Concurrent and predictive validity of dynamic assessments of early literacy skills: A systematic review and meta-analysis

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

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

[2] Rehabilitation Sciences Institute, University of Toronto

Author Note

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

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

The authors declare no conflicts of interest.

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

Email: e.wood@utoronto.ca

Abstract

Traditional static literacy assessments evaluate acquired knowledge and are prone to floor effects. These tools are also developed almost 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 potential alternative, and more equitable approach to evaluating critical early literacy skills of phonological awareness, sound-symbols knowledge, and decoding. This systematic review and meta-analysis examined the concurrent validity of DAs of early literacy with their static equivalents, and their predictive validity longitudinally with later word reading outcome measures both overall across all populations, and specifically with bilingual

and at-risk groups. Thirty studies were identified through searching 5 databases, and the grey literature. Included studies provided a correlation between a dynamic and concurrent static assessment, or a dynamic and a later reading outcome measure. Results of the first random effects meta-analysis suggested that overall, there was a strong relationship between dynamic and static assessments (r=.58). Subgroup analysis revealed that there were significant differences (p=.0012) between DAs of distinct early literacy skills, with decoding (r=.72) and phonological awareness (r=.50) measures demonstrating greater degrees of correlation with their static counterparts, compared to DAs of sound-symbol knowledge (r=.34). The outcomes of the second random effects meta-analysis indicate that there is a similarly strong relationship between DAs and word reading outcome measures overall (r=.58). Subgroup analyses did not reveal significant differences (p=.0593) in the predictive association between DAs of phonological awareness (r=.55) and decoding (r=.58). There were insufficient studies to conduct separate analyses for bilingual and at-risk populations. However, a narrative review suggests that the magnitude of the effect sizes from individual studies conducted with these populations are in line with overall correlational findings. There is also some evidence to suggest that DAs have the capacity to explain a significant amount of variance in later word reading outcomes in bilingual (7-11%) and at-risk groups (7-21%). Future studies should examine the validity of DAs specifically for use with well-defined bilingual and at-risk groups, as these are the populations who potentially have the most to gain from these measures.

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

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 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.

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 this statistic has only worsened as a result of the COVID-19 pandemic. School closures, and remote learning during the formative years of early literacy are expected to result in a loss of 67% in literacy skills for nearly 1.5 billion kindergarten-aged students around the world and 5.5. million students in Canada (UNESCO, 2021; Aurini & Davies, 2021). The resulting educational gap is cause for great concern. Inadequate literacy skills are associated with notable negative life outcomes, such as poor physical and mental health (Wolf et al., 2005; Kim et al., 2014; Daniel et al., 2006), reduced academic attainment (Ritchie & Bates, 2013), restricted socioeconomic mobility, and increased rates of poverty, homelessness, and incarceration (Barwick & Siegel, 1996; Shelley-Tremblay et al., 2007). Research has established that early identification and intervention are key in mitigating longstanding literacy difficulties and their associated adverse effects (Lundberg, 1994; OHRC, 2022).

Traditional Static Literacy Assessment

A range of traditional standardized assessment tools (e.g., CTOPP-2, Wagner et al, 2013; PAT-2:NU, Robertson & Salter, 2017; TAPS-4, Martin et al., 2018) are commonly used by clinicians, educators, and researchers. These so-called “static assessments” (SAs) attempt to 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 problems are associated with these traditional SAs. First, the majority 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 that mask a child's capacity to perform the skill. For example, if a bilingual kindergarten-aged child who speaks Tamil at home and has just started learning English in school is administered a traditional, static English test evaluating their knowledge of letter sounds, they will perform much more poorly than a child who has grown up speaking exclusively English. This is not because they lack the ability to learn letter sound relationships, but because they have had significantly less exposure to English letters and their sounds than their English monolingual peers, and may even be a result of their not understanding the test's English instructions. These language effects associated with SAs often result in misidentification of literacy difficulties in bilingual populations (Bedore & Peña, 2008; Petersen & Gillam, 2015).

The second, related issue associated with these traditional SAs is floor effects (Catts et al., 2009). Floor effects are commonly observed in traditional assessment of early literacy because these tools attempt to quantify a child's acquired ability in an area with which they have limited to no experience when they begin school. For example, many kindergarten-aged children would be likely to perform poorly on a SA of word reading ability at the start of the year, simply because they have had limited experience with reading. When most children who take a test perform poorly,

testers are unable to differentiate those who truly are at risk for 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 an at-risk child with few literacy experiences, by suggesting that their lack of knowledge is predictive of their future ability. Early and accurate identification facilitates access to intervention, which can prevent the later negative effects associated with reading difficulties. When a test identifies most students as at-risk, it becomes impossible to direct the limited available resources to those who truly need them to prevent these adverse effects.

Dynamic Assessment as an Alternative

Dynamic assessment (DA) is an alternative to this traditional SA paradigm. While SA endeavors to measure a child's acquired knowledge via a neutral, uninvolved examiner's administration of test items and collection of binary correct/incorrect responses, DAs attempt to measure the ability to learn through an interactive approach to testing that uses training, scaffolding, prompting and feedback. This testing method much more closely resembles real world teaching and learning experiences and rejects the false dichotomy between assessment and intervention, conceptualizing the two as existing along a single continuum rather than as two separate processes (Grigorenko & Sternberg, 1998; Poehner, 2008). DA has been used in various domains, yet, still lacks wide usage and uptake because its administration methods and approach to grading tend to be less structured and unstandardized, making it difficult to quantify the validity and reliability of various DAs, especially in settings that rely on generalizable percentiles when evaluating early literacy (Poehner, 2008; Orellana, 2019). Despite this critique, DA is a potential alternative to SA for equitably assessing bilingual and at-risk populations, who are often

disadvantaged by traditional assessment formats (Vellutino et al., 1998; Orellana et al., 2019). This is because evaluating the ability to learn a skill, like reading, through a process that more closely resembles classroom instruction may allow typically developing bilingual children to overcome the language effects of traditional SA testing and demonstrate their true potential while also allowing for accurate identification of those who are truly at-risk based on poor performance on DAs even with testing, feedback, and prompts.

Previous Systematic Reviews in Dynamic Assessment

Given these promising findings, interest in DA has been growing. The use of DA has been evaluated by several systematic reviews and meta-analyses. Caffrey et al. (2008) reported that DAs were slightly more consistent in predicting outcomes than SAs across the domains of literacy, cognition, language, and mathematics. This review offers support for the use of DAs across domains but does not provide a specific synthesis on its validity in relation to evaluation of early literacy skills. Additionally, the authors merged at-risk and bilingual children together in their analysis when stratifying studies across four population groups. Bilingualism is not inherently a risk-factor, and these two populations should therefore be considered separately to provide a more accurate understanding of the validity of DAs for use with these groups.

More recently, two systematic reviews from 2019 and 2022 examined the use of DAs for the diagnosis of developmental language disorder (DLD) in bilingual children (Orellana et al., 2019; Hunt et al., 2022). Orellana et al. (2019) synthesized sensitivity and specificity values from six articles to determine the diagnostic accuracy of these tools. The authors found suggestive evidence supporting the use of DAs to identify DLD in these groups. However, only peer-reviewed primary studies published in English, whose participants were bilingual English-other speakers, were included in this review. By disregarding the grey literature, the risk of publication bias was

increased and by excluding studies published in other languages and with other bilingual profiles the authors perpetuate the notion that English is the normative language for study and address only the utility of DA for Spanish/English bilinguals (Thornton, 2000; Hamel, 2007). In the 2022 article, Hunt et al., reviewed 10 studies; 9 of which claim to have successfully identified DLD in children under the age of 12. Like Orellana et al. (2019), Hunt et al. (2022) focused specifically on tools used in the diagnosis of oral language difficulties and did not include DAs evaluating literacy skills. Although Hunt et al. (2022) findings support DA as a reliable measure for identifying DLD, the small number of studies and sample sizes reveal an area of research that requires attention and the need for more focus on multilingual populations.

Dixon et al. (2022a; 2022b) published two papers exploring 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 the first study (2022a) the researchers identified 17 studies that used DA to measure reading skills between at least two timepoints. Overall, they found that DAs of phonological awareness, decoding, and morphological awareness can account for variance in the growth of different outcome measures (OMs), ranging from 1-33%. The second study yielded 14 papers examining the utility of DA to classify reading disorder. Results of this review indicate that DA can account for unique variance in predicting later reading disorder, particularly when the test construct is similar to reading (e.g., decoding vs. working memory) and when predicting abilities proximally vs distally (e.g., in early vs later school years). The authors also report that in some instances DA demonstrated adequate classification accuracy independently, but in others this accuracy was enhanced by concurrent use of SAs (2022b). While the first review provides support for the use of DA in predicting future reading abilities its' scope is limited to evaluating DA's potential for accounting for growth in reading

skills beyond SA through hierarchical regression. The second paper similarly only included articles that addressed diagnostic accuracy of DAs through use of sensitivity and specificity measures, while excluding papers that examine use of DA for other purposes using varying statistical techniques. Because of the heterogeneity in the predictor and outcomes variables used in the regression analyses in review 1, and the variability in the diagnostic measures and definitions of reading disorder used in study 2, authors were not able to conduct anything syntheses of their findings and instead provided a narrative review only. Finally, much like Orellana et al. (2019), in both instances Dixon et al. (2022a; 2022b) only included peer-reviewed English journal articles.

The current study

This systematic review and meta-analysis will address the gaps discussed above. First, we will specifically examine the concurrent validity of DAs of early literacy skills with their equivalent SAs, and the predictive validity of DAs with word reading outcomes. The construct of early literacy in this review was informed by the subskills that comprise word recognition ability in the evidence-based reading model - Scarborough's Reading Rope (2001). Scarborough states that three components are essential for word recognition or word reading ability; (i) phonological awareness (PA)– the ability to identify and manipulate parts of speech, (ii) 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) decoding – the ability to apply PA and SSK skills to sound out real or made-up words. Regarding the objective examining predictive validity of DA as it relates to reading outcomes, reading in this review was defined as single word or nonword reading. Secondly, the authors will specifically investigate whether DAs of early literacy demonstrate consistent validity across population groups, including not only typically developing monolinguals, but also bilingual children and those at-risk for reading difficulties. While Orellana

(2019) and Hunt (2022) examined the diagnostic accuracy of DA for identifying DLD in bilinguals, their work did not examine the validity of DAs of early literacy and did consider use of DA for at-risk children. Furthermore, this review differs from that of Dixon et al. (2022) in that it is not restricted to analyzing studies that determine DA's ability to uniquely predict variance in growth of reading skills beyond SAs. Investigating the general predictive and concurrent validity of DA in early reading skills gives more insight on whether DA is a potential substitute to SA. This insight is integral in developing novel literacy tools that optimize resources and help remedy the literacy crisis our global community faces. Also, due to our broader range of focus, we were able to analyze validity as measured by correlation coefficients, which permitted inclusion of studies not identified by Dixon, yielding sufficient studies with common effect sizes to facilitate conducting a meta-analysis. Finally, to reduce potential publication bias, we expanded our search to include unpublished work. Also, to avoid perpetuating an English monolingual focus, we included articles written in other languages (French and Spanish).

Research Questions

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

1. B) Do dynamic assessments of early literacy skills demonstrate concurrent validity with static assessments of early literacy skills within population groups defined by their language (monolingual vs. bilingual) or reading status (typically developing vs. at-risk or diagnosed with difficulty)?

2. A) Do dynamic assessments of early literacy skills demonstrate predictive validity with reading outcome measures (single word reading) across all populations?

2. B) Do dynamic assessments of early literacy skills demonstrate predictive validity with reading outcome measures (single word reading) within population groups defined by their language or reading status?

Method

Eligibility Criteria

Study inclusion criteria were decided upon in advance and outlined in the systematic review and meta-analysis screening protocol available on Open Science Framework (Wood & Molnar, 2021):

(i) Only research articles that are primary in nature were included. Case reports, commentaries and editorials were excluded, as well as systematic reviews and books or book chapters. Articles published in peer-reviewed journals and unpublished grey literature from preprint repositories and reports or dissertations from google scholar searches were also included.

(ii) 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. 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 early literacy skills and (i) a SA of early literacy skills at the same time point, AND/OR (ii) an outcome measure of reading ability at a later

timepoint. Included studies reported correlation coefficients to quantify relationships between DA and SA early literacy measures and/or DA early literacy and 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, but only works published or generated prior to January 31st, 2022, were included in this review.

Search Strategy and Information Sources

The initial search was carried out 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. Following the initial database search, 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 full list of search terms used in each database see 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 a research assistant 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 a research assistant 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 that conduct research in the field of early literacy.

At each stage, all studies were rated by two trained research assistants. Disagreements were resolved by the first author. A team of ten reviewers participated in the article identification process, and as a result, 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 or practical. 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

Total number of participants (after attrition in the case of longitudinal studies), the percentage of males vs. females in the sample 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. This information was essential to compare the concurrent and predictive validity of DAs according to population type (e.g., do DAs demonstrate consistent predictive validity for both monolinguals and bilinguals, or for typically developing vs. at-risk

groups?) This was particularly relevant for studies that included children diagnosed with reading disorders in terms of how the diagnosis was made, or for studies that included bilingual children, to inform if and how bilingualism was characterized in the sample.

Measures

To verify the concurrent validity of DAs of early literacy (PA, SSK, Decoding) with equivalent SAs of early literacy (PA, SSK, Decoding), and the predictive validity of distinct types of DAs of early literacy (PA, SSK, Decoding) with reading outcome measures, information about the characteristics of each assessment was extracted.

Dynamic Assessment(s) Coders reported the name of the DA if one was provided, the early literacy 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 early literacy 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. SAs in this review are tests that are conducted at the same time point (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., The Comprehensive Test of Phonological Processing-2), 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 names of any assessments used (e.g., CTOPP-2), the early literacy 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-III) and the specific subtests of the measure (e.g., Word Attack). Coders also indicated the reading skills evaluated by these measures (e.g., word reading accuracy, nonword reading accuracy, word reading fluency etc.).

[1]Coders also reported whether the DA was administered in the traditional "in-person" format, or via computer and which type of DA was used, the Train/Test or Graduated Prompt format. Results of the comparisons between in-person and computerized DA and train/test vs. Graduate prompts DA will be reported in a separate forthcoming study regarding the relationship between the different administration methods and formats of DAs of early literacy skills and word reading measures.

Effect Sizes

The correlation coefficients representing the relationships between the DAs and SAs, and/or the DAs and OMs were extracted from correlational matrices to conduct the correlational meta-analyses examining concurrent and predictive validity, respectively. 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 (e.g., a DA of PA and a word reading fluency and nonword reading accuracy OM) coders were instructed to extract all relevant correlations coefficients at this stage. Following review of all extracted coefficients, the first, second and last author 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 primarily made based on which measure was most consistently used across studies. For example, word reading accuracy was reported in every 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 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 the following 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.

Following selection, the coefficient representing the effect size was input into a Table for presentation of findings from single studies and an excel document in preparation for synthesis.

Quality Appraisal Assessment

All included studies were assessed by two trained independent reviewers on their quality 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.

Regarding participants, coders rated whether the participant sample was adequately described in terms of age, gender breakdown, language and reading status and demographic characteristics. Coders evaluated the dynamic assessment domain by rating whether the DA was described with sufficient detail in terms of the skills evaluated, the format of the test, the prompting and scoring used, and administration process. Coders also recorded whether the task used to evaluate the early literacy skill(s) in the DA was developmentally appropriate for the population. When assessing the domain related to the reference standards (SAs and OMs) coders evaluated whether the studies employed developmentally appropriate tools for evaluating early literacy skills and reading and writing outcome measures. They also rated whether psychometric properties of reliability and validity of the reference measures used were reported. In evaluating flow and timing, coders rated whether all participants were included in the analyses and noted whether authors explained and accounted for reasons for loss to follow up and attrition if this occurred. 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 (34-66%) or high quality (67-100%). Only studies rated as medium and high quality were included in the analyses.

Analyses

A random effects model was employed in the meta-analyses to account for clinical and methodological variance between studies (Borenstein et al., 2010). An excel document with the following information; each study's author, the number of participants, the Pearson correlation coefficient representing the effect size, and the type of early literacy skill evaluated by the DA was uploaded into R studio for analysis (R Core Team, 2021). First, the 'metacor' package was used to conduct a Fisher Z transformation of Pearson's correlation coefficients into Z scale scores (Laliberté, 2019). This was necessary as meta-analytic approaches assume that the sampling distribution of observed outcomes are normally distributed. However, when a study sample size is small, and the magnitude of the correlation is large, the sampling distribution is skewed (Silver & Dunlap, 1987). When correlations are transformed to a Z scale, this normalizes the distribution (Corey et al., 1998). 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. The result is a weighted correlation coefficient that represents the overall mean effect size.

Heterogeneity between studies was examined using the Q, I^2 and τ^2 statistics. The Q statistic indicates the ratio of observed variation to within study variance and is a measure of how much heterogeneity can be attributed to between study differences (Cochran, 1954). A significant Q statistic suggests that studies do not share a single common effect size (West et al., 2010). The Q statistic is sensitive to the number of studies included and their precision and should therefore not be interpreted alone (Harrer et al., 2021). I^2 represents the proportion of observed variance in effects that is attributed to actual difference between studies, rather than variance within studies or due to chance (Higgins & Thompson, 2002). An I^2 of 0-25% represents low heterogeneity, 25-50% is indicative of moderate heterogeneity, and suggests 75%-100% substantial heterogeneity (Higgins & Thompson, 2002). This measure is less sensitive to the number of included studies, but still sensitive to their precision (Borenstein et al., 2017). The τ^2 statistic is an estimate of the between study variance and is insensitive to the number and precision of included studies (Harrer et al., 2021). It is used to calculate Tau, which is an estimate of the standard deviation of true effect sizes across studies (Harrer et al., 2021). The Sidik Jonkman estimator was used to calculate Tau in these analyses (Sidik & Jonkman, 2005). Typically, all 3 heterogeneity statistics are calculated and reported to provide a robust picture of the degree heterogeneity between studies. However, these statistics are not without flaws, chief among them, that they are difficult to interpret. For this reason, we also calculated and reported prediction intervals. Prediction intervals are easier to understand, and represent the range in which we would expect an effect size from a new, relevant study to fall, based on the included, available evidence (Harrer et al., 2021; Spineli & Pandis 2020). If the prediction interval falls completely on the positive side of 0, and does not cross it, we can

expect future relevant studies to demonstrate positive effects, regardless of the varying effects (IntHout et al., 2016; Harrer et al., 2021).

Heterogeneity statistics determine whether there is significant heterogeneity between studies and quantify an estimate of how much variance can be attributed to true between study differences as opposed to sampling error (Harrer et al., 2021). Baujat plots (presented for the concurrent and predictive validity syntheses in Figure 10 and 11 in Appendix 3) even permit identification of which studies contribute most to overall heterogeneity (Baujat et al., 2002). However, these statistical analyses and plots do not shed light on why this heterogeneity might exist. Subgroup analyses, also referred to as moderator analyses, where studies are grouped based on a specific characteristic, employ a mixed model (random effects within subgroups, and fixed effects between subgroups) and allow researchers to determine whether differences in effect sizes are due to sampling error, or because of these moderators (Harrer et al., 2021). In subgroup analyses, it is theorized that the included studies come from different subgroups, and that each subgroup has a separate, overall effect (Harrer et al., 2021). Subgroup analyses are mixed effects models because they include a random effects model for within subgroups and a fixed effects model for between groups (Harrer et al., 2021). A subgroup analysis by DA type (phonological awareness, sound-symbol knowledge, and decoding) was planned a priori in this study and conducted for concurrent and predictive validity syntheses.

To examine the potential risk of publication bias, funnel plots for each analysis were generated in R studio (R Core Team, 2021). Funnel plots present effect sizes of included studies on the horizontal axis with their corresponding standard error on the vertical axis and the pooled effect size is represented by a dashed line in the center (Harrer et al., 2021). Funnel plots are inspected visually. When data points are centered symmetrically around the pooled effect line in

an upside-down funnel shape, there is minimal potential risk of publication bias (Harrer et al., 2021). However, because interpretation of these plots is subjective, an additional measure, Egger's regression test, a test of significance, was conducted to examine funnel plot asymmetry (Egger et al., 1997). A significant Egger's test indicates plot asymmetry and potential for publication bias in the meta-analysis (Harrer et al., 2021).

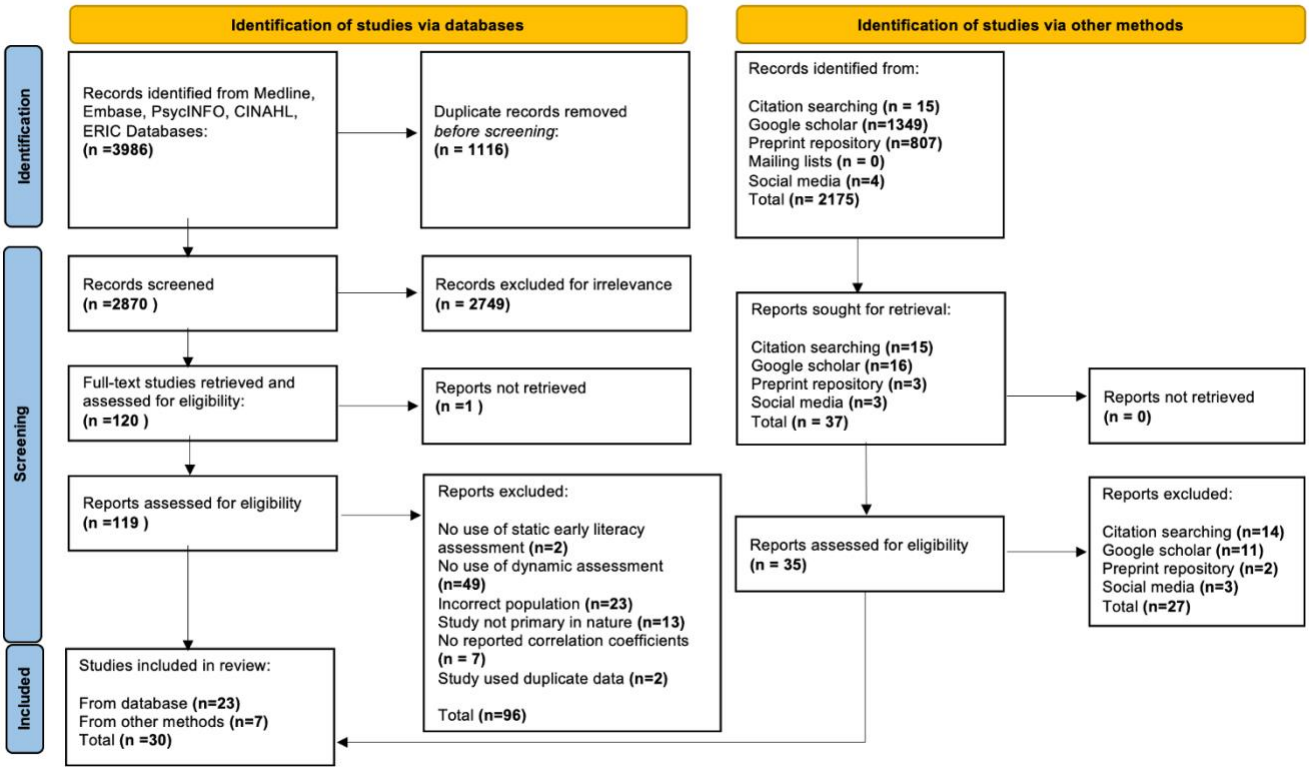
Results

Study Selection

The initial database searches produced 3,986 articles which were uploaded to Covidence. The software automatically detected and removed 1,116 duplicate articles leaving 2,870 titles and abstracts for review. Of these articles, 119 were reviewed at the level of full text, and 23 articles were identified for inclusion. Next, the 807 articles identified from the preprint repositories search 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 1,349 articles were identified via forward Google Scholar searching of the 24 already included articles. Over three rounds of iterative searching, a total of 16 relevant articles were identified, of which 11 were excluded and 5 included. Subsequently, an ancestral search of the 29 identified articles was completed. This yielded 15 potential articles which were reviewed at the level of full text. 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, 30 articles were identified for inclusion in the systematic review and meta-analysis. The process of study identification is visualized in the PRISMA flowchart below.

Figure 1.

PRISMA flowchart of literature search



Study Characteristics

See Appendix 2 for Tables 1-3, which provide the following information:

- **Table 1.** ***General Study Characteristics***
- **Table 2.** ***Characteristics of Assessments***
- **Table 3.** ***Quality Appraisal of Included Studies***

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

Correlations representing the relationship between DAs and equivalent SAs of early literacy are reported in Table 4 below.

Table 4.

Effect Sizes Representing Concurrent Validity Between Dynamic Assessments (DA) and Static Assessments (SA) of Early Literacy Skills

Study	N	DA-PA – SA-PA	DA-SSK SA-SSK	DA-Dec SA-Dec
Aravena, S., Tjims, J. Snellings, P. & van der Molen, M.W. (2013)	42			0.52***
Aravena, S., Tjims, 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			.54*
Cho, E., Compton, D.L., Gilbert, J., Steacy, L.M., Collins, A.A. & Lindström, E.R. (2017)	105			-.43*
Compton, D.L., Fuchs, D., Fuchs,	355			-.59***

L.S., Bouton, B. Gilbert, J.K., Barquero, L.A., Cho, E. & Crouch, R.C. (2010)			
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*
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., Heikkila, R., Nieminen, L., Rantanen, V. Latvala, J-M. & Richardson,U. (2020)	723		.49***
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., Ghesquiere, P. & Vandermosten, M. (2018)	84		.125
Liu, C., Hoa Chung, L.C., Wang, L.C. & Liu, D. (2021)	203		.09
Lu, Y-Y. & Hu, C-F. (2019)	50	.76***	
Sittner Bridges, M. & Catts, H.W. (2011) Sample 1 [b]	90	.840*	
Sittner Bridges, M. & Catts, H.W. (2011) Sample 2 [c]	96	.591*	
Spector, J. (1992)	38	.43**	
Teeuwen, E. (2020)	284		.572***
Yap, D.F-F. (2018)	99	.82***	
Zumeta, R.O. (2010)	37	.63**	

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

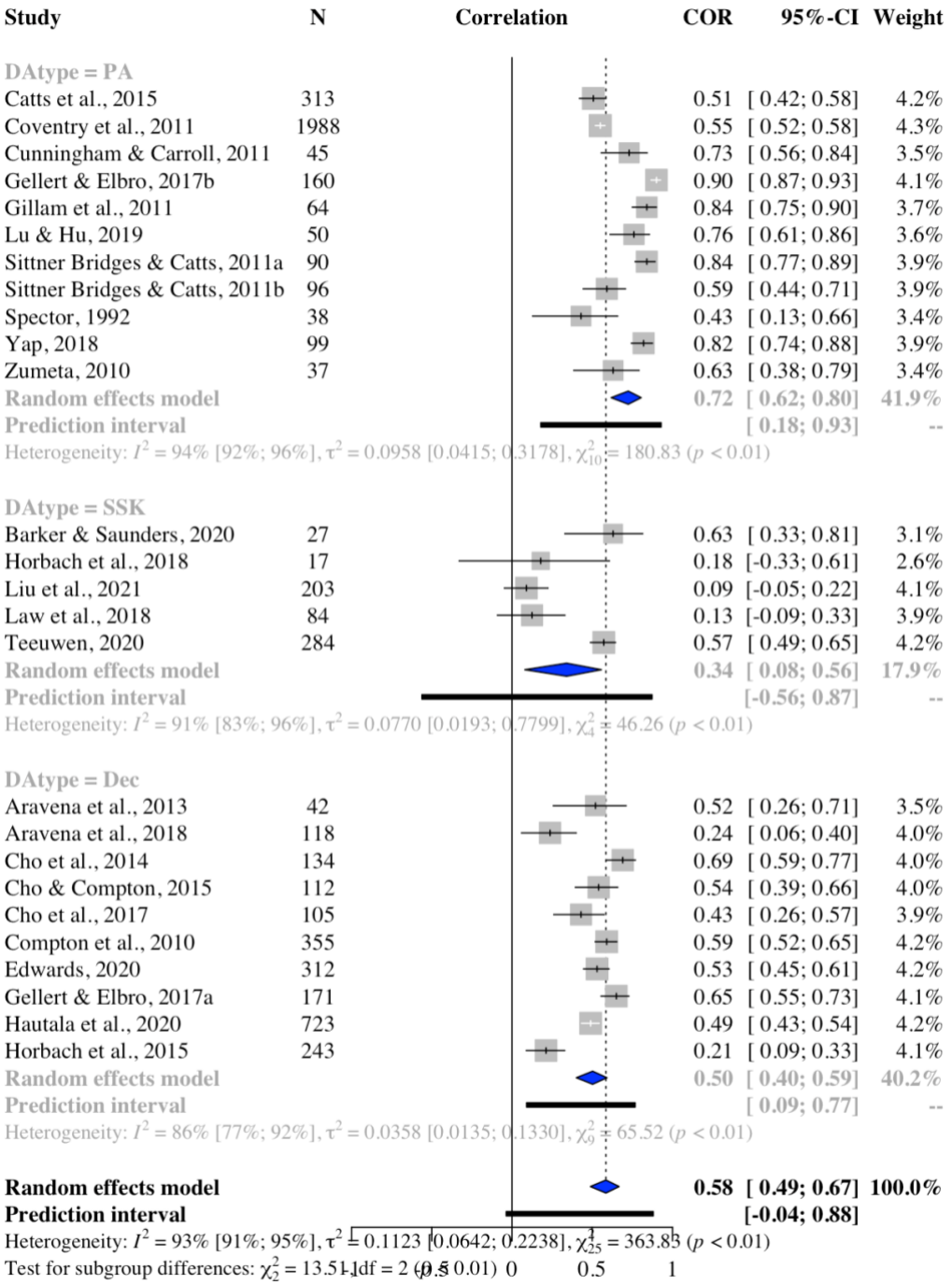
Analysis of Concurrent Effect Sizes

Overall Concurrent Validity DA-SA

Results of the random-effects meta-analysis examining the concurrent validity between DAs and SAs of early literacy are displayed in Figure 2. The random effects meta-analysis found that the overall mean effect size is large (r=.58, 95%CI = [0.49-0.67]) suggesting that DAs are strongly correlated with SAs. The prediction interval however, ranged from g=-0.04 to 0.88, crossing 0 and indicating that negative correlation effect sizes cannot be ruled out for future studies. Significant heterogeneity was found (Q=363.83, p<.01) and between study variance was estimated at $\tau^2 =$ 0.1123 (95% CI=0.0642-0.2238] with an I^2 value of 93% (95% CI= 91-95%]).

Figure 2.

Forest plot of random effects meta-analysis examining the concurrent validity between dynamic and static assessments of early literacy.



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

Concurrent Validity Subgroup Analysis

Given the significant heterogeneity, the moderator of DA type was included and found to be significant (Q=13.51, df=2, p<.005). Results of the subgroup analysis are reported in Figure 3. According to this mixed effects model, there is a significant variation between subgroups of DA type (phonological awareness, sound-symbol knowledge, and decoding) and their concurrent validity with SAs of equivalent constructs.

DA-PA concurrent validity with SA-PA

The overall mean effect size representing the concurrent validity between DAs and SAs of phonological awareness is large (r=.72, 95%CI= [0.62-0.80]). The outcomes of the analysis suggest that there is a strong correlation between DA-PA and DA-SA. The prediction interval (95%CI= [0.18-0.093]), does not cross 0 indicating that negative correlation effect sizes are unlikely in future studies.

DA-SSK concurrent validity with SA-SSK

The overall mean effect size representing the concurrent validity between DAs and SAs of sound-symbol knowledge is moderate (r=.34, 95%CI= [0.08-0.56]). The outcomes of the analysis suggest that there is a modest correlation between DA-SSK and SA-SSK. The prediction interval

crosses 0 (95%CI = [-0.56-0.87]) indicating that negative correlation effects sizes cannot be ruled out in future studies. However, this result should be interpreted with caution due to the limited number of studies in this subgroup.

DA-Dec concurrent validity with SA-Dec

The overall mean effect size is large (r=.50, 95%CI= [0.40-0.5]). The outcomes of the analysis suggest that there is a strong correlation between DA-decoding and SA-decoding. The prediction interval does not cross 0 (95%CI = [0.09-0.77]) indicating that negative correlation effect sizes are unlikely in future studies.

It should be noted 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=292.61, df=23, p<.0001), which indicates that there are likely other moderators beyond DA type that are impacting heterogeneity.

Figure 3.

Results of subgroup/moderator analyses

	g	95%CI	*p*	*I2*	95%CI	Prediction interval	*p subgroup*
DA type							.0012
Phonological Awareness (PA)	.72	0.62-0.80	<.01	0.94	0.92-0.96	0.18-0.93	
Sound-Symbol Knowledge (SSK)	.34	0.08-0.56	<.01	0.91	0.83-0.91	-0.56-0.87	
Decoding (Dec)	.50	0.40-0.59	<.01	0.86	0.77-0.92	0.09-0.77	

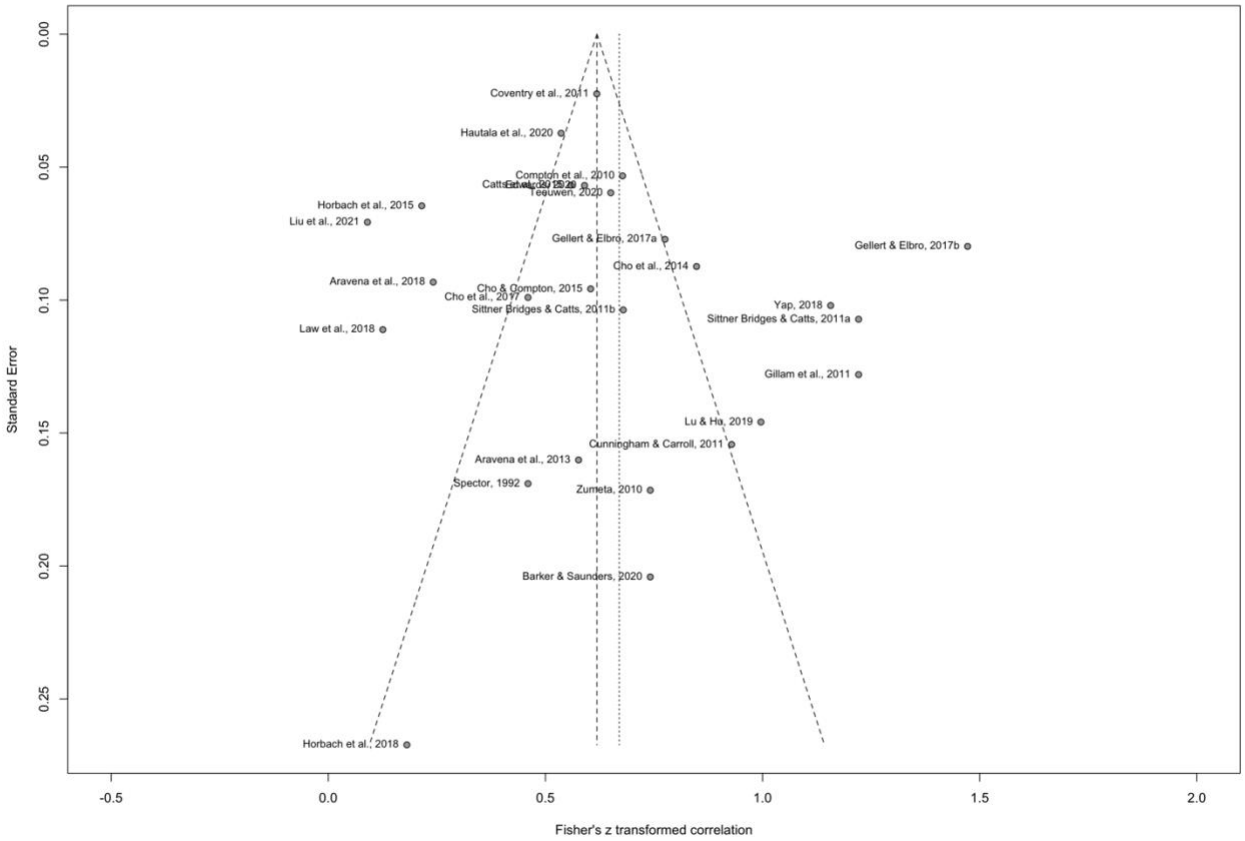
Risk of Publication Bias

Two additional analyses were conducted to examine potential publication bias. First, funnel plots were generated. Visual inspection of the funnel plot does not suggest asymmetry, Second, Egger's test was calculated and found to be not significant (Intercept = 1.084, 95%CI= [-1.61 - -

3.78], p=.437). There is no indication of funnel plot asymmetry and minimal risk of publication bias.

Figure 4.

Funnel plot of studies included in the meta-analysis of the concurrent validity between dynamic and static assessments of early literacy skills.



Note. In the funnel plot, 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).

Question 2A: Do dynamic assessments of early literacy skills demonstrate predictive validity with reading outcome measures (single word reading) across all populations?

Correlations between each of the three dynamic assessment skills of early literacy and the outcome of word reading are reported in Table 5 below.

Table 5.

Effect Sizes Representing Predictive Validity Between Dynamic Assessments (DA) of Early Literacy Skills and Word-Reading (WR) Outcomes Measures

Study	N	DAPA-WR	DASSK-WR	DADec-WR
Caffrey, E. (2006 sample 1)	120			-.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**		
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. & Günther, T. (2018)	17		.855**	
Osa Fuentes, P.M. (2003)	164	.64**		
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) Sample 1 [b]	90	.516*		
Sittner Bridges, M. & Catts, H.W. (2011) Sample 2 [c]	96	.426*		
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

Analysis of Predictive Effect Sizes

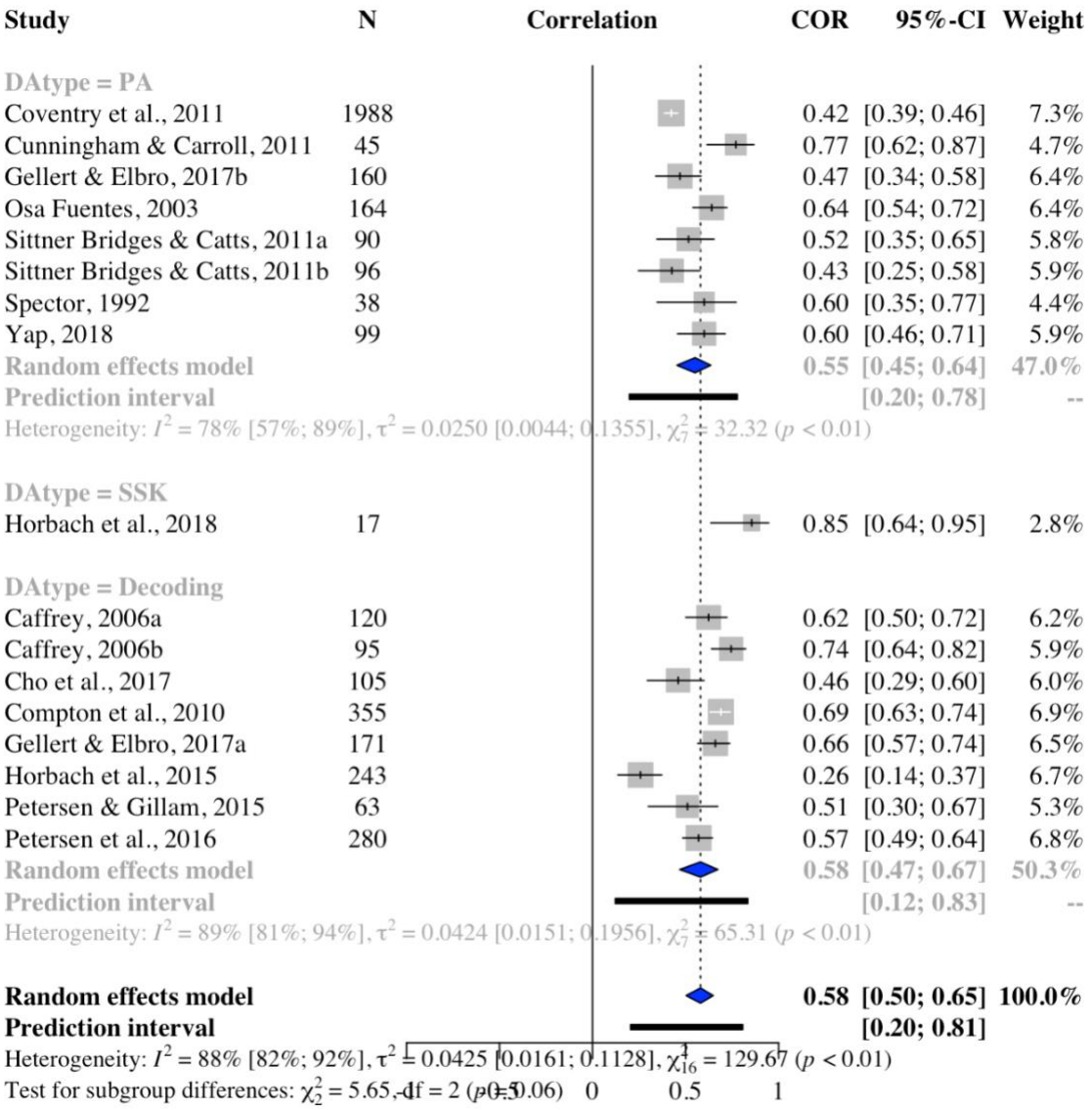
Results of the random-effects meta-analyses examining the predictive validity of dynamic assessments of early literacy tasks with later word reading outcomes are presented Figure 6.

Overall Predictive Validity DA-WR

The random effects meta-analysis found that the overall mean effect size is large (r=.58, 95%CI = [0.50-0.65]) suggesting that DAs are strongly correlated with word reading outcome measures. Additionally, the prediction interval ranged from g=0.20 to 0.81, and did not cross 0, indicating that future relevant studies are likely to find a positive correlation between the two. Significant heterogeneity was found (Q=129.67, p<.01) and between study variance was estimated at $\tau^2 =$ 0.0425 (95% CI=0.0161-0.1128] with an I^2 value of 88% (95% CI= 82-92%]).

Figure 5.

Forest plot of random effects meta-analysis examining the predictive validity of dynamic assessments of early literacy with word reading outcome measures.



Note. Study names, 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 = Dec. The grey box associated with each study represents the weight allocated to each effect size, while the horizontal

line that extends from either side of the box is a measure of the confidence interval (95%). The solid vertical line is the line of no effect while the dashed vertical line represents the significant overall mean effect size. The blue diamonds are an indication of the overall confidence interval, and the black bar represents the prediction interval. Figure drawn in R using `metacor` package (R Core Team, 2021; Laliberté, 2019).

Predictive Validity Subgroup Analysis

Given the significant heterogeneity, the moderator of DA type was included but not found to be significant (Q=5.65, df=2, p<.0593). This mixed effects model (see Figure 5) determined that there was not significant variation between subgroups of DA type (phonological awareness and decoding) and their predictive validity with tests of word reading outcomes as determined by correlation coefficients. The subgroup of SSK could not be compared due to a lack of studies. Result of the subgroup analysis are reported in Figure 6.

DA-PA predictive validity with word reading

The overall mean effect size representing the predictive validity between DAs of phonological awareness and word reading outcome measures is large (r=.55, 95%CI= [0.45-0.64]). The outcomes of the analysis suggest that there is a strong correlation between DA-PA and WR-OM. The prediction interval (95%CI= [0.20-0.78]), does not cross 0 indicating that negative correlation effect sizes are unlikely in future relevant studies.

DA-SSK predictive validity with word reading

The overall mean effect size representing the predictive validity between DAs of sound-symbol knowledge and word reading outcome measures is large (r=.61, 95%CI= [-0.28 - 0.94]). The outcomes of the analysis suggest that there is a strong correlation between DAs of SSK and

WR-OMs. However, this finding is based only on two studies, with a limited number of participants. Results should be interpreted with caution.

DA-Dec predictive validity with word reading

The overall mean effect size representing the predictive validity between DAs of decoding and word reading outcome measures is strong (r=.62, 95%CI= [0.54-0.69]). The outcomes of the analysis suggest that there is a strong correlation between DA-Dec and WR-OM. The prediction interval does not cross 0 (95%CI = [0.33-0.80]) indicating that negative correlation effect sizes are unlikely in future studies.

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=97.63, df=14, p<.0001), which indicates that there are likely other moderators beyond DA type that are impacting heterogeneity.

Figure 6.

Results of subgroup/moderator analysis

	g	95%CI	*p*	*I2*	95%CI	Prediction interval	*p subgroup*
DA type							.0593
Phonological Awareness (PA)	.55	0.45-0.64	<.01	0.78	0.57-0.89	0.20-0.78	
Sound-Symbol Knowledge (SSK)	-	-	-	-	-	-	-
Decoding (Dec)	.58	0.47-0.67	<.01	0.89	0.81-0.92	0.12-0.83	

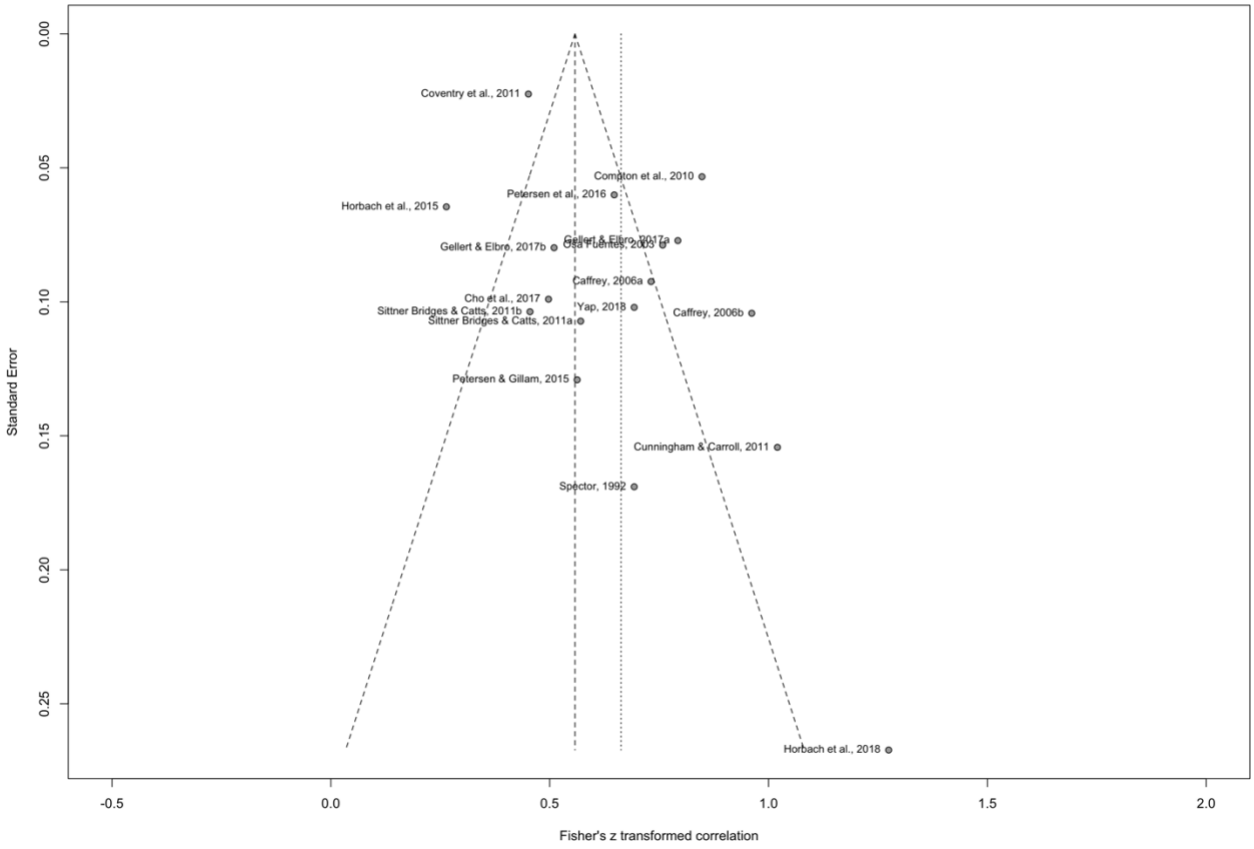
Risk of Publication Bias

Figure 7 was generated to examine publication bias in the predictive validity analysis. Visual inspection of the funnel plot suggests potential asymmetry, and Egger's test was significant for the presence of plot asymmetry (Intercept = 2.696, 95%CI [0.65 - 4.75-, p=.021).

Inspection of the plot reveals that there are several small studies with positive effects included in the analysis (e.g., Cunningham & Carroll, 2011; 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).

Figure 7.

Funnel plot of 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 funnel plot, 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).

Question 1B & 2B: Do dynamic assessments of early literacy skills demonstrate concurrent validity with static assessments of early literacy skills, and predictive validity with word reading outcome measures within population groups defined by their language (monolingual vs. bilingual) or reading status (typically developing, at-risk, diagnosed with difficulty)?

Initially, syntheses of concurrent and predictive validity for each population group defined by their language and reading status were planned (i.e., 6 groups typically developing/at-risk/diagnosed + monolingual/bilingual). However, there were insufficient studies identified to carry out any of these analyses. For this reason, narrative summaries of the findings for specific population groups of interest are reported below.

Bilingual Groups

Only 8/35 studies included bilingual participants and of those only 4 defined or described their bilingual participants and explicitly examined their performance separately from monolinguals. These 4 studies could not be grouped together in a meaningful way for statistical analyses. In this section, we will narratively elaborate on findings reported on the use of DA with bilinguals and extend this narrative analysis beyond correlational observations to provide a deeper understanding of the applications of DA for use with this group.

Concurrent validity of DAs with SAs of early literacy

2 studies reported concurrent correlation coefficients between DAs and SAs of the same early literacy construct. Lu & Hu (2019) and Yap (2019) both employed DAs of phonological awareness skills with bilinguals. Lu & Hu conducted their study with Cantonese/English sequential bilinguals in Taiwan, and Yap with primarily Chinese/English, Malay/English (and other language/English) sequential bilinguals in Singapore. Both reported strong correlations (r=.76, r=.82 respectively) between their DA and an SA of PA skills. These findings are consistent with the overall analyses of concurrent validity of DA-PA with SA-PA (r=.72).

Predictive validity of DAs with word reading outcomes

2 studies also reported predictive correlation coefficients between DAs and outcome measures of word reading. Petersen & Gillam (2015) reported a strong correlation between their DA of decoding and later word reading ability (r=.51) for use with sequential Spanish/English bilingual children in the United States. Yap (2019) also reported strong correlation between the DA-PA test administered in kindergarten and reading outcomes in grade 1 (r=.60) in her Singapore study with heterogeneous sequential bilingual groups. These findings are in line with the overall findings of the syntheses of the predictive validity of DA-Dec and DA-PA with word reading outcomes (r=.58 and r=.55 respectively). 1 study conducted in Hong Kong with sequential Cantonese/English bilinguals, (Chow, 2014), did not report concurrent correlations between DA and an equivalent SA, or use a longitudinal approach to examine predictive validity. However, the authors report cross-sectional correlations between their DA of SSK and Chinese and English word reading (r=.22* and r=.28** respectively). These findings related to English word reading are consistent with those reported in the synthesis regarding predictive validity of DA-SSK and word reading which indicated a moderate average correlation (r=.37). However, the findings related to

DA SSK, and Chinese word reading suggest a weaker relationship between these two measures. In summary, individual correlation coefficients from studies conducted with various sequential bilingual groups, representing the concurrent validity of DAs with SAs and the predictive validity of DAs with word reading outcomes are consistent with the results of the overall syntheses.

Unique predictive capacity of DAs beyond SAs

3 of the 4 studies conducted with bilinguals employed hierarchical regression analyses to examine the added predictive value of DA beyond SA in determining word reading outcomes. Chow et al., (2014), found that after controlling for nonverbal reasoning, phonological memory, and SA PA performance, their DA-SSK predicted an additional 7% in English word reading. Interestingly, these researchers employed two different tasks, a DA-SSK (which they term a visual-pronunciation paired associate learning (PAL) task), and another PAL task which was described as visual-semantic in nature. This visual-semantic PAL task uniquely predicted an additional significant 6%in Chinese word reading after controlling for all other variables but did not have predictive value in English word reading. Conversely, the visual-pronunciation PAL task (in our study described as a DA-SSK) only had predictive value for determining English word reading, and not Chinese word reading. In the hierarchical regression performed by Yap (2019), when age, SES, language dominance group, receptive and expressive language scores, word reading scores and SA PA scores were entered first, the DA PA did not add unique prediction to the outcome variable of letter word identification for the whole sample of children (n=99). However, when this same analysis was run including only children identified as at-risk (n=36), the DA PA task was the only significant predictor of word identification scores and accounted for a significant 11% of variance in outcomes. This finding points to the importance of differentiating at-risk and bilingualism. Finally, Lu & Hu (2019) also conducted a hierarchical regression analysis, but with

English real word spelling as the outcome measure. They found that after controlling for SA PA, their DA PA task accounted for 8% of variance in real word spelling. When DA PA was entered before SA PA, the SA task did not account for additional unique predictive variance.

Overall, the findings of the limited studies employing hierarchical regression analyses with heterogeneous bilingual populations indicate that DA can account for between 7% and 11% of variance in English word reading, and 8% variance in English word spelling. No studies examined word reading or spelling ability in other languages.

Performance differences on SAs and DAs

Though never a primary research question, 3 studies used either t test or ANOVAs to compare performance of bilinguals on DAs and SAs or to performance of monolinguals. Yap (2019) conducted a t- test to examine whether the average scores of bilingual children in their sample were significantly different from the average scores of the monolingual children in the normative sample. They found that the sequential heterogeneous bilingual children performed significantly poorer on 2 SA PA tasks from the CTOPP (elision and blending) than the monolingual children used in the normative sample. Gellert & Elbro (2017a) included 61 bilingual children in their total sample of 1 DOI: 10. They compared the average performance on all study measures of the bilingual children and the monolingual children using a t test and found that the monolinguals only performed significantly better on the vocabulary test. However, the authors did not specify what languages these children spoke or describe the nature of their bilingualism, making it difficult to meaningfully interpret these findings. Finally, Horbach et al., (2018), report no significant average performance differences on a DA SSK task between monolingual German speakers and multilinguals. The authors specify that all multilingual children (11 of the 17 in the

final sample), had minimally 2 years of exposure to German but do not report the languages that these children spoke.

Overall, based on a limited number of studies, bilinguals perform more poorly or similarly to monolinguals on SA measures of early literacy, and similarly to monolinguals on DA measures of early literacy. However, these findings should be interpreted with caution given that bilingual groups were typically poorly defined in terms of the languages spoken, the nature of their bilingualism, their age of acquisition or their proficiency levels.

At-Risk for and Diagnosed with Reading Difficulty Groups

15/35 studies only included participants who were typically developing (TD) in their reading abilities. Of the other 20 studies, 17 included participants at-risk (AR) for reading difficulty and 3 included participants already diagnosed with a reading disorder. Given the limited number of studies conducted with children diagnosed with dyslexia or other reading disorders, these two population groups were merged in this narrative analysis. Many studies included AR or diagnosed groups merged with the groups when reporting their correlational findings. Therefore, as with the bilingual group, we extend this narrative analysis beyond correlation coefficients for a more thorough examination of the use of DAs with at-risk and diagnosed groups.

Concurrent validity of DAs with SAs of early literacy

3 studies report separate concurrent correlation coefficients between DAs and SAs for AR groups. Cho et al., (2014), found a strong correlation between their DA-Dec and a SA-Dec (r=.69**) in at-risk populations. Two studies conducted by Aravena et al., (2013; 2018), reported separate coefficients for children in their study who were diagnosed with dyslexia. In the 2013 study, the authors indicate that their DA-Dec correlated strongly with a SA-Dec (r=.52), but they

report only a weak correlation between the DA-Dec and SA-Dec in the 2018 study (r=.237). These results are mostly consistent with the overall concurrent validity between DA-Dec and SA-Dec synthesized in this review (r=.50). It is unclear why these correlations from the two Aravena studies differ, as the participants, methods of diagnosing dyslexia, and the measurement tools utilized were similar in both studies.

Predictive validity of DAs with word reading outcomes

2 studies report separate predictive correlation coefficients between DAs and word reading outcomes. Petersen & Gillam (2015) conducted DA-Dec with AR bilingual participants and found that this measure correlated strongly with later word reading (r=.51). Spector (1992) conducted a DA-PA and found strong correlations between this measure and later reading outcomes in at-risk monolinguals (r=.60). These findings are similar to those calculated in the meta-analyses of the overall predictive validity of DA-Dec and DA-PA with word reading (r=.62, r=.55 respectively). Overall, with the limited number of studies that report separate correlation coefficients for AR or diagnosed groups, no major differences were found in terms of concurrent or predictive validity of DAs for use with AR or diagnosed populations.

Unique predictive capacity of DAs beyond SAs

3 studies employed separate statistical analyses with the at-risk groups to analyze whether DA added unique, significant predictive value beyond other predictor variables, including traditional SAs. Horbach et al., (2015) employed a hierarchical regression to determine that after controlling for IQ, early reading ability, static PA scores, short term memory and rapid automatic naming performance, their DA-SSK added a significant 8% unique variance to later word reading outcomes for the at-risk nonreaders group but did not add unique variance to word reading outcomes for the typically developing early readers group. Spector (1992) utilized a stepwise

multiple linear regression approach and reported that after controlling for SA PA scores, spelling scores and receptive vocabulary outcomes, the DA PA accounted for an additional significant 21% unique variance in spring reading performance in the study's at-risk sample. Finally, Cho et al., (2014) created a conditional model to predict variance in growth in reading skills. They found that after entering two SAs of decoding, the DA Dec was still a significant predictor of word reading growth, accounting for 7% of the final level if growth variance. In this model, the DA-Dec was the only significant predictor of linear growth. 5 other studies conducted regression analyses to examine the added predictive capacity of DA but merged TD and AR or diagnosed participants in their analyses and are therefore not discussed here. Overall, the results of these 3 studies suggest that DA (of various early literacy constructs) can account for between 7-21% significant unique variance in later word reading performance.

Performance differences on SAs and DAs

4 studies compared the average performance of AR or diagnosed groups to TD groups on SAs and DAs using t tests or ANOVAs. Aravena et al., (2013) conducted t tests to examine whether TD and dyslexic children performed significantly differently on their DA of SSK or DA of Decoding. The findings suggest that TD children on average had greater accuracy on the SSK component of the tool, read significantly faster and significantly more words in the DA Decoding portion than the children diagnosed with dyslexia. In a follow-up 2018 study, Aravena et al., (2018) used a 2-way ANOVA to demonstrate that the dyslexic group on average made more errors in the implicit training section of the DA than the TD group when timed, and again found that the TD group read significantly more words at a faster rate than the dyslexic group. In their 2015 study, Horbach et al., conducted a two-way repeated measures ANOVA to examine whether the early TD early readers group and the AR nonreaders group performed differently on their DA SSK and

decoding tasks. They found a significant main effect of group, indicating that in general, early readers performed better than nonreaders. Follow up t-tests suggest that these TD early readers outperformed the AR nonreaders on all DA tasks. Finally, Teeuwen (2010), conducted a MANOVA to compare performance differences between TD and AR groups on both DA and SA tasks. Teeuwen reports significant performance differences on the SA SSK and PA tasks but no significant performance differences on the accuracy or reaction measures of the DA-SSK task.

In summary, the results of 3 studies are mixed. 2/3 found that TD and AR risk groups perform differently on DA tasks, while 1 study suggests they do not. Additionally, only one study compared TD and AR performance on SA tasks and reported significant difference in outcomes.

Discussion

We conducted a systematic review and meta-analysis on the concurrent and predictive validity of dynamic assessment (DA) of phonological awareness (PA), sound-symbol knowledge (SSK) and decoding for use overall and with populations stratified by reading and language status. Thirty articles met the inclusion criteria, and 23 articles were identified in the initial database search and the preprint, forward and ancestral searches yielded an additional 7 papers. All articles were appraised for quality.

The results of the overall meta-analysis examining the *concurrent validity* of DAs of early literacy with their equivalent static assessment (SA) counterparts suggest that there is a strong correlation between the two types of tests (r=.58). Moderator analysis suggests that there are significant differences in effect sizes for DAs of different literacy skills, with DA-PAs demonstrating the strongest correlation (r=.72) with their SA counterpart, followed by DA-Dec (r=.50) and DA-SSKs (r=.34). It is possible that this is because PA is a well-defined construct with

an established hierarchy of skills that can be consistently evaluated across languages (Anthony & Francis, 2005). Furthermore, because of decades of research conducted on the role of PA as a literacy precursor, there appears to be a greater degree of consistency in the tasks researchers use to evaluate PA dynamically and statically across studies. Specifically, most researchers used a DA and an SA of a phoneme level task, either a simple task like segmentation or a complex task like deletion (e.g., give examples). Additionally, DA-PAs and equivalent SA-PAs were both exclusively conducted in-person. The use of consistent tasks and administration methods across DAs and SAs should result in higher correlations between the two measures. Conversely, assessments of SSK and decoding were characterized by a greater degree of variability in task type and method of administration between DAs and SAs (e.g., a DA of SSK which evaluates via computer the ability to learn a novel symbol's sound and the SA which evaluates letter name knowledge in-person). Overall, the results of this synthesis provide suggestive evidence for the concurrent validity of DA-PA and DA-Decoding with their SA equivalents.

A second meta-analysis examining the *predictive validity* of DAs of early literacy with later word reading outcome measures (OMs), found that overall, there is a strong correlation between the two (r=.58). Again, moderator analysis with subgroups defined by DA type was conducted but, in this instance, determined that there were not significant differences in effect sizes for DAs of PA and decoding. In this analysis, DA-Dec demonstrated the strongest correlation with later word reading outcome measures (r=.58), but this overall weighted effect size was not significantly different from that of the DA-PA with word reading (r=.55). Only one study was included in the DA-SSK subgroup, so an overall effect size could not be calculated. This single study with only 17 participants cannot be generalized. Regarding DA-Dec, it is logical that a DA test that evaluates ability to learn how to decode words would demonstrate 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 in a DA, and alphabetic word reading as an OM. There was also variation in the administration. In some instances, the DA-Dec measure was conducted via computer, while the OM was done in-person. Despite this, a strong correlation was found between DA-Dec and word reading outcomes, providing suggestive evidence of the predictive validity of these DA-Dec tests.

It should be noted that in the concurrent validity analysis, the prediction interval calculated in this analysis crossed 0. This indicates that future relevant studies may demonstrate a negative correlation between DAs and SAs. This is likely a result of the small effect sizes documented for relationships between DA SSK and SA SSK. Furthermore, in both analyses there was significant heterogeneity in the studies included, which was not accounted for by DA type moderator analyses. It is possible that this heterogeneity was a result of difference in participant characteristics, such as language or reading status, or languages spoken. However, subsequent moderator analyses using these variables as subgroups were not conducted due to the inability to separate these factors from one another in a single participant (e.g., participants can be typically developing AND monolingual and defining them by only their reading status in an analysis negates the potential role of their language status). For this reason, and because of a lack of studies conducted with children who were not typically developing (TD) monolinguals, a narrative review of the findings related to the concurrent validity of DAs for use with bilinguals and at-risk (AR) groups was conducted.

Comparisons of the correlational findings reported in studies examining the concurrent and predictive validity of DAs with *bilingual groups* were consistent with the overall findings from each analysis, meaning the effect sizes from individual studies conducted with bilinguals did not

differ markedly from the overall weighted effect sizes representing concurrent and predictive validity of DAs with SAs and OMs. However, only 8 of 35 studies included bilinguals in their sample, and of those 8, only 4 conducted separate analyses of bilinguals to allow for comparison with monolinguals. In addition, most studies poorly characterized their bilingual participants. Frequently, the nature of bilingualism was not defined (i.e., sequential vs. simultaneous) and little to no information was provided about the languages spoken, the age of acquisition or proficiency levels. It is possible that this is because for most studies, the role of language status in the outcomes of DA and SA testing was not the primary research question. Many studies included bilingual children out of convenience and conducted analyses only to determine whether their overall average scores differed significantly from the monolingual groups, to merge the two groups and proceed with addressing their primary research questions. Of the studies that did define and examine bilingual performance, 3 studies used hierarchical regression models to examine whether DA added unique variance beyond SA in predicting word reading outcomes. The findings were promising and suggested that DAs add 7-11% unique variance in predicting word reading outcomes, and 8% unique variance in word spelling outcomes beyond traditional static measures. There is also some evidence from 3 studies indicating that monolinguals and bilinguals perform differently on SAs but not on DAs. Specifically, two studies suggest that monolinguals outperform bilinguals on SA measures of phonological awareness and vocabulary, and one study reported that the groups do not perform differently on a DA of SSK. Because of the limited number of studies, it is impossible to determine whether these differences can be attributed to the capacity of DAs and SAs to accurately capture bilingual ability. It is possible that bilinguals underperform on static tests developed for monolinguals but are less likely to do so on DAs which are designed to capture the ability to learn rather than acquired knowledge. Additional research is required to examine this

possibility. In summary, the findings from these few studies are limited, but encouraging. Importantly, though DA has been touted as an evidence-based practice for use with bilinguals, this population continues to be neglected in research related to the application of these tools when it comes to evaluation of early literacy skills.

A similar pattern of findings emerged for *the at-risk (AR) group*. Only 3 studies included children diagnosed with a reading disorder, and while more studies included participants at-risk for a reading disorder (17), many of these did not conduct separate statistical analyses with AR groups. The correlational findings from 3 studies that examined the concurrent validity of DAs and SAs with AR participants were mostly consistent with the overall synthesis results which found a strong correlation between DAs and SA. There was one notable exception, the 2018 Aravena et al., study, which reported a much weaker relationship between a DA-Dec and equivalent SA-Dec (r=.24), than the overall weighted mean effect size (r=.50). It is unclear why this finding differs so markedly from the findings of the previous Aravena study (2013), which recruited similar participants, used similar methods and measures, and found a much stronger correlation between DA and SA (r=.52). Regarding the predictive validity of DAs with at-risk groups, the effect sizes of the individual studies were strong, and consistent with the overall finding of a strong correlation between DAs and reading OMs (r=.58). Reading status factored into author's primary research questions much more frequently than language status. However, the same pattern of inadequate definition observed in the description of bilinguals was also noted in the way researchers defined and identified who was at-risk. At times, researchers did not outline how the at-risk status was decided, and individual authors employed different methods to determine this status. This heterogeneity makes it difficult to interpret and generalize the findings across studies in a meaningful way. Despite this, the trends that emerged from the narrative

analysis are encouraging. 3 studies reported that diverse types of DAs of early literacy added unique predictive value, ranging from 7-21%, beyond SAs in determining later word reading ability for at-risk groups. 4 other studies examined whether at-risk and TD populations perform similarly on DAs and SAs. 3/4 studies provide evidence indicating that at-risk/diagnosed and TD children perform differently on DAs of SSK and decoding (Aravena et al., 2013;2018; Horbach, 2015). One study provided conflicting results and reported that the at-risk group and TD children did not perform significantly differently on the DA-SSK in their study (Teeuwen, 2010). In summary, at-risk status is poorly defined and myriad methods are used to determine who fits this criterion in the current body of research. Many papers do not conduct separate analyses with at-risk groups to allow for comparison with TD peers. The results from the limited number of studies that do analyze at-risk populations separately, provide some evidence to suggest that DAs add unique predictive value beyond SAs for this group, and that at-risk groups do perform differently on DAs than TD children, pointing to the potential for this type of task to differentiate at-risk children from their peers.

Limitations

Despite a comprehensive database and grey literature search, it is possible that relevant articles were not identified. For example, much like with the synonym computerized adaptive testing (CAT) we also encountered relevant studies that used the term “paired associate learning” (PAL). While we reran our search using the term CAT, we ultimately elected not to include the term PAL as a synonym. This was primarily because our research team agreed that not all PAL tasks are inherently dynamic in nature, and secondarily because the volume of articles yielded with this term added was substantial. Rerunning our search including PAL would have meant expending

resources on screening the thousands of PAL articles. This was not only not feasible, but also not necessarily likely to result in the identification of new, relevant articles. As stated by Dixon (2022), future reviews may wish to include papers focused on PAL tasks and compare these methods with DAs and SAs.

Additionally, it is possible that some 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. The preprint repository search produced several articles in Portuguese, Korean and Mandarin, that may also have been relevant. Future review papers 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 outside in non-Western European languages might be included.

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 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. This allowed us to address select limitations of other previous DA review papers (e.g., Orellana, Hunt, Dixon) who chose different outcome measures, and consequently identified fewer studies and in some cases were unable to conduct a meta-analysis of the findings. However, because of this choice, the results can only provide insight into the associations or relationships between DAs and SAs or OMs, rather than the causal role of DA. 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 in the domain of study.

Finally, to examine the purported potential of DA for use with bilingual and at-risk populations, we planned to conduct separate syntheses examining the validity of its' use with these groups. However, even when using the most common effect size measure (correlation coefficients), there were still insufficient number of studies that conducted separate analyses with either of these groups to justify a separate synthesis of the results. Bilinguals were only included in 8 out of 35 studies, and only 4 of those studies accurately described the group or provided a separate analysis of their performance. At-risk participants were included in 17 studies, and diagnosed children were the focus of 3 papers, but again, researchers in all but 7 studies grouped these children together with typically developing participants in the correlational or regression analyses. As a result, we were only able to narratively synthesize the findings. To provide a more robust account of the use of DAs with these populations, we extended our narrative review to include a description of the findings related to the unique predictive value of DAs beyond SAs, and the performance differences of these groups on DAs and SAs. Finally, including studies with a broad range of types and formats of DAs, SAs and OMs used, and participant characteristics resulted in significant heterogeneity in the analyses. While we were able to determine through subgroup analyses that DA types differed significantly in their mean weighted effect sizes, inclusion of this moderator did not account for the heterogeneity observed. Due to overlap in participant characteristics in studies (e.g., inclusion of at-risk and typically developing children who were mono and bilingual) we were not able to conduct subsequent moderator analyses to examine the role of language or reading status.

Acknowledgements

This study was funded by in part by a Social Sciences and Humanities Research Council Canadian Graduate Scholarship-Master's Grant to the first author and by a Natural Sciences and Engineering Research Council Discovery Grant to the last author (RGPIN-2019-06523).

Author Notes

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

References

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

American Speech-Language-Hearing Association (2004). *Knowledge and skills needed by speech-language pathologists and audiologists to provide culturally and linguistically appropriate services*. www.asha.org/policy

American Speech-Language-Hearing Association (n.d.). *Written language disorders*. www.asha.org/Practice-Portal/Clinical-Topics/Written-Language-Disorders/

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

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

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

**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.

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).

**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., 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

**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

Cochran, W.G. (1954). Some methods for strengthening the common X2 tests. *Biometrics, 10*(4): 417-51.

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., ... & 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 early literacy and related skills* (Doctoral dissertation, University of Warwick).

Cunningham, A., & Carroll, J. (2011). Age and schooling effects on early literacy and phoneme awareness. *Journal of Experimental Child Psychology, 109*(2), 248-255. DOI: 10.1016/j.jecp.20 DOI: 10.12.005

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*, 002221942211175 DOI: 10.

Dumas, D., McNeish, D., & Greene, J. A. (2020). Dynamic measurement: A theoretical–psychometric paradigm for modern educational psychology. *Educational Psychologist*, *55*(2), 88-105.

**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.

Fuchs, D., Compton, D. L., Fuchs, L. S., Bouton, B., & Caffrey, E. (2011). The construct and predictive validity of a dynamic assessment of young children learning to read: Implications for RTI frameworks. *Journal of learning disabilities*, *44*(4), 339-347. DOI: 10.1177/0022219411407864

Furlong, L. M., & Serry, T. A. (2022). An exploratory study of speech-language pathologists' clinical practice in the literacy domain: Comparing onsite practices with telepractice services during COVID-19. *International Journal of Speech-Language Pathology*, 1-13. DOI: 10.1080/17549507.2022.2030410

**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

Good, R. H., Gruba, J., & Kaminski, R. A. (2002). Best Practices in Using Dynamic Indicators of Basic Early Literacy Skills (DIBELS) in an Outcomes-Driven Model.

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

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

Hamel, R. E. (2007). The dominance of English in the international scientific periodical literature and the future of language use in science. *Aila Review, 20*(1), 53-71. DOI: 10.1075/aila.20.06ham

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.

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

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

Hunt, E., Nang, C., Meldrum, S., & Armstrong, E. (2022). Can Dynamic Assessment Identify Language Disorder in Multilingual Children? Clinical Applications from a Systematic Review. *Language, Speech, and Hearing Services in Schools*, *53*(2), 598-625. DOI: 10.1044/2021_LSHSS-21-00094

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

Kim, B. S., Lee, D. W., Bae, J. N., Chang, S. M., Kim, S., Kim, K. W., ... & Cho, M. J. (2014). Impact of illiteracy on depression symptomatology in community-dwelling older adults. *International psychogeriatrics*, *26*(10), 1669-1678.

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

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).*

LEADERSproject. (2020). *Understanding assessment: Assessment materials- dynamic vs. static Assessment.* https://www.leadersproject.org/2013/03/01/assessment-materials-dynamic-vs-static-assessment/

**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

**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.

Ministry of Education. (2021). *Education in Ontario: policy and program direction: Policy/program memorandum 8.* https://www.ontario.ca/document/education-ontario-policy-and-program-direction/policyprogram-memorandum-8

Montoya, Silvia. "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. 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

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

Orellana, C. I., Wada, R., & Gillam, R. B. (2019). The use of dynamic assessment for the diagnosis of language disorders in bilingual children: A meta-analysis. *American Journal of Speech-Language Pathology*, *28*(3), 1298-1317. DOI: 10.1044/2019_AJSLP-18-0202

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

**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

**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

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

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

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.

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

Sidik, K., & Jonkman, J. N. (2005). Simple heterogeneity variance estimation for meta-analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 54*(2), 367-384. DOI: 10.1111/j.1467-9876.2005.00489.x

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.

Speech-Language Pathology Audiology Canada. (2020). *Speech-language pathology service delivery models in schools.* https://www.sac-oac.ca/sac-work/position-papers-and-guidelines

Spineli, L.M., Pandis, N. (2021). Prediction interval in random-effects meta-analysis. *Statistics and Research Design 157*(4):586-588.

Statistics Canada. (2016). *Census in Brief: Bilingualism among Canadian children and youth.* https://www150.statcan.gc.ca/n1/daily-quotidien/191216/dq191216c-eng.htm

Statistics Canada. (2011). *Census in Brief: Linguistic characteristics of Canadians.* https://www12.statcan.gc.ca/census-recensement/2011/as-sa/98-314- x/98-314-x2011001-eng.cfm

Statistics Canada. (2022). *The Daily — While English and French are still the main languages spoken in Canada, the country's linguistic diversity continues to grow*. Statcan.gc.ca. https://www150.statcan.gc.ca/n1/daily-quotidien/220817/dq220817a-eng.htm

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

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

Thornton, A., & Lee, P. (2000). Publication bias in meta-analysis: its causes and consequences. *Journal of clinical epidemiology*, *53*(2), 207-216. DOI: 10.1016/s0895-4356(99)00161-4

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

Vellutino, F. R., Scanlon, D. M., & Tanzman, M. S. (1998). The case for early intervention in diagnosing specific reading disability. *Journal of School Psychology*, *36*(4), 367-397.

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.

West, S. L., Gartlehner, G., Mansfield, A. J., Poole, C., Tant, E., Lenfestey, N., ... & Lohr, K. N. (2011). Comparative effectiveness review methods: clinical heterogeneity.

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

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

**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 Univer

Figure 1.

PRISMA flowchart of literature search

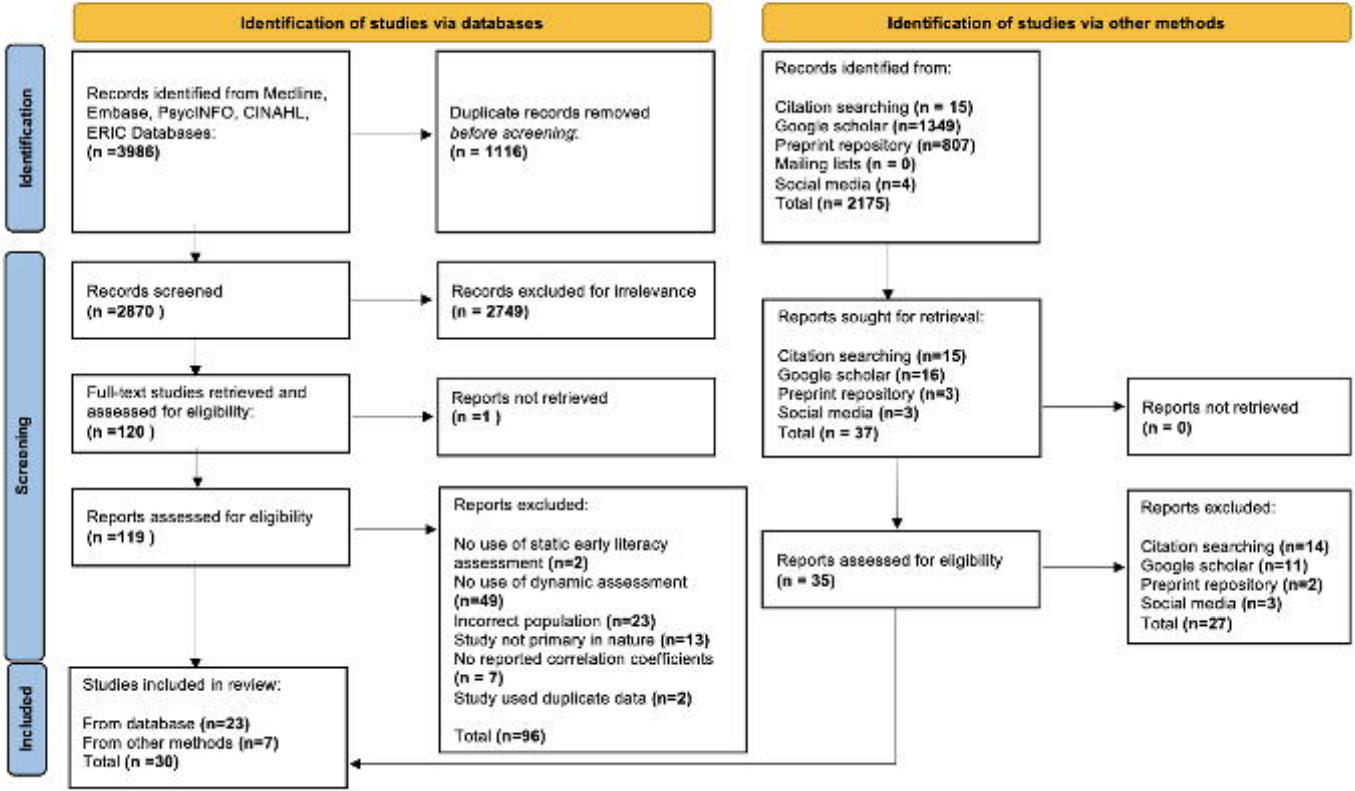


Figure 2.

Forest plot of random effects meta-analysis examining the concurrent validity between dynamic and static assessments of early literacy.

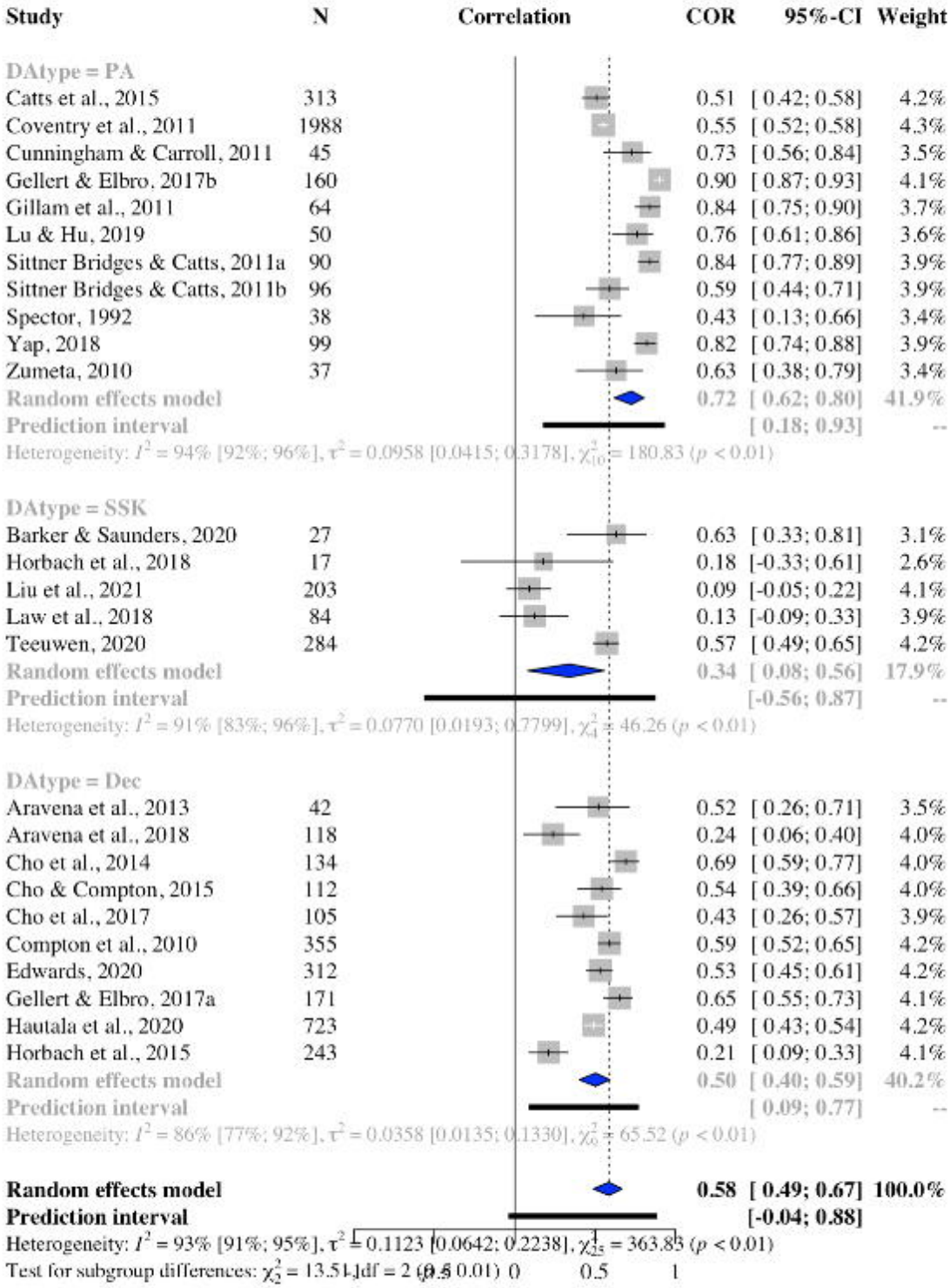


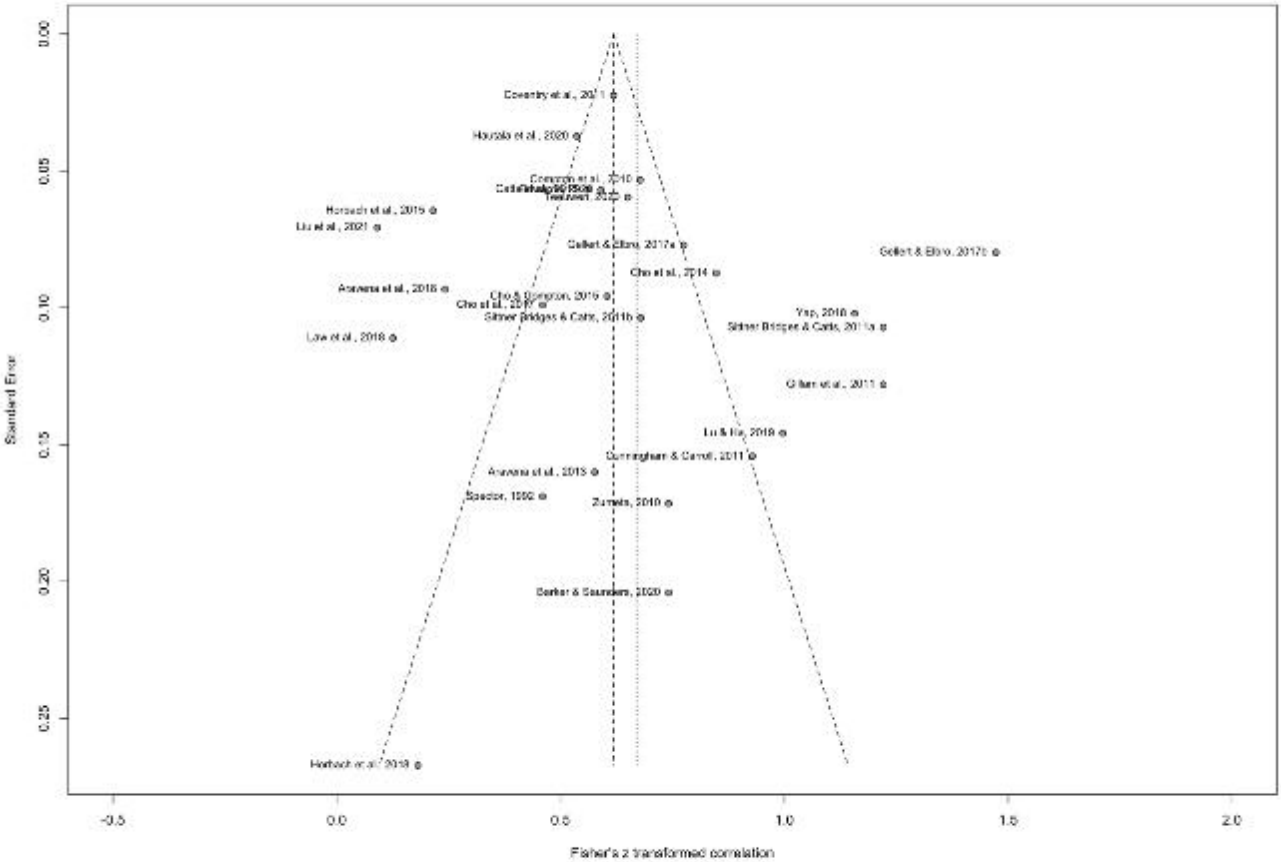
Figure 3.

Results of subgroup/moderator analyses

	g	95%CI	*p*	*I2*	95%CI	Prediction interval	*p subgroup*
DA type							.0012
Phonological Awareness (PA)	.72	0.62-0.80	<.01	0.94	0.92-0.96	0.18-0.93	
Sound-Symbol Knowledge (SSK)	.34	0.08-0.56	<.01	0.91	0.83-0.91	-0.56-0.87	
Decoding (Dec)	.50	0.40-0.59	<.01	0.86	0.77-0.92	0.09-0.77	

Figure 4.

Funnel plot of studies included in the meta-analysis of the concurrent validity between dynamic and static assessments of early literacy skills.



Note. In the funnel plot, 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).

Figure 5.

Forest plot of random effects meta-analysis examining the predictive validity of dynamic assessments of early literacy with word reading outcome measures.

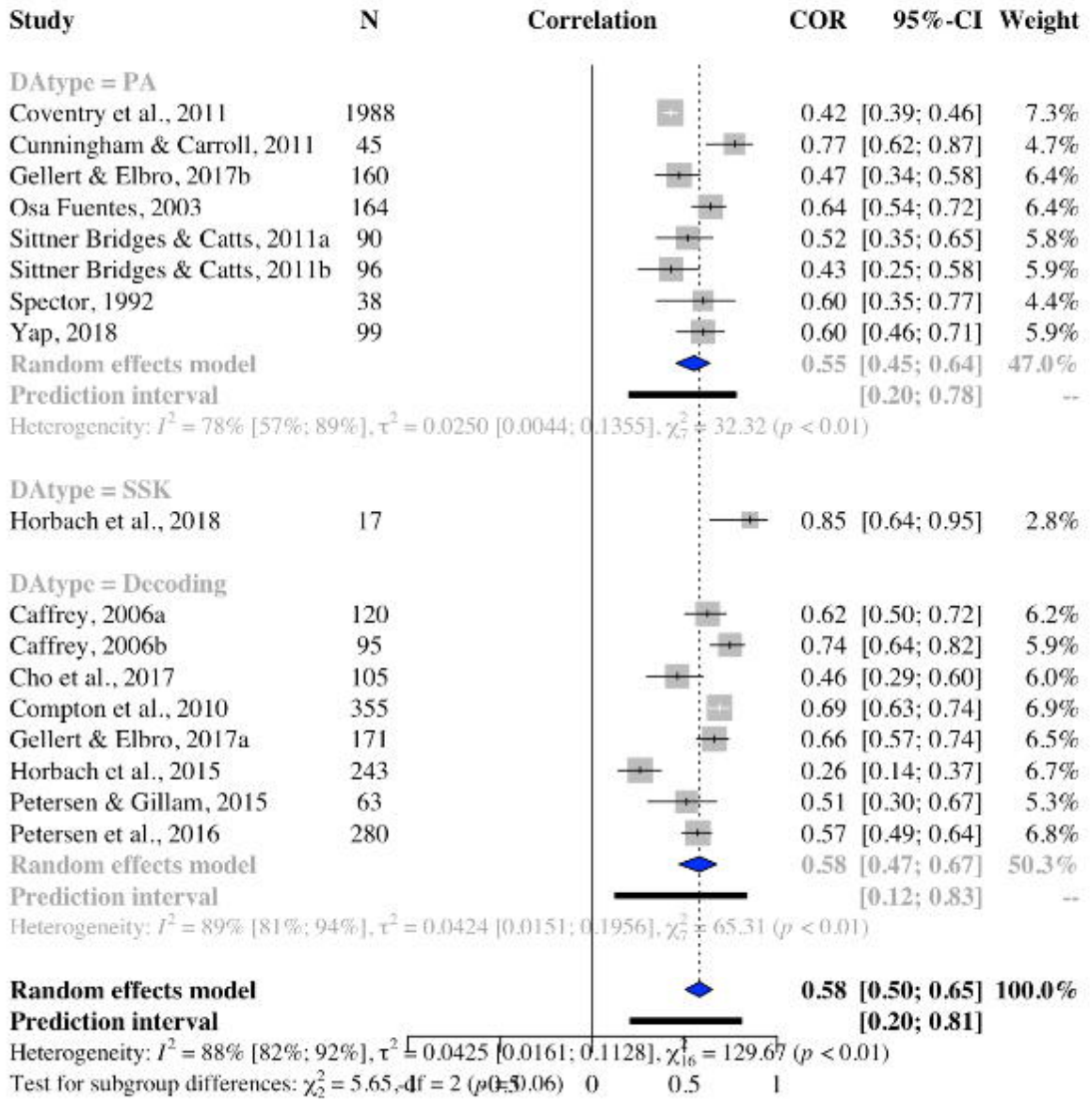
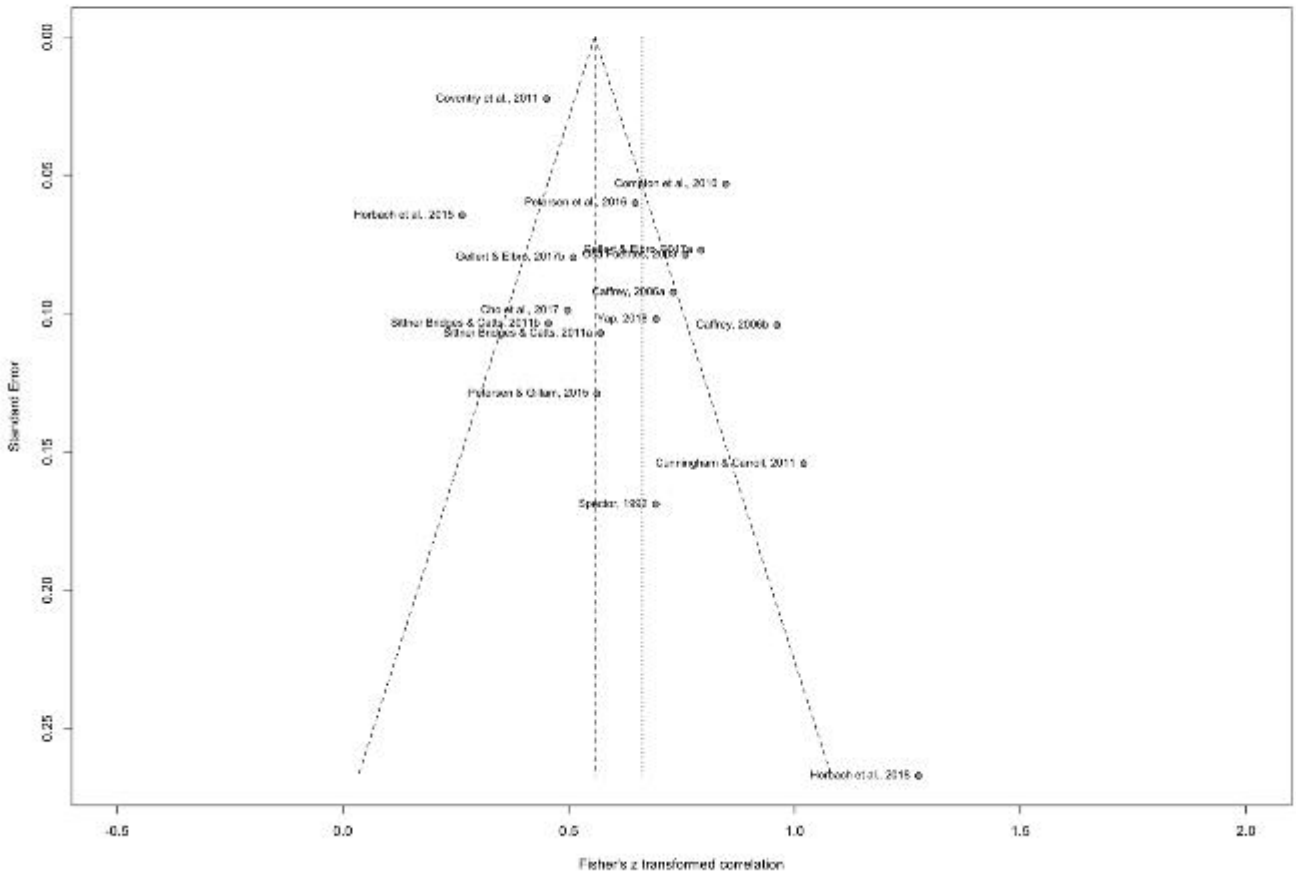


Figure 6.

Results of subgroup/moderator analysis

	g	95%CI	*p*	*I2*	95%CI	Prediction interval	*p subgroup*
DA type							.0593
Phonological Awareness (PA)	.55	0.45-0.64	<.01	0.78	0.57-0.89	0.20-0.78	
Sound-Symbol Knowledge (SSK)	-	-	-	-	-	-	-
Decoding (Dec)	.58	0.47-0.67	<.01	0.89	0.81-0.92	0.12-0.83	

Figure 7.

Funnel plot of studies included in the meta-analysis of the predictive validity of dynamic assessments of early literacy skills with word reading outcome measures