgroups stratified by language and reading status.

Table 1.

Participant Characteristics

	N	Mean start age*	Mean % males**
Total (min-max)	6696 (9-1988)	5;8 (4;1 – 9;9)	50.92% (41%-67%)
Language Status			
Monolinguals	3812	6;5	51.1%
Bilinguals	231	5;3	49.1%
Mixed language	2653	5;2	48.03%
Reading Status			
Typically developing	3445	5;2	50.85%
At-risk	348	7;7	52.4%
Mixed reading	2903	6;4	49.8%

Note. N= Number of participants

* 23/34 studies reported mean age of participants at study onset

**26/34 studies reported mean % males in their participant sample

Languages Spoken and Language Status

The majority of studies were conducted in the United States (n=17), followed by the Netherlands (n=3), Denmark (n=2), China (n=2), England (n=2), Germany (n=2), Belgium (n=1), Finland (n=1), Singapore (n=1), Spain (n=1) and Taiwan (n=1). Of the 34 studies, 24 included monolinguals only. The languages spoken by the monolinguals included English (n=15), Danish (n=2), Dutch (n=2), Mandarin (n=2), Finnish (n=1), Spanish (n=1) and German (n=1). Six studies included both monolingual and bilingual populations in their analyses. Of those 6, only 1 specified the languages spoken by both the mono and bilinguals in their study (English monolinguals and Spanish/English bilinguals). All 5 other studies did not provide linguistic details about the included bilingual groups but did indicate that the monolinguals spoke Danish (n=2), German (n=1), English (n=1), or English, Swedish or Norwegian

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

(n=1). Finally, five studies included only bilingual groups. Of those, three were conducted with Spanish/English bilinguals, 1 with Mandarin/English bilinguals and 1 with English/Chinese, Malay, or other bilinguals. Please see Table 2 below for further information about the language status and languages spoken by participants in included studies.

Reading Status

Seventeen of the 34 studies included exclusively participants who were typically developing. Two studies conducted their correlational analyses separately for participants diagnosed with dyslexia, and 3 exclusively included participants at-risk for reading difficulty. Given the limited number of studies conducted with exclusively at-risk children or those diagnosed with difficulty, this group was merged. Finally, 12 studies included a mix of typically developing and at-risk groups in their analyses. Table 2 provides more information about participants' reading status in included studies.

Dynamic Assessments

Fourteen of the included studies evaluated phonological awareness via syllable or phoneme identification (n=5), syllable or phoneme deletion (n=4) and segmentation (n=5) tasks. Five DAs evaluated sound-symbol knowledge via novel symbol-sound or symbol-syllable learning tasks. Fifteen DAs evaluated decoding via either nonwords (n=11) or real words (n=4) and using novel symbols (n=7) or known letters (n=8). Four studies used implicit feedback via a game in their DA, while the remaining 30 employed explicit verbal and/or visual feedback. Table 2 provides additional information about characteristics of DAs in the included studies.

Static Assessments and Word Reading Outcome Measures

Of the 29 studies that reported a correlation between a DA and an SA, the majority used a norm-referenced tool as the SA (n=15), followed by researcher developed tools (n=12) and standardized screening tools (n=2). The majority of the 18 longitudinal studies included in this review used a norm-

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

referenced tool at follow-up to evaluate word reading abilities (n=14). Fewer used researcher developed tools (n=3), or standardized screening tools (n=1). The most common word reading outcome measure was the Word Identification subtest (n=6) from the Woodcock Reading Mastery Test- Revised, Normative-Update. Table 2 below provides additional information regarding characteristics of SAs and word reading outcome measures in the included studies (e.g., names of tests and subtests).

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 (Years; months)	Grade T1	Sex % Males	Lang Status	Read Status	Study Timing	DA Skill	DA Feedback	SA Skill / Subtest/ Test	WR-OM 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	Mixed TD & Dx.	Cross-sectional	Decoding rate real words of novel symbols	Implicit	Decoding words rate / 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	Mixed TD & Dx.	Cross-sectional	Decoding real words of novel symbols	Implicit	Decoding words 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	Letter-sound knowledge / Curriculum based measure	-
Caffrey, E. (2006) Sample 1	United States	25	Not stated	K	52%	ML English	TD	Longitudinal	Decoding nonwords	Explicit verbal	-	Word Reading/ WRAT
Caffrey, E. (2006) Sample 2	United States	95	Not stated	G1	57%	ML English	TD	Longitudinal	Decoding nonwords	Explicit verbal	-	Word Reading/ WRAT
Catts, H.W., Nielsen, D.C., Sittner Bridges, M., Liu, Y. S., &	United States	103TD 263AR	Not stated	K	51.4%-TD 57% AR	ML English	Mixed TD & AR	Cross-sectional	PA syllable & phoneme deletion real words	Explicit verbal	Phoneme matching /Sound matching/ CTOPP	-

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Bontempo, D.E. (2015)												
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	Decoding nonwords / Word attack/ WRMT-RNU	-
Cho, E. & Compton, D.L. (2015)	United States	112	6;8	G1	57.1%	ML English	TD	Cross-sectional	Decoding nonwords of novel symbols	Explicit verbal	Decoding nonwords /Word attack WRMT-RNU	-
Cho, E., Compton, D.L., Gilbert, J.K., Steacy, L.M., Collins, A.A. & Lindstrom, E.R. (2017)	United States	54 TD 51 AR	6;7 TD 6;10 AR 6;8	G1	55%	ML English	Mixed TD & AR	Cross-Sectional Longitudinal	Decoding nonwords of novel symbols	Explicit verbal	Decoding nonwords / Word attack / WRMT-R-NU	Word identification /WRMT-R NU
Compton, D. L., Fuchs, D., Fuchs, L. S., Bouton, B., Gilbert, J. K., Barquero, L. A., Cho, E., Crouch, R.C. (2010)	United States	140TD 214AR	Not stated	G1	53.45%	ML English	Mixed TD & AR	Cross-Sectional Longitudinal	Decoding nonwords	Explicit verbal	Word reading fluency / Researcher developed tool	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	Mixed ML English Norwegian Swedish BL not specified	TD	Cross-Sectional Longitudinal	PA phoneme identification real words	Explicit verbal	Composite: Phoneme blending, matching, elision, rhyme / Researcher developed tool	Composite: Sight word efficiency & Phonemic Decoding efficiency / TOWRE
Cunningham, A. & Carroll, J.	England	45	5;0 K 5.2 youn	K G1	Not stated	ML English	TD	Cross-sectional Longitud	PA phoneme	Not specified	Phoneme deletion /PAT	Word reading / BAS 2

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

(2011) * **			g G1 5;11 old G1					inal	segmentation nonwords			
Edwards, A. (2020)	United States	312	Not stated	G1	Not stated	ML English	TD	Cross-Sectional	Decoding nonwords	Explicit verbal	Decoding nonwords / Word attack / WJ-III	-
Gan, Y., Zhang, J., Kahrabi-Yamato, L., Su, Y., Zhang J., Jiang, Y., Hui, Y., & Li, H. (2022)	China	135	6; 8 males 7;2 femal es	G1	54%	ML Mandarin	TD	Cross-Sectional Longitudinal	Decoding compound ideographic characters	Explicit verbal	Character recognition / Researcher developed tool	Character recognition/ Researcher developed tool
Gellert, A.S. & Elbro, C. (2017a)	Denmark	89 TD 82 AR	6;11	K	Not stated	Mixed ML Danish-110 BL not specified- 61	Mixed TD & AR	Cross-Sectional Longitudinal	Decoding nonwords of novel symbols	Explicit verbal	Word reading / Researcher developed tool	Word reading / Researcher developed tool
Gellert, A.S. & Elbro, C. (2017b)	Denmark	84 TD 76 AR	6;4	K	48%	Mixed ML Danish -101 BL not specified- 59	Mixed TD & AR	Cross-Sectional Longitudinal	PA phoneme identification real words	Explicit verbal	Phoneme identification / Researcher developed tool	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	Phoneme deletion/ Researcher developed tool	-
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	Mixed TD & AR	Cross-Sectional	Decoding words	Explicit visual and auditory	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	Mixed TD & AR	Cross-Sectional Longitudinal	Decoding nonwords of novel symbols	Train: Explicit verbal Test: None	Read nonsense syllables / Researcher developed tool	Word reading accuracy and speed / SLRT-II

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Horbach, J. Weber, K. Opolony, F. Scharke, W. Radach, R. Heim, S. & Günther, T. (2018)	Germany	17	5;0	Pre-K	41%	Mixed ML German-6 BL not specified 11	TD	Cross-Sectional Longitudinal	SSK learn novel symbol-syllables	Train: Explicit verbal Test: None	Letter name and sound knowledge / Researcher developed tool	Word reading accuracy and speed / SLRT-II
Law, J.M., De Vos, A., Vanderauwera, J., Wouters, J. Ghesquière, P/ & Vandermosten, M. (2018)	Belgium	64 TD 20Dx.	8;3	G3	Not stated	ML Dutch	Mixed	Cross-Sectional	SSK learn novel symbol-sounds	Implicit	Orthographic knowledge judgement task / Researcher developed tool	-
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	Orthographic knowledge task/ 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	Phoneme elision / TOPSS	-
Lu, Y. & Hu, C (2019)	Taiwan	50	Not stated	G4	Not stated	BL Mandarin/ English	TD	Cross-Sectional	PA phoneme segmentation nonwords	Explicit verbal and visual	Phoneme segmentation / Researcher developed tool	-
Osa Fuentes, P.M. (2003)	Spain	164	5;6	Infantil (Pre-K)	45.7%	ML Spanish	TD	Longitudinal	PA syllable segmentation and matching	Explicit verbal	-	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	-	Word identification / WRMT-RNU
Petersen, D.B, Allen, M.M. &	United States	280	Not stated	K	Not stated	Mixed ML	Mixed TD & AR	Longitudinal	Decoding nonword	Explicit verbal	-	Sight Word Efficienc

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Spencer, T.D. (2016)						English BL Spanish/ English			s			y / TOWRE
Sittner Bridges, M. & Catts, H.W. (2011) Sample 1	United States	90	Not stated	K	Not stated	ML English	TD	Cross-Sectional Longitudinal	PA syllable and phoneme deletion	Explicit verbal	Syllable and phoneme deletion / SDT	Word identification / WRMT-RNU
Sittner Bridges, M. & Catts, H.W. (2011) Sample 2	United States	96	Not stated	K	Not stated	ML English	Mixed TD & AR	Cross-Sectional Longitudinal	PA syllable and phoneme deletion	Explicit verbal	Phoneme identification / Initial sound fluency / DIBELS	Word Identification / WRMT-NU
Spector, J. (1992)	United States	38	5;11	K	Not stated	ML English	AR	Cross-Sectional Longitudinal	PA phoneme segmentation	Explicit verbal	Phoneme segmentation / Yopp Singer task	Word recognition / San Diego Quick Assessment List
Teeuwen, E. (2020)	The Netherlands	284	5;5	K	52%	ML Dutch	Mixed TD & AR	Cross-Sectional	SSK learn novel symbol-sounds	Implicit	Letter name knowledge / Researcher developed tool	-
Wyman Chin, K.R. (2018)	United States	9	Range 3;2 – 7;6	Not reported	66%	BL Spanish/ English	TD	Cross-Sectional	PA syllable discrimination nonwords	Explicit verbal and visual	Phoneme elision / TOPSS	-
Yap, D.F-F. (2018)	Singapore	99	5;3	K	43.4%	BL English/ Chinese (75%) -Malay (12%) -Other (13%)	TD	Cross-Sectional Longitudinal	PA phoneme segmentation	Explicit verbal	Composite: Phoneme elision, blending, matching /CTOPP	Letter-Word identification / WJ-III
Zumeta, R.O. (2010)	United States	37	6;1	K	51.4%	Mixed ML English BL not specified	Mixed TD & AR	Cross-sectional	PA phoneme segmentation	Explicit verbal	Composite: Phoneme elision, blending, matching/CTOPP	-

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

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, WR = word reading, OM=Outcome Measure, BAS=British Ability Scales, CTOPP= Comprehensive test of phonological processing, DIBELS = Dynamic indicators of basic early literacy skills, PAT=Phonological awareness test, SDT= Static Deletion Task, SLRT-III=Salzburger lese und rechtschreibtest, TALE = Test de Análisis de la Lecto-escritura, TOPPS = Test of Phonological Processing, 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,

* 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

Quality Appraisal

All 32 articles were appraised using the modified Johanna Briggs Institute quality appraisal tool (Moola et al., 2020). The average rating was 10.6/12 or 88%. The most common reason for deduction of points was inadequate description of participants (either based on age, sex, language or reading status). No studies were rated as low quality (0-33%), 1 article (Teeuwen, 2020), was rated as medium quality (33-67%), and the remaining 31 articles were rated as high quality (68-100%). No studies were excluded from analysis based on their quality appraisal rating. See Appendix 2 Table 3 for quality appraisal ratings of included articles.

Research Question 1A: Do dynamic assessments of word reading skills (phonological awareness (PA), sound-symbol knowledge (SSK), and decoding), demonstrate concurrent validity with static assessments of word reading skills (PA, SSK, decoding) across all populations?

As Table 5 indicates, 28 articles including 29 studies reported correlations between a DA and an SA of an equivalent construct. Thirteen studies examined PA, 5 examined SSK, and 10 examined decoding. Correlations representing the relationship between DAs and equivalent SAs of word reading skills are reported in Table 5 below.

Table 5.

Effect Sizes Representing Concurrent Validity Between Dynamic Assessments and Static Assessments of Word Reading Skills

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Study	N	Effect sizes between DAs and SAs of Phonological Awareness	Effect sizes between DAs and SAs of Sound-Symbol Knowledge	Effect sizes between DAs and SAs of decoding
Aravena, S., Tijms, J. Snellings, P. & van der Molen, M.W. (2013)	42			.52***
Aravena, S., Tijms, J. Snellings, P. & van der Molen, M.W. (2018)	71			.237*
Barker, R.M. & Saunders, K.J. (2020)	27		.63***	
Catts, H.W., Nielsen, D.C., Bridges, M.S., Liu, Y.S. & Bontempo, D.E. (2015)	313	.507**		
Cho, E., Compton, D.L, Fuchs, D., Fuchs, L.S. & Bouton B. (2014)	134			-.69*
Cho, E. & Compton, D.L. (2015)	112			.64*
Cho, E., Compton, D.L., Gilbert, J., Steacy, L.M., Collins, A.A. & Lindström, E.R. (2017)	105			-.55*
Compton, D.L., Fuchs, D., Fuchs, L.S., Bouton, B. Gilbert, J.K., Barquero, L.A., Cho, E. & Crouch, R.C. (2010)	355			-.59***
Coventry, W.L., Byrne, B., Olson, R.K., Corley, R. & Samuelsson, S. (2011)	1988	.550***		
Cunningham, A. & Carroll, J. (2011)[a]	45	.73**		
Edwards, A. 2020	312			-.53*
Gan, Y., Zhang, J., Kahrabi-Yamato, L., Su, Y., Zhang J., Jiang, Y., Hui, Y., & Li, H. (2022)	135			.68**
Gellert, A.S. & Elbro, C. (2017a)	171			.65**
Gellert, A.S. & Elbro, C. (2017b)	160	.90**		
Gillam, S.L., Fargo, J., Foley, B. & Olszewski, A. (2011)	64	-.84***		
Hautala, J., Heikkilä, R., Nieminen, L., Rantanen, V. Latvala, J-M. & Richardson,U. (2020)	723			.49***

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Horbach, J., Scharke, W., Cröll, J, Heim, S. & Günther, T. (2015)	243			.212
Horbach, J, Weber, K., Opolony, F., Scharke, W., Radach, R., Heim, S. & Günther, T. (2018)	17		.179	
Law, J.M., De Vos, A., Vanderauwera, J., Wouters, J., Ghesquiére, P. & Vandermosten, M. (2018)	84		.125	
Liu, C., Hoa Chung, L.C., Wang, L.C. & Liu, D. (2021)	203		.09	
Loreti, B. (2015)	10	.87**		
Lu, Y-Y. & Hu, C-F. (2019)	50	.76***		
Sittner Bridges, M. & Catts, H.W. (2011)	90	.840*		
Sample 1 [b]				
Sittner Bridges, M. & Catts, H.W. (2011)	96	.591*		
Sample 2 [c]				
Spector, J. (1992)	38	.43**		
Teeuwen, E. (2020)	284		.572***	
Wyman Chin, K.R. (2018)	9	.49		
Yap, D.F-F. (2018)	99	.82***		
Zumeta, R.O. (2010)	37	.63**		

Note. DA= dynamic assessment, SA=static assessment, 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

Meta-analysis of Concurrent Effect Sizes

The effect sizes from these twenty-nine studies in Table 5 were included in the analyses. Results of the random-effects meta-analysis examining the concurrent validity between DAs and SAs of word

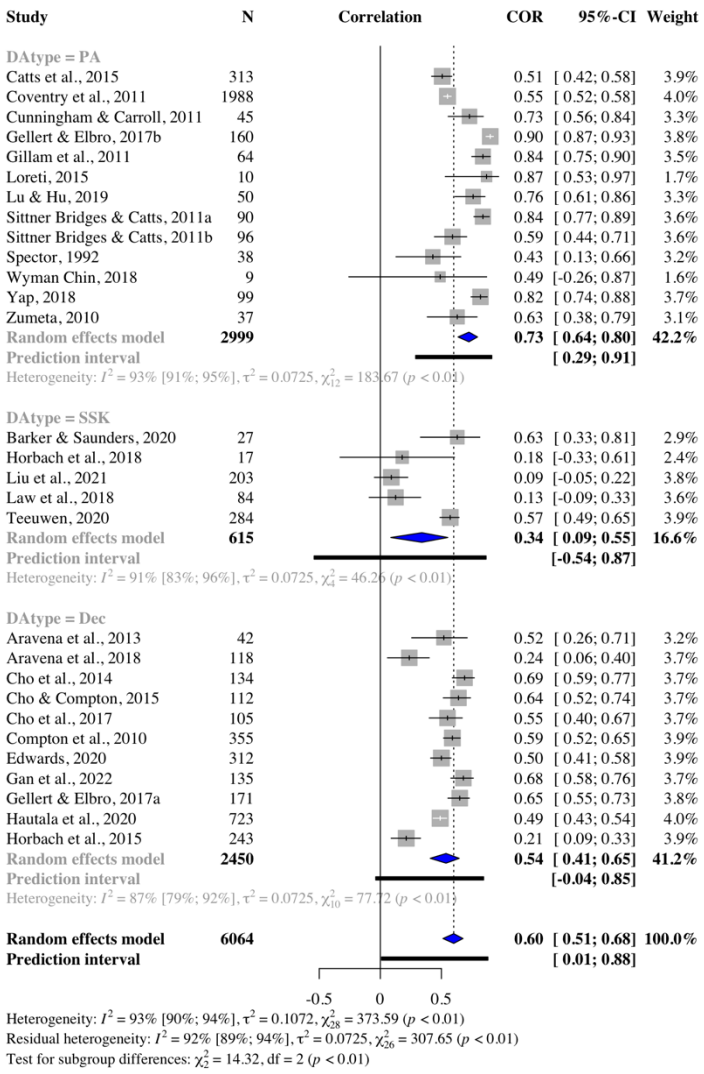
medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

reading skills are reported in Figure 4. The overall mean effect size is large (r=0.60, 95%CI = [0.51-0.68]) suggesting that overall DAs are strongly correlated with SAs. Given the significant heterogeneity, the moderator of DA type was included and found to be significant (Q=14.32, df=2, p<0.01). Results of the subgroup analysis are also displayed in Figure 4. Outcomes of this mixed effects model suggest that there is a significant variation between subgroups of DA type (PA, SSK, and decoding) and their concurrent validity with SAs of equivalent constructs. Overall mean effect sizes representing the relationship between DAs of PA and decoding, and their SA counterparts are strong, while the effect size for DAs of SSK is moderate. Furthermore, the prediction intervals for the SSK and decoding subgroups cross 0 indicating that future relevant studies examining relationships between SAs and DAs of SSK and decoding may demonstrate a negative correlation. Importantly, the prediction interval for the PA and decoding subgroups do not cross 0, suggesting that future studies are likely to report positive correlations between these DAs and SAs (Harrer et al., 2021; IntHout et al., 2016). Therefore, the results indicate that DAs of PA skills are most likely to be valid alternatives to traditional SAs, followed by DAs of decoding, which demonstrate strong concurrent correlations a slightly larger prediction interval, and finally DAs of SSK which are associated with weaker overall mean effect sizes and the largest prediction interval. However, it is also important to note that the test for within group heterogeneity in the mixed effects model was still found to be significant, even with DA type as a moderator (Q=307.65, df=25=6, p<0.01), which indicates that there are likely other moderators beyond DA type that are impacting heterogeneity.

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

Figure 4.

Forest Plot of Random Effects Meta-Analysis Examining the Concurrent Validity Between Dynamic and Static Assessments of Word Reading Skills



Note. Study names=Study, sample size =N, effect sizes =COR, and 95% confidence intervals =CI (95%) are reported as well as the type of DA = Dynamic assessment, phonological awareness = PA, sound-symbol knowledge = SSK or Decoding. 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 bars represent the prediction intervals. Figure drawn in R Studio using `metacor` package (Laliberté, 2019; R Core Team, 2021; Laliberté).

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Research Question 1B: Do dynamic assessments of word reading skills demonstrate predictive validity with reading outcome measures (single word reading) across all populations?

Sixteen articles including 18 studies reported correlations between a DA and a later word reading outcome measure. Eight examined PA, 1 examined SSK, and 9 examined decoding. Table 6 below shows correlations between each of the three DA skills and word reading outcomes.

Table 6.

Effect Sizes Representing Predictive Validity Between Dynamic Assessments of Word Reading Skills and Word-reading Outcome Measures

Study	N	Effect sizes between DAS of Phonological Awareness and WR outcomes	Effect sizes between DAs of Sound-Symbol Knowledge and WR outcomes	Effect sizes between DAs of decoding and WR outcomes
Caffrey, E. (2006 sample 1)	25			-.624**
Caffrey, E. (2006 sample 2)	95			-.745**
Cho, E., Compton, D.L., Gilbert, J., Steacy, L.M., Collins, A.A. & Lindström, E.R. (2017)	105			-.46*
Compton, D.L., Fuchs, D., Fuchs, L.S., Bouton, B. Gilbert, J.K., Barquero, L.A., Cho, E. & Crouch, R.C. (2010)	355			-.69***
Coventry, W.L., Byrne, B., Olson, R.K., Corley, R. & Samuelsson, S. (2011)	1988	.423***		
Cunningham, A. & Carroll, J. (2011) [a]	45	.77**		
Gan, Y., Zhang, J., Kharabi-Yamato, L., Su, Y., Zhang J., Jiang, Y., Hui, Y., & Li, H. (2022)	135			.70**
Gellert, A.S. & Elbro, C. (2017a)	171			.66**
Gellert, A.S. & Elbro, C. (2017b)	160	.47**		
Horbach, J, Scharke, W., Cröll, J, Heim, S. & Günther, T. (2015)	243			.258*
Horbach, J, Weber, K., Opolony, F., Scharke, W., Radach, R., Heim, S. &	17		.855**	

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Günther, T. (2018)			
Osa Fuentes, P.M. (2003)	164	.36**	
Petersen, D.B., Gillam, R.B. (2015)	63		.51**
Petersen, D.B. Allen, M.M., Spencer, T.D. (2016)	280		.57**
Sittner Bridges, M. & Catts, H.W. (2011)	90	.516*	
Sample 1 [b]			
Sittner Bridges, M. & Catts, H.W. (2011)	96	.426*	
Sample 2 [c]			
Spector, J. (1992)	38	.60**	
Yap, D.F-F. (2018)	99	.60***	

Note. DA= dynamic assessment, WR= word reading, PA = phonological awareness, SSK = sound-symbol knowledge, Dec = decoding, 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

Meta-Analysis of Predictive Effect Sizes

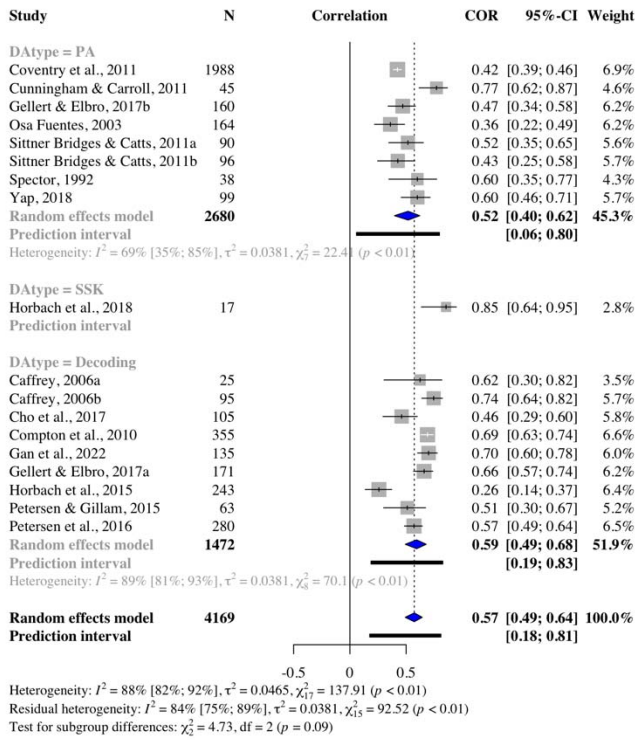
Results of the random-effects meta-analyses examining the predictive validity of DAs of word reading skills with later word reading outcome measure are presented Figure 5. The overall mean effect size is large (r=0.57, 95%CI = [0.49-0.64]) suggesting that DAs are strongly correlated with word reading outcome measures. Given the significant heterogeneity, the moderator of DA type was included but not found to be significant (Q=4.73, df=2, p=0.09). The results of this mixed effects model, also displayed in Figure 5, indicate that there is not significant variation between subgroups of DA type (PA and decoding) and their predictive validity with tests of word reading outcomes as determined by

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

correlation coefficients. The subgroup of SSK could not be compared due to a lack of studies. The overall prediction interval and those representing the relationship between DAs of PA and decoding and later word reading outcomes do not cross 0, indicating that future studies are likely to indicate positive correlations between DAs of PA and decoding and later word reading outcomes (Harrer et al., 2021; IntHout et al., 2016). The results of this analysis provide suggestive evidence for the predictive validity of DAs of PA and decoding with word reading outcomes. However, again it should be noted that the test for withing group heterogeneity in the mixed effects model was still found to be significant, even with DA type as a moderator (Q=92.52, df=15, p<0.01), which indicates that there are likely other moderators beyond DA type that are impacting heterogeneity.

Figure 5.

Forest Plot of Random Effects Meta-Analysis Examining the Predictive Validity of Dynamic Assessments of Word Reading Skills with Word Reading Outcome Measures



medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

Note. Study names=Study, sample size =N, effect sizes =COR, and 95% confidence intervals =CI (95%) are reported as well as the type of DA = Dynamic assessment, phonological awareness = PA, sound-symbol knowledge = SSK or Decoding. 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 bars represent the prediction intervals. Figure drawn in R using `metacor` package (Laliberté, 2019; R Core Team, 2021).

Research Question 2A: Do dynamic assessments of word reading skills demonstrate concurrent validity with static assessments of word reading skills, and predictive validity with word reading outcome measures within population groups defined by their language status (monolingual vs. bilingual)?

Concurrent validity –Subgroup analysis language status

Of the 29 studies that reported a correlation coefficient between a DA and an SA of an equivalent construct 20 studies included exclusively monolinguals, 4 studies included exclusively bilinguals, and 5 studies included mixed language status populations in their analyses. Table 2 presents further study details. A subgroup analysis by language status was planned a priori. Results are reported below in Table 7. The outcomes of this mixed effects model suggests that there is significant variation between subgroups defined by language status (monolingual, bilingual or mixed mono and bilingual populations) in terms of the concurrent validity of DAs with their equivalent construct SA (Q=6.54, df=2, p=0.04). The effect sizes were largest for bilingual groups, followed by mixed language groups and monolingual groups. Prediction intervals crossed 0 for all three subgroups in this analysis. This outcome runs counter to what was expected. Given that SAs are typically designed for monolinguals, we hypothesized that the overall mean effect size between DAs and SAs would be larger for this group, while we might observe smaller effect sizes between the two for bilinguals who are prone to underperform on SAs. It is possible that this outcome is a result of other factors that cannot be accounted for in a simple subgroup analysis. For example, the limited number of studies in the BL group (n=4)

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

included only typically developing children, while the many studies in the ML group (n=20) included typically developing and at-risk children, as well as those diagnosed with dyslexia. It is possible that this, or other unknown factors, are contributing to this difference.

Table 7.

Results of Subgroup Analyses by Language Status for Concurrent Validity

	Included studies	*g*	95%CI	*I2*	*I2* 95%CI	p value heterogeneity	Prediction interval	p value subgroup
Language Status								0.04
Monolingual	20	0.54	0.44-0.63	0.90	0.87-0.93	<0.01	-0.04 - 0.85	
Bilingual	4	0.78	0.58-0.89	0.03	0.0-0.85	0.38	-0.46 - 0.99	
Mixed	5	0.72	0.53-0.84	0.97	0.95-0.98	<0.01	-0.51-0.98	

Predictive validity- Subgroup analysis language status

Of the 18 studies that reported a correlation coefficient between a DA and a later word reading outcome measure, 11 included exclusively monolinguals, 2 studies included exclusively bilinguals, and 5 studies included mixed mono and bilingual populations. Table 2 lists further study details. Subgroup analysis by language status was planned a priori and results are reported in Table 8. The outcomes of this mixed effects model suggest that there is no significant variation between subgroups defined by language status (monolingual, bilingual or mixed) in terms of the predictive validity of DAs with later word reading outcome measures (Q=0.03, df=2, p=0.98). However, the prediction interval could not be calculated for the bilingual group due to limited number of studies and crossed 0 for mixed language groups but not the monolingual group, suggesting that future relevant studies may be most likely to document positive correlations between DAs and word reading outcomes for monolinguals.

Table 8.

Results of Subgroup Analysis by Language Status for Predictive Validity

	Included studies	*g*	95%CI	*I2*	*I2* 95%CI	p value heterogeneity	Prediction interval	p value subgroup
Language Status								0.98
Monolingual	11	0.57	0.46-0.66	0.89	0.82-0.93	<0.01	0.10-0.76	

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Bilingual	2	0.56	0.27-0.76	0	-	0.43	-
Mixed	5	0.58	0.42-0.71	0.88	0.75-0.94	<0.01	-0.15-0.90

Research Question 2B: Do dynamic assessments of word reading skills demonstrate concurrent validity with static assessments of word reading skills, and predictive validity with word reading outcome measures within population groups defined by their reading status (typically developing vs. at-risk)?

Concurrent validity –Subgroup analysis reading status

Of the 29 studies that reported a correlation coefficient between a DA and an SA of an equivalent construct 14 included exclusively participants who were typically developing (TD) in their reading abilities 4 included participants who were exclusively at-risk or diagnosed with reading difficulty and 11 included mixed reading status populations in their analyses. Refer to Table 2 for study details. Subgroup analysis by reading status was planned a priori and results are reported in Table 9. The outcomes of this mixed effects model suggest that there is no significant variation between subgroups defined by reading status (typically developing, at-risk or mixed) in terms of the concurrent validity of DAs with their equivalent construct SA counterparts (Q=3.41, df=2, p=0.18). However, the prediction interval crossed 0 for both the at-risk and mixed language groups, but not the typically developing group, suggesting that future relevant studies may be most likely to document positive correlations between DAs and SAs for TD children.

Table 9.

Results of Subgroup Analysis by Reading Status for Concurrent Validity

	Included studies	*g*	95%CI	*I2*	*I2* 95%CI	p value heterogeneity	Prediction interval	p value subgroup
Reading Status								0.18
Typically developing	14	0.68	0.56-0.77	0.92	0.89-0.95	<0.01	0.09-0.92	
At-risk	4	0.49	0.19-0.70	0.87	0.68-0.95	<0.01	-0.78-0.97	
Mixed	11	0.54	0.40-0.67	0.94	0.91-0.96	<0.01	-0.13-0.88	

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Predictive validity – Subgroup analysis reading status

Of the 18 studies that reported a correlation coefficient between a DA and a later word reading outcome measures, 9 studies included exclusively participants who were typically developing, 2 studies included participants who were at-risk or diagnosed with reading difficulty and 7 included populations with mixed reading abilities in their analyses. Please see Table 2 for study details. Subgroup analysis by reading status was planned a priori and results are reported in Table 10. The outcomes of this mixed effects model suggest that there is no significant variation between subgroups defined by reading status (typically developing, at-risk or mixed) in terms of the predictive validity of DAs with later word reading outcome measures (Q=1.35, df=2, p=0.51). Furthermore, the prediction interval did not cross 0 for any subgroup suggesting that future relevant studies are likely to document positive correlations between DAs word reading outcomes for all subgroups regardless of reading status.

Table 10.

Results of Subgroup Analysis by Reading Status for Predictive Validity

	Included studies	*g*	95%CI	*I2*	*I2* 95%CI	p value heterogeneity	Prediction interval	p value subgroup
Reading Status								0.51
Typically developing	9	0.62	0.50-0.72	0.89	0.81-.94	<0.01	0.15-0.86	
At-risk	2	0.55	0.25-0.76	0	-	0.54	-	
Mixed	7	0.53	0.40-0.64	0.88	0.79-0.93	<0.01	0.08-0.80	

Risk of Publication Bias

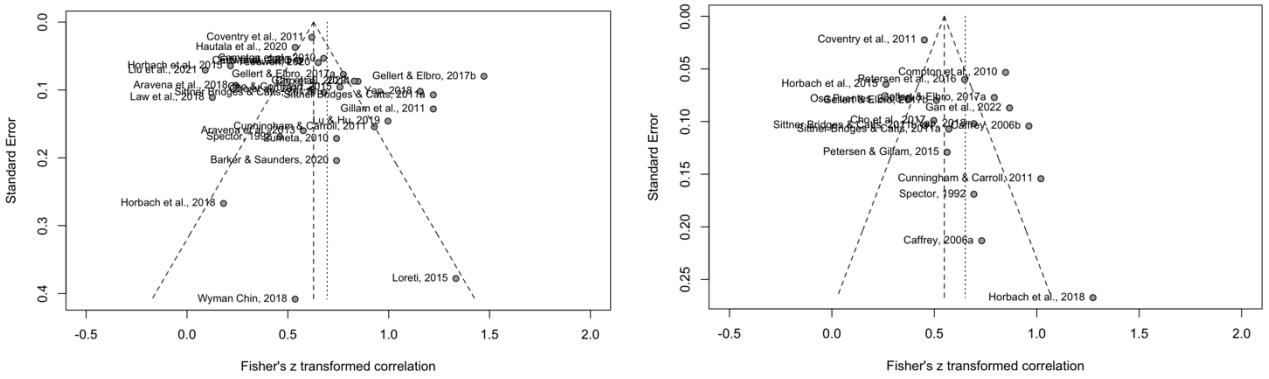
Two additional analyses were completed to examine potential publication bias for both the concurrent and predictive validity analyses. First, funnel plots were generated. Visual inspection of the funnel plot for the concurrent validity analysis does not suggest asymmetry (Figure 10). Second, Egger's test was calculated and not found to be significant (Intercept = 1.282, 95%CI= [-1.04 - 3.61], p=.289). Therefore, we conclude that there is no indication of funnel plot asymmetry and minimal risk of publication bias in the analysis for concurrent validity. However, visual inspection of the funnel plot for the predictive validity analysis (Figure 11) suggests potential asymmetry, and Egger’s test was

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

significant for the presence of plot asymmetry (Intercept = 2.4, 95%CI [0.37 - 4.43-, p=.034). Inspection of the plot reveals that there are several small studies with positive effects included in the analysis (e.g., Caffrey, 2006a; Spector, 1992; Horbach et al., 2018), but an absence of smaller studies with negative effects. This suggests that there is a possibility that small studies with negative effects either were not written up, published, or identified in the grey literature search (Lee & Hotopf, 2012).

Figures 10 & 11.

Funnel Plots of Studies Included in the Meta-analysis of the Concurrent Validity Between Dynamic and Static Assessments of Word Reading Skills (Left) And Predictive Validity Between Dynamic Assessments of Word Reading Skills and Word Reading Outcome Measures (Right)



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).

Discussion

Based on 34 studies from 32 articles, we conducted a systematic review and meta-analysis of the concurrent and predictive validity of dynamic assessments (DA) of word reading skills, including phonological awareness (PA), sound-symbol knowledge (SSK) and decoding.

Our meta-analysis examining the *concurrent validity* of DAs with their equivalent static assessment (SA) counterparts suggests that there is a strong correlation between the two types of tests (r=.60). However, subgroup analysis suggests that there are differences in effect sizes for DAs of different literacy skills: DA-PAs demonstrated the strongest correlation and (r=.73) with their SA

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

counterparts and narrowest positive prediction interval, followed by DA-Dec (r=.54) and DA-SSKs (r=.34). There are two possible reasons DAs of PA might demonstrate the strongest concurrent validity with their SA counterparts. First, PA is a well-defined construct with an established hierarchy of skills that is consistent across languages (Anthony & Francis, 2005). Independent of orthography, PA tasks range in complexity of manipulation from simple (e.g., blending) to complex (e.g., manipulation) and in size of unit of speech from larger (e.g., syllables) to smaller (e.g., phonemes). In the included articles, there was a great degree of consistency between the tasks researchers use to evaluate PA dynamically and statically across studies (e.g., a DA of phoneme segmentation compared an SA of phoneme segmentation). Conversely, DAs of SSK and decoding were characterized by a greater degree of variability in task type between DA and SA (e.g., a DA of SSK of learning novel-symbol sound correspondences compared to an SA of letter naming knowledge). Additionally, DAs of PA are likely to correlate more strongly with SAs because they were almost exclusively conducted in-person. This is relevant because most SAs were also conducted in-person and consistency in administration may produce stronger correlations between the two. DAs of SSK, however, were often administered via a computer program and compared to an in-person SA which could have resulted in weaker correlations between the two (e.g., Barker & Saunders, 2020; Horbach et al., 2018).

A second meta-analysis found that that DAs of word reading skills have strong predictive validity with word reading outcome measures (OMs) (r=.57). These findings are consistent with Caffrey et al.,'s 2008 review which documented overall strong effect sizes between DAs and outcome measures across domains. Our findings corroborate Caffrey's but provide more specific insight into the use of DAs of distinct word reading skills. In the subgroup analysis by DA type, DA-Dec demonstrated the strongest correlation with later word reading outcome measures (r=.59), but this overall weighted effect size was not significantly different from that of the DA-PA with word reading (r=.52). In terms of DA-

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

SSK, only one study was included and could therefore not be compared in subgroup analysis. These findings, which provide suggestive evidence for the validity of DAs of PA and decoding in predicting later word reading outcomes are in line with outcomes of Dixon et al.'s systematic review which found that both DAs of PA and decoding accounted for significant unique variance in later word reading ability (between 1-33% across included studies; Dixon et al., 2022a).

It is unsurprising that DA-Dec showed the strongest correlation in this analysis. A DA that evaluates the ability to learn how to decode words should demonstrate strong predictive validity with later word reading outcomes because these two constructs are similar. It is worth noting that this strong correlation was documented even with variation in the types of DA-Dec and SA-Dec tasks used. For example, some studies used novel symbol nonword decoding task in the DA, and alphabetic real word reading task as an OM and still reported strong correlations between the two (e.g., Gellert & Elbro 2017). These findings complement results from a second systematic review from Dixon et al., which found that DAs targeting constructs that more closely resemble reading ability demonstrated higher classification accuracy in determining reading disorder than those that were more distal (e.g., DAs of decoding better predicted reading status than DAs of morphological awareness).

Significant residual heterogeneity was observed in both the concurrent and predictive validity analyses, even after including the moderator of DA type. Subgroup analyses examining the role of language and reading status were planned a priori to further examine heterogeneity. Regarding *language status* only 9 of 34 studies included bilinguals in their sample, and of those 9, only 4 conducted separate analyses of bilinguals to allow for comparison with monolinguals, while 5 grouped mono and bilinguals together in their analyses. In terms of concurrent validity, subgroup analysis by language status results suggest significant differences in mean effect sizes between groups, with studies conducted with bilinguals demonstrating the strongest correlations between DAs and SAs (r=.78), followed by mixed

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

populations (r=.72) and monolinguals (r=.54). However, these results should be interpreted with caution for two reasons. First there were a limited number of studies included in each of the mixed and bilingual subgroups (n=5 and n=4 respectively) relative to the monolingual group (n=20), a finding which is consistent with previous DA reviews (Dixon et al., 2022a, 2022b). Secondly, the bilingual children in all 4 studies were all typically developing, while the monolingual children varied in terms of their reading status across the 20 studies in this subgroup. These differences could have implications for strength of correlations between DAs and SAs. Participant characteristics cannot be separated from each other in a subgroup analysis but are practically important. A different pattern emerged in the predictive validity analysis by language subgroup, with no significant differences observed between subgroups, suggesting for DAs of word reading demonstrate consistent predictive validity with later reading outcomes across groups regardless of their language status. It should be noted that the nature of bilingualism was often poorly defined in included studies (e.g., sequential vs. simultaneous) and little to no information was provided about the languages spoken, the age of acquisition or proficiency levels. A notable exception is Petersen and Gillam (2015) who evaluated their Spanish/English bilinguals using the Bilingual English Spanish Oral Screener (BESOS; Peña et al., 2009) and reported the number of participants who were English or Spanish dominant versus balanced bilinguals, as well as the average years of expressive language experience in English, Spanish, or bilingual settings.

Subgroup analyses were also conducted to examine whether mean effects sizes differed given participant *reading status* (typically developing vs. at-risk readers). Across the reviewed studies, reading status factored into research questions much more frequently than language status and generally researchers described how at-risk status was defined in their studies. However, only 5 studies conducted separate correlational analyses with at-risk groups or children diagnosed with reading difficulty. An additional 12 studies included mixed groups of participants, and 17 were conducted exclusively with

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

typically developing children. Results of the subgroup analysis for concurrent and predictive validity did not suggest any significant differences in mean effect sizes for groups stratified by their reading status.

The results of the subgroup analyses for predictive validity of DAs stratified by language and reading status differ from findings documented in previous reviews. Specifically, Caffrey et al., (2008) reported that mean effect sizes for DAs and outcome measures (across domains) were strongest for populations with disabilities (r=0.59), followed by typically developing children (r=0.42) and those who were at-risk (r=0.37), while we documented strongest effect sizes for typically developing children (r=0.62), then at-risk groups (r=0.55), then mixed ability groups (r=0.53). This difference may be attributed to the fact that we included only studies examining DAs of word reading skills rather than DAs across cognitive domains or could be a result of separating at-risk and bilingual populations, who were grouped together in the previous review.

Limitations

Despite a comprehensive database and grey literature search, it is possible that relevant articles were not identified. It is possible that relevant articles were not included because of their language of publication. We included articles published in English, French and Spanish, because these were the languages that we were able to read and extract data from. However, the preprint repository search produced several articles in Portuguese, Korean and Mandarin, that may have also been relevant. Future reviews conducted in the field of bilingual literacy would benefit from purposeful inclusion or cross-cultural and linguistic collaborations with team members who speak and can read other languages so that studies published in non-Western European languages are included. Also, despite the suggestion by Dixon et al., (2022) to consider including PAL in future systematic reviews of DA, our team elected to not include this term in this review primarily because PAL tasks are inherently dynamic in nature.

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

Finally, we chose correlation coefficients as our measure of effect size to represent the concurrent validity between DAs and SAs and predictive validity between DAs and OMs. Like in the Caffrey et al., review (2008), this choice was made because correlation coefficients were the most observed effect size across studies and allowed for inclusion of a greater number of studies. However, because of this choice, the results can only provide insight into the associations between DAs and SAs or OMs. Though other measures of effect size, like regression coefficients, are better suited to determine causality, the variety in statistical analyses, choice of predictor and outcome variables and study design factors made conducting a regression meta-analysis infeasible.

Conclusions and Clinical Implications

Despite these limitations, the results of this systematic review and meta-analysis provide suggestive evidence for the concurrent and predictive validity of DAs of phonological awareness (PA) and decoding. DAs of PA and decoding demonstrated strong overall mean effect sizes with SAs of equivalent constructs and with later word reading outcomes. In contrast, there is less evidence to support the validity of DAs of SSK, given that the overall effect size representing the concurrent validity of DAs of SSK with equivalent SAs was only moderate, and there were insufficient studies to conduct a subgroup analysis for the predictive validity of DAs of SSK with later word reading outcomes. Results were mixed regarding the concurrent validity of DAs with SAs across population subgroups, but outcomes indicate that DAs demonstrate strong predictive validity with word reading outcomes across populations, regardless of their language (monolingual/bilingual) or reading status (typically developing/at-risk).

Given these findings, it may be useful for clinicians to incorporate DAs of PA and decoding into their practice. For bilingual children, for whom there are limited word reading screening tools available, DAs may be a suitable alternative to SAs for evaluating and predicting future literacy skills. For

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

monolingual children, clinicians may consider using a DA in addition to an SA. Although this requires additional time, inclusion of a DA in an early literacy assessment battery permits comparison between the outcomes of the two measures. This comparison may help differentiate those who are truly struggling or at-risk for later difficulty from those who lack the skills to perform on the SA, and may be worth the extra time required to administer the DA.

Future Directions

Future research should examine the validity and use of DAs of word reading skills with well-defined bilingual populations and children who are at-risk for or diagnosed with reading difficulties. DA is often purported to be a less biased method for evaluating the abilities of linguistically diverse children, and as a useful tool for differentiating between those who are truly at-risk for difficulty and those who perform poorly as a result of lack exposure to literacy experiences prior to schooling (e.g., Prath, 2020; Saenz & Huer, 2003). However, this review identified that the majority of studies examining DA to date have been conducted with monolingual children (n=24/34), who are typically-developing (n=17/34).

It will also be critical that these future studies adequately describe the characteristics of their bilingual populations. We noted in this review that while researchers typically described how at-risk status was defined in their articles, even in the rare instances where bilingual children were included in studies, the nature of their bilingualism was poorly defined or not described at all. In the future, researchers should include information about which languages are spoken by bilinguals, report whether these languages were learned simultaneously or sequentially and provide some metric of their proficiency in each language. Without this information, results cannot be meaningfully interpreted for use with any bilingual group.

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Beyond language status, it may also be valuable for future studies to examine how other demographic variables contribute to performance on both DAs. For instance, previous studies have indicated that factors like sex, race, and socioeconomic status (SES) contribute to performance on traditional SAs of word reading skills. In the studies included in this review, sex distributions were reported for most studies, but data on participants racial identities and SES backgrounds was limited. It is worth investigating whether DAs have the potential to reduce sex and gender or racial bias in addition to linguistic bias. To achieve this, researchers must consider these intersecting factors in their research design and methodology.

Author Notes

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

Acknowledgements

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

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

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

References

Anthony, J. L., & Francis, D. J. (2005). Development of phonological awareness. *Current directions in psychological Science*, *14*(5), 255-259. DOI: 10.1111/j.0963-7214.2005.00376.x

**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. DOI: 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. DOI: 10.1177/0022219417715407

Arias, G., & Friberg, J. (2017). Bilingual language assessment: contemporary versus recommended practice in American schools. *Language Speech and Hearing Services in Schools*, *1(*48), 1-15. DOI: 10.1044/2016_LSHSS-15-0090. PMID: 27788525.

Aurini, J., & Davies, S. (2021). COVID□19 school closures and educational achievement gaps in Canada: Lessons from Ontario summer learning research. *Canadian Review of Sociology/Revue Canadienne de sociologies*, *58*(2), 165-185. DOI: 10.1111/cars.12334

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

**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. DOI: 10.1080/07434618.2020.1737965

Barwick, M. A., & Siegel, L. S. (1996). Learning difficulties in adolescent clients of a shelter for runaway and homeless street youths. *Journal of research on adolescence*, *6*(4), 649-670.

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. DOI: 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. DOI: 10.2167/beb392.0

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

Borenstein, M., Higgins, J. P. T., Hedges, L. V., & Rothstein, H.R. (2017). Basics of meta-analysis: I2 Is not an absolute measure of heterogeneity. *Research Synthesis Methods*, *8*(1): 5-18. DOI: 10.1002/jrsm.1230

** 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).

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

**Caffrey, 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.

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.

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. DOI: 10.1177/0022219408326219

**Catts, H. W., Nielsen, D. C., Bridges, M. S., Liu, Y. S., & Bontempo, D. E. (2015). Early identification of reading disabilities within an RTI framework. *Journal of learning disabilities, 48*(3), 281-297. DOI: 10.1177/0022219413498115

**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. DOI: 10.1177/0022219412466703

**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/ DOI: 10.1016/j.lindif.2014. DOI: 10.004

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

**Cho, 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. DOI: 10.1177/0022219415599343

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement, 20*(1), 37-46.

**Compton, D. L., Fuchs, D., Fuchs, L. S., Bouton, B., Gilbert, J. K., Barquero, L. A., Cho, E., & 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. DOI: 10.1037/a0018448

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

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

** 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 word reading and phoneme awareness. *Journal of Experimental Child Psychology*, *109*(2), 248-255. DOI: 10.1016/j.jecp.20 DOI: 10.12.005

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

Daniel, S. S., Walsh, A. K., Goldston, D. B., Arnold, E. M., Reboussin, B. A., & Wood, F. B. (2006). Suicidality, school dropout, and reading problems among adolescents. *Journal of learning disabilities*, *39*(6), 507-514. DOI: 10.1177/00222194060390060301

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. DOI: 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*, DOI: 10.1177/00222194221117510

D'Souza, C., Kay-Raining bird, E., & Deacon, H. (2012). Survey of Canadian speech-language pathology service delivery to linguistically diverse clients. *Canadian Journal of Speech-Language Pathology and Audiology*, *36*(1), 18-39.

**Edwards, A. (2020). *Predictor Importance in Future and Concurrent Predictions of Oral Reading Fluency*. (Course report, Florida State University).

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. DOI: 10.1136/bmj.315.7109.629

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

**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

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Chinese children from grades 1-2. *Scientific Studies of Reading*, 1-17. DOI: 10.1080/10888438.2022.2143271

**Gellert, A. S., & Elbro, C. (2017). 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. DOI: 10.1177/0022219415609185

**Gellert, A. S., & Elbro, C. (2017). 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.

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

**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. DOI: 10.1016/j.jcomdis.20 DOI: 10.11.003

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

Grosjean, F. (2010). Bilingual. In *Bilingual*. Harvard university press.

Harrer, M., Cuijpers, P., Furukawa, T. A., & Ebert, D. D. (2021). Doing meta-analysis with R: A hands on guide. Boca Raton, FL, and London: Chapmann & Hall/CRC Press.

**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.

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta□analysis. *Statistics in medicine*, *21*(11), 1539-1558. DOI: 10.1002/sim.1186

Hogan, T. P., Catts, H. W., & Little, T. D. (2005). The relationship between phonological awareness and reading. *Language Speech and Hearing Services in Schools*, *36*(4), 285-293.

Høien, T., Lundberg, I., Stanovich, K. E., & Bjaalid, I. K. (1995). Components of phonological awareness. *Reading and writing*, *7*(2), 171-188.

**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. DOI: 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*, 1716. DOI: 10.3389/fpsyg.2018.01716

IntHout, J., Ioannidis, J. P. A., Rovers, M. M., & Goeman, J. J. (2016). Plea for routinely presenting prediction intervals in meta-analysis. *British Medical Journal Open 6*(7). http://dx.doi.org/ DOI: 10.1136/bmjopen-2015-010247

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

**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*, *1393*. DOI: 10.3389/fpsyg.2018.01393

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

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)*.

**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. DOI: 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).

**Lu, Y. Y., & Hu, C. F. (2019). Dynamic assessment of phonological awareness in young foreign language learners: Predictability and modifiability. *Reading and Writing, 32*(4), 891-908.

Lundberg, I. (1994). Reading difficulties can be predicted and prevented: A Scandinavian perspective on phonological awareness and reading. In C. Hulme & M. Snowling (Eds.), *Reading development and dyslexia* (pp. 180–199). Whurr Publishers.

Martin, N., Brownell, R., & Hamaguchi, P. (2018). *Test of auditory processing skills (TAPS-4).* Pro-Ed.

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

Montoya, S. (2018). "Defining literacy." In *GAML Fifth meeting*, pp. 17-18. 2018.

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, Munn Z (Editors). *JBI Manual for Evidence Synthesis.* JBI, 2020. DOI: https://doi.org/ DOI: 10.46658/JBIMES-20-08

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

Moretti, G. A. S., & Frandell, T. (2013). *Literacy from a right to education perspective.* United Nations Educational, Scientific and Cultural Organization (UNESCO). https://unesdoc.unesco.org/ark:/48223/pf0000221427

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

**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* DOI: 2021;372:n71. doi: 10.1136/bmj.n71

Peña, E. D., Bedore, L. M., Guitiérrez-Clellen, V. F., Iglesias, A., & Goldstein, B. (2009). *Bilingual English Spanish Oral Screener.* Unpublished manuscript.

**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. DOI: 10.1177/0022219414538518

**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. DOI: 10.1177/0022219413486930

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

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

Prath, S. (2020, June 8). *Dynamic Assessment: What we need to know*. Bilinguistics. https://bilinguistics.com/dynamic-assessment/

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

Ritchie, S. J., & Bates, T. C. (2013). Enduring links from childhood mathematics and reading achievement to adult socioeconomic status. *Psychological science, 24*(7), 1301-1308. DOI: 10.1177/0956797612466268

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

Saenz, T. I., & Huer, M. B. (2003). Testing strategies involving least biased language assessment of bilingual children. Communication Disorders Quarterly, 24(4), 184-193.

Scarborough, H. S. (1998). Predicting the future achievement of second graders with reading disabilities: contributions of phonemic awareness, verbal memory, rapid naming, and IQ. Ann. Dyslexia 48, 115–136. DOI: 10.1007/s11881-998-0006-5

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). New York, NY: Guilford Press.

Shelley-Tremblay, J., O'Brien, N., & Langhinrichsen-Rohling, J. (2007). Reading disability in adjudicated youth: Prevalence rates, current models, traditional and innovative treatments. *Aggression and Violent Behavior, 12*(3), 376-392. DOI: 10.1016/j.avb.2006.07.003

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Silver, N. C., & Dunlap, W. P. (1987). Averaging correlation coefficients: should Fisher's z transformation be used? *Journal of applied psychology, 72*(1), 146.

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.

**Spector, J. E. (1992). Predicting progress in beginning reading: Dynamic assessment of phonemic awareness. *Journal of educational psychology, 84*(3), 353.

**Teeuwen, E. (2020). Assessment of letter− speech sound learning in children with familial risk for dyslexia in kindergarten. (Bachelor's thesis, University of Amsterdam).

United Nations Educational, Scientific and Cultural Organization. (2017). *More than half of children and youth worldwide 'not learning'*-UNESCO. https://www.un.org/sustainabledevelopment/blog/2017/09/more-than-half-of-children-and-youth-worldwide-not-learning-unesco/

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

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

Wolf, M. S., Gazmararian, J. A., & Baker, D. W. (2005). Health literacy and functional health status among older adults. *Archives of internal medicine*, *165*(17), 1946-1952. DOI: 10.1001/archinte.165.17.1946

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

Wood, E., & 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. Retrieved from osf.io/bcghx

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

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

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

Appendix 1.

Figure 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. OR (comput* adapt* test*).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. OR (comput* adapt* test*).tw,kf.	*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 OR (comput* adapt* test*).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 OR Comput* adapt * test*

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Figure 2.

Search terms for concept 2 – Literacy

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

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

			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))	Spell OR Spelling OR Decode OR decoding

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

Appendix 2.

Table 3.

Quality Appraisal of Included Studies

Study	Participants	Dynamic Assessment		Static Assessment &/ WR Outcome Measures		Flow and Timing		Statistical Analyses	Overall Score and /12 & %	Overall Appraisal
	Were participant characteristics adequately described? (1)	*Was the DA adequately described? (2)*	*Was the test appropriate for evaluating early literacy skills? (2)*	*Was the SA or WR outcome measure adequately described? (2)*	*Was the SA or OM valid and reliable for evaluating early literacy skills?* (2)	*Were all participants included in all 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	10(83%)	High
Caffrey, E. (2006)	1	2	2	2	2	0	1	1	11(92%)	High
Catts, H.W., Nielsen, D.C., Sittner Bridges, M., Liu, Y. S., & Bontempo, D.E. (2015)	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,	1	2	2	2	2	0	1	1	11(92%)	High

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

A.A. & Lindstrom, E.R. (2017)										
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
Gan, Y., Zhang, J., Kharabi-Yamato, L., Su, Y., Zhang J., Jiang, Y., Hui, Y., & Li, H. (2022)	1	1	2	2	2	1	1	1	12(100%)	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. & Olszewski A. (2011)	1	2	2	2	2	1	1	1	12(100%)	High
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

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

Horbach, J., 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
Lu, Y. & Hu, C (2019)	0	2	2	2	2	1	1	1	11(92%)	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
Teeuwen, E. (2010)	1	2	2	2	0	0	1	0	8(67%)	Medium
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
Zumeta, R.O. (2010)	1	2	2	2	2	1	1	1	12(100%)	High

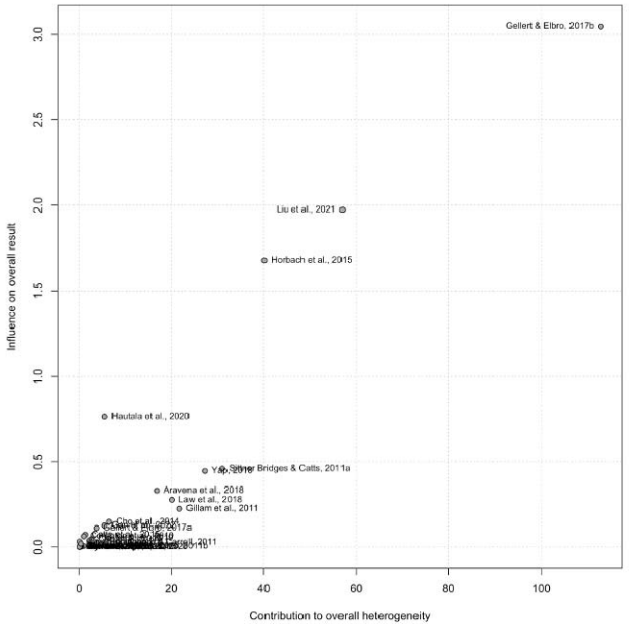
Note. DA= Dynamic assessment, SA= Static assessment, WR = word reading, Low = 0-33%, Medium=34-66%, High=67-100%

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Appendix 3.

Figure 12.

Baujat plot for studies included in the meta-analysis of the concurrent validity of dynamic and static assessments of early literacy skills



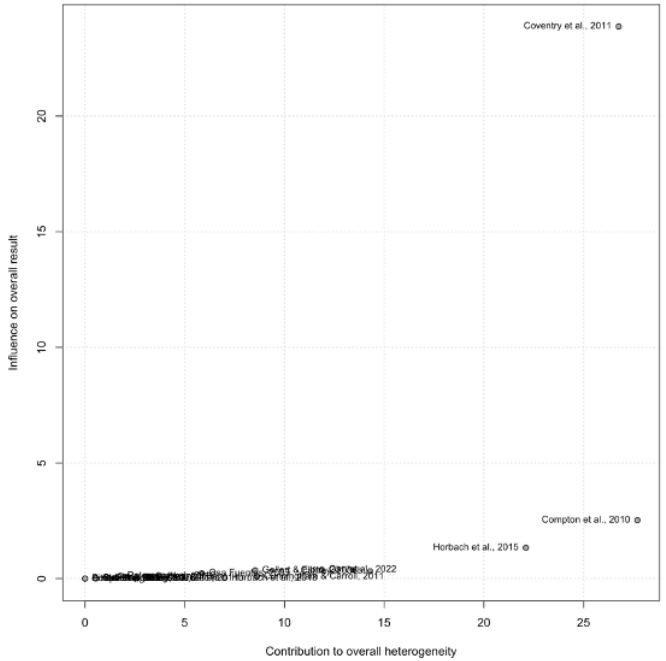
Note. In the Baujat plot, individual contribution to overall heterogeneity is represented on the horizontal axis, and the influence on overall result on the vertical axis. Studies with greatest influence are found in the top right quadrant of the figure. Drawn in R using the 'metacor'

package (R Core Team, 2021; Laliberté, 2019).

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Figure 13.

Baujat plot for studies included in the meta-analysis of the predictive validity of dynamic assessments of early literacy skills with 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).

medRxiv preprint doi: https://doi.org/10.1101/2022.09.19.22279942; this version posted January 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .