Article:
Dixon, C, Oxley, E, Nash, H orcid.org/0000-0002-4357-945X et al. (1 more author) (2022) Does Dynamic Assessment Offer An Alternative Approach to Identifying Reading Disorder? A Systematic Review. Journal of Learning Disabilities. 222194221117510-. ISSN 0022-2194

https://doi.org/10.1177/00222194221117510

Abstract

Traditional static tests of reading and reading-related skills offer some ability to predict future reading performance, though such screeners may misclassify children with or at risk of reading disorder (RD). Dynamic assessment (DA) is an alternative approach which measures learning potential and may be less dependent on learning background. A systematic review was carried out to examine the ability of DA to classify children with or at risk of RD. A database search yielded 14 eligible articles, assessing DA of decoding, phonological awareness, and working memory. Results suggest that DA explains unique variance in the prediction of later RD status, and although models with a single dynamic predictor sometimes achieved good classification accuracy, this was enhanced somewhat by the addition of static predictors. Higher classification accuracy was found for DA targeting constructs more proximal to reading, particularly decoding, but the predictive power of DA of decoding and phonological awareness appeared to wane with increasing age as static measures explained more variance in outcomes. Some evidence emerged that DA provides benefits over static tests for the prediction of RD in bilingual students, though no studies examined outcomes by administration format or orthographic depth. Limitations and suggestions for future work are discussed.

Keywords: dynamic assessment, reading disorder, classification, decoding, phonological awareness

Does Dynamic Assessment Offer an Alternative Approach to Identifying Reading Disorder? A Systematic Review

A good level of reading skill provides a solid foundation for educational achievement and a range of life outcomes (Morrisroe, 2014). Despite dramatic increases in the global literacy rate in recent years, concerns continue to be raised over the proportion of school children in developed economies failing to attain basic proficiency in reading (Save the Children, 2015; Schleicher, 2019). DSM-5 (American Psychiatric Association, 2013) defines Specific Learning Disorder (SLD) as a difficulty in the acquisition and use of an academic skill (e.g., reading, writing, or mathematics) that has persisted for at least six months despite appropriate intervention. Individuals with reading impairment may encounter difficulties in reading accuracy, fluency, and comprehension, as well as written expression and mathematical reasoning. Despite heterogeneity and comorbidity of such difficulties, two types of reading impairment profile are commonly found in school-aged children. Dyslexia is characterised by slow and effortful word reading, and affects 5-17% of children depending on diagnostic criteria (Grigorenko et al., 2020). The causal mechanism for dyslexia is thought to involve a phonological impairment which results in delays in decoding and orthographic learning (Bailey et al., 2004; Melby-Lervåg et al., 2012), but does not affect linguistic comprehension. Conversely, specific reading comprehension impairment, with prevalence of around 5-11% (Kelso et al., 2020), is characterised by accurate and fluent decoding but difficulty in comprehending text. These *poor comprehenders* present with a range of oral language impairments, particularly in vocabulary knowledge (Landi & Ryherd, 2017), but their reading difficulties may be more likely to go unnoticed in the classroom. We use the term reading disorder (RD) here to refer to children with a specific learning disorder affecting

their reading skills in either decoding or comprehension though again, it should be noted that a range of different criteria may be used to determine RD.

Identification of Specific Learning Disorder

The classification of SLD continues to be the subject of debate in the literature and as alluded to above, prevalence rates vary according to diagnostic criteria. Where once an IQ-achievement discrepancy was routinely employed as a diagnostic criterion, this practice was later challenged by research showing lack of qualitative differences in the reading difficulties of children with and without low levels of IQ (Stanovich & Siegel, 1994). Two alternative approaches to SLD identification which take account of individual differences in children's learning processes are response to intervention (RTI) and dynamic assessment (DA). Within an RTI framework, children's rate of learning progress throughout a period of regular classroom instruction may be used to gauge risk for SLD (Fuchs & Fuchs, 2006). Although RTI measures children's progress over time in a continuous fashion, it has been criticised for its 'wait-to-fail' approach and is still subject to statistical issues imposed by arbitrary cut-offs for assigning 'non-responder' status (Burns & Senesac, 2005). The present review focuses on DA, a framework conceptually related to RTI but requiring a much shorter period of time to implement. Before a detailed discussion of DA, some key aspects of screening for SLD will be outlined below.

The practice of educational screening typically involves the administration of standardised tests of achievement in order to identify children with or at risk of RD. A major motivating factor behind such testing is the efficacy and cost-effectiveness of intervention when provided as early as possible in the reading acquisition process (Torgesen, 2000). Screening is fraught with prognostic difficulty, however, with significant risks of under- and overidentification. Criteria used to assess the accuracy of screening measures include sensitivity – the proportion of true positives – and specificity – the proportion of true

negatives. Although there is a trade-off to be had between these two indices of classification accuracy, sensitivity of 80% is widely considered a minimum desirable threshold (Glover & Albers, 2007). This trade-off is analysed statistically using receiver operating characteristic (ROC) curve analysis, in which an area under the curve (AUC) represents the likelihood that a correct diagnosis will be applied among any randomly selected pair of individuals (Petersen et al., 2016). AUC ranges between 0 and 1, where values of .5 or under indicate no discrimination, while values of .7 to .8 and .8 to .9 represent *acceptable* and *excellent* discrimination, respectively (Hosmer et al., 2013). The optimal cut-off score for a screener is one which maximises the true positive rate while minimising the false positive rate (Streiner & Cairney, 2007).

The accuracy of screening is affected by several factors, including the number and particular combination of tests in a screening battery, the point at which screening is conducted, and the number of times a screener is administered (Glover & Albers, 2007; O'Connor & Jenkins, 1999; Poulsen et al., 2017). Screeners consisting of traditional measures of reading-related skills may yield poor identification accuracy for two reasons. Firstly, screeners of early-acquired code-based skills such as grapheme-phoneme correspondence rules are likely to yield floor effects if administered at or shortly after the onset of formal reading instruction (Catts et al., 2009). Secondly, by focusing on developed ability, traditional screeners are insensitive to variation in children's home learning experiences, which is particularly problematic for children from culturally and linguistically diverse (CLD) backgrounds (Peña & Halle, 2011). For these children, low performance may be a result of different learning opportunities and experiences as opposed to a learning disorder per se, potentially inflating the false positive rate of a test (Tzuriel, 2000). While sensitivity and specificity may be improved with larger screening batteries, this strategy calls for more time and resources (Compton et al., 2010). Timing considerations in screening

programmes are particularly important given the importance of early identification and intervention for children with RD; indeed, extended screening batteries and the progress monitoring approach of RTI both require significant time commitments, prompting consideration of alternative approaches such as dynamic assessment (Grigorenko, 2009).

Dynamic Assessment

Dynamic assessment (DA) is an approach to psychological testing that grew out of dissatisfaction with traditional static, standardised, norm-referenced cognitive ability tests. DA distinguishes between developed ability and latent capacity; or in other words, between the skills that have been acquired up until the day of testing, and the potential to learn new skills given assistance (Grigorenko & Sternberg, 1998). A core aspect of DA is therefore the measurement of examinees' ability to respond to teaching or intervention in an effort to predict future performance. The origin of DA can be traced to multiple sources, though the theoretical foundation of much work is predicated on Vygotsky's *zone of proximal development* (see Dumas et al., 2020 for a review).

DA shifts the focus of static tests from the product of learning to the process of learning by incorporating explicit feedback into the assessment procedure in order to measure psychological processes of change. This violates the neutral and detached nature of the examiner-examinee relationship in static testing, though does not obviate the standardisation of dynamic tests. Data arising from dynamic testing are therefore interpreted in an *ideographic* rather than norm-referenced fashion; that is, in terms of the comparison of within-individual change from pretest to posttest as a result of teaching (Haywood & Lidz, 2007, p.12). Due to its substantive focus on learning potential, DA is likely to be particularly appropriate for individuals from CLD backgrounds, as well as for children in the early stages of skill acquisition, where, as discussed above, static tests are more likely to yield undesirable statistical properties for the prediction of future performance.

The term *dynamic assessment* subsumes a number of major frameworks including structural cognitive modifiability, learning potential testing, testing-the-limits, *Lerntest*, and graduated prompts (see Sternberg & Grigorenko, 2002 for a review). These frameworks, among others, have been categorised in terms of task procedures. For instance, in Haywood's (1997) nomenclature, DA involves either (i) restructuring the test situation, (ii) learning within the test, or (iii) metacognitive intervention. Sternberg and Grigorenko (2002) propose an additional fourth category, namely *training a single cognitive function*, which represents the focus of the present review. Examples of this approach include DA of constructs such as working memory and phonological awareness (Sternberg & Grigorenko, 2002). Along different lines, Caffrey et al. (2008) distinguish between clinically-oriented DA which seeks to measure and remediate poor cognitive functioning, and research-oriented DA which focuses solely on measurement of a particular skill in a standardised and time-limited fashion.

Despite the purported advantages of dynamic tests to measure learning potential, their adoption among educational psychologists is low (Hill, 2015). DA has been criticised for its concept fuzziness, time-consuming nature, and questionable psychometric properties, particularly in relation to the reliability and interpretation of pre-post-test gain scores (Grigorenko & Sternberg, 1998). These criticisms notwithstanding, DA has been shown to offer advantages in the prediction of future academic performance. In a mixed methods review of 24 studies, Caffrey et al. (2008) considered the predictive validity of dynamic tasks of nonverbal reasoning, language, and working memory. Although static and dynamic tests correlated similarly with various achievement measures, dynamic measures offered additional advantages in their ability to identify children likely to respond to instruction, to classify bilingual children with and without language disorder, and to explain variance in outcomes over and above that explained by static tests.

DA is concerned with learning potential, and therefore it is particularly appropriate in the study of reading which is itself a learning process. In recent years there has been increasing interest in DA of reading, with a number of studies converging in areas of instrument design, analytical procedure, and research questions. However, at the time of writing we are unaware of any systematic reviews of studies using DA to classify children with or at risk of RD. The current review therefore uniquely contributes to the literature by examining the ability of DA of reading and reading-related constructs to correctly classify children with or at risk of RD, using rigorous methods to assess the quality of studies within.

Present Study

We conducted a systematic review to answer two research questions. Our first research question concerned the extent to which DA of reading and reading-related constructs is able to accurately identify children, concurrently or longitudinally, with or at risk of RD. We sought to answer this question through the extraction and comparison of classification metrics such as sensitivity, specificity, and AUC values. Where studies explicitly compared the classification accuracy of static and dynamic measures, we examined the additional predictive accuracy offered by DA. Our second research question considered the role of potential moderating factors on the ability of DA to classify RD. Firstly, given that DA is purported to be particularly appropriate for children from CLD backgrounds, we considered differential classification accuracy for bi-/multilingual children compared to monolingual peers. Secondly, DA has many possible methodological configurations and learning potential may be operationalised in different ways. Therefore, we considered the prevalence of different DA administration formats across the different domains of reading and reading-related skills targeted (pretest-teach-posttest and graduated prompts; see below). Thirdly, orthographies vary in the consistency of speech to sound mappings, with transparent orthographies primarily consisting of one-to-one relationships, and opaque orthographies

containing numerous one-to-many and many-to-one relationships (Ziegler et al., 2010). Given evidence for particular challenges posed by opaque orthographies in learning to read (Seymour et al., 2003), we also considered orthography as a moderating factor.

Method

Literature Search

We searched the electronic databases PsycInfo, Web of Science, ERIC, LLBA, and Medline (14/5/2021) using the following search terms: (child* OR under-18) AND ("dynamic assessment" OR "dynamic test" OR "dynamic task" OR "mediated learning" OR "mediated assessment" OR "interactive assessment" OR "testing the limits" OR "learning potential") AND ("read* dis*" OR "read* impair*" OR "read* delay*" OR "read* difficult*" OR "dyslex*" OR "read* comprehension impair*" OR "read* comprehension difficult*" OR "poor comprehender*"). In addition to database searches, we conducted backward citation searches of reference lists of all articles included at the full-text screening phase.

Inclusion Criteria

To be considered for inclusion in the review, articles had to conform to the following criteria: (a) uses a dynamic assessment of reading or a reading-related skill to classify participants as at-risk or not-at-risk of RD, defined according to reading performance either concurrently or longitudinally (studies merely comparing the performance of different groups of good and poor readers on dynamic assessments were excluded); (b) reports empirical data and appropriate statistical information for determining classification accuracy such as area under the curve, sensitivity, specificity or associated metrics (studies seeking only to investigate the statistical reliability of dynamic assessments were excluded); (c) participants aged 18 years or under; (d) peer reviewed and published in English (with no restrictions on the language of the assessment itself). Note that dynamic assessment was operationalised as any testing procedure within which explicit teaching and/or feedback was provided, which

participants were given the opportunity to act upon (e.g., repeated attempts at the same stimuli or application of a teaching phase to novel items). We imposed no selection criteria on publication date, sample characteristics, or the way in which studies classified RD. Results of the literature search are presented in Figure 1.

----Figure 1 here----

Coding of Studies

Studies were coded for information about research design (cross-sectional or longitudinal) and demographics (country, sample size, age, gender, second language learner status, and orthography). DA procedures were coded according to the construct in which participants were trained, format of administration, and computerisation. As discussed in the introduction, DA exists in multiple instantiations: given the exclusive focus of the present review on reading, all studies synthesised herein are characterised by Sternberg and Grigorenko's (2002) *training a single cognitive function* category of DA, and all are considered *research-oriented* (Caffrey et al., 2008). The DA format of each study was coded as *pretest-teach-posttest* (PTP) if it employed minimally a training and posttest phase and provided feedback during the training phase. DA format was coded as *graduated prompts* (GP) if it fulfilled the criteria for PTP but also employed a graduated set of hints for each incorrect response and incorporated the number of prompts required into the operationalisation of learning potential. All screening and data extraction was carried out independently by the first two authors. For database searches, agreement reached 97.6% for title and abstract screening and 100% for full-text screening; for backward citation searches, agreement for full-text screening reached 97.1%. Each author extracted data from 50% of the

studies in the final sample, with the remaining 50% being checked for accuracy by the other author. Disagreements at each stage were logged and resolved through discussion.

Quality Assessment

All studies were critically appraised according to the Quality Assessment Tool for Studies of Diverse Designs (QATSDD) instrument (Sirriyeh et al., 2012). The QATSDD contains 16 quality indicators such as *explicit theoretical framework* and *fit between research question and method of analysis* which are scored on a 4-point Likert scale from 0 (no mention at all) to 3 (described in full). Final scores here are based on the 14 indicators relevant to quantitative studies (no qualitative studies were included in the review). Study quality is expressed as a percentage out of a maximum score of 42. The quality of each study was assessed independently by the first two authors, yielding a weighted Kappa statistic of .738 ($p < .01$), representing *substantial* agreement (Landis & Koch, 1977). Applying the methodology of Murphy & Unthiah (2015), for all disagreements within 1 point we selected the lower of the two scores. Disagreements of 2 points or more were discussed and resolved by the first two authors. Mean study quality was judged to be 66.3% (min = 50%; max = 73.8%).

Results

An initial search yielded a total of 959 database records and 2 records arising from backward citation searches (see Figure 1). After removing duplicate records, screening of abstracts and full-text articles resulted in 14 articles eligible for inclusion in the review, published between 1994 and 2020 and representing 15 individual studies (Table 1). The majority of studies followed participants longitudinally (n = 12; 80%), typically administering dynamic measures at the first time point alongside other static predictors to predict future RD status. Studies were carried out mostly in the USA (n = 10; 67%), followed by Denmark (n = 3; 20%), Canada (n = 1; 7%), and the Netherlands (n = 1; 7%). The age

range of children was 5 years 5 months to 10 years 9 months; however, some studies did not report participant age range, instead opting for school year (see Table 1). Three studies reported that all participants were monolingual, seven reported some proportion of children acquiring a second or additional language in the home (ranging from 35% to 100%), and the six remaining studies did not report language learner status. Where data are reported, samples largely consisted of equal proportions of male and female participants. Median sample size was 158 (range: 57–600).

DA protocols trained children in decoding (n = 8; 53%), phonological awareness (PA; n = 5; 33%) and working memory (n = 2; 13%). Eight studies employed a graduated prompts (GP) format, while the remaining seven employed a pretest-train-posttest (PTP) format. For the most part, DAs were administered in a pencil-and-paper format (n = 13; 87%), with only two studies employing computerised DA. Study characteristics, along with summaries of DA procedures and methods for RD classification can be found in Table 2 in supplementary material. Results are presented below according to the construct targeted by DA (decoding, PA, and working memory). Subsequently, the moderating factors of language status, administration format, and orthography are considered.

---Table 1 here---

Construct Targeted by Dynamic Assessment

Decoding

Eight studies examined DA of decoding for the purposes of classification. Of the seven longitudinal studies in this sample, the majority administered DAs in the earliest stages of reading instruction in kindergarten, either following children to the end of the first grade (Gellert & Elbro, 2017b; Petersen et al., 2016; Petersen & Gillam, 2015), end of second grade (Gellert & Elbro, 2018), or in one case as far as the end of fifth grade (Petersen et al., 2018). In contrast, two longitudinal studies administered DAs later on, in first grade, either

following children until the end of the first grade (Cho et al., 2020) or the end of second grade (Compton et al., 2010). Three studies used the Predictive Early Assessment of Reading and Language (PEARL; Petersen & Gillam, 2015; Petersen et al., 2016, 2018), in which children are taught to decode four nonsense words. Similarly, Compton et al. (2010) taught children to decode nonsense words, in this case using three different decoding strategies (see Table 2). Four of the studies administered DAs using novel orthographies unfamiliar to participants (i.e., different to that of the language of school instruction), including Hebrew (Aravena et al., 2018), Mandarin (Cho et al., 2020), and novel letter shapes (Gellert & Elbro, 2017b, 2018). Typically, studies classified children as at-risk for RD on the basis of performance on word accuracy and/or fluency composites, though it should be noted that certain studies oversampled at-risk participants (Aravena et al., 2018; Compton et al., 2010; Gellert & Elbro, 2017b, 2018; Petersen & Gillam, 2015) and therefore cut-off criteria are not directly comparable.

As a group, DA of decoding studies reported moderate to high classification accuracy of their overall logistic regression models in terms of AUC, sensitivity, and specificity. Where a dynamic decoding variable was entered alone or with other predictors, models typically yielded AUCs of .8 or above, and sensitivity and specificity similarly at or above 80% (Cho et al., 2020; Gellert & Elbro, 2017b, 2018; Petersen & Gillam, 2015; Petersen et al., 2016, 2018). Crucially, a number of studies reported higher classification accuracy where a dynamic predictor was entered into a model *after* static measures of letter knowledge, naming fluency and PA, with AUCs in the .8 to .9 *excellent* range (Compton et al., 2010; Gellert & Elbro, 2017b, 2018; Petersen et al., 2018). Note that Aravena et al. (2018) and Petersen and Gillam (2015) were the only decoding studies not to add static measures to their classification models. Even in the most comprehensive batteries, DA variables were found to be significantly and uniquely predictive of RD status, for instance in Compton et al. (2010)

when entered in a model after word identification fluency, rapid digit naming, PA, and oral vocabulary (AUC = .953, sensitivity = 90.7%) and in Gellert and Elbro (2017b) after letter knowledge, PA, rapid naming, early word reading, vocabulary, and non-verbal IQ (AUC = .89, sensitivity = 80%).

Some other interesting results emerged from DA of decoding studies. Firstly, Compton et al. (2010) contrasted the classification accuracy of a baseline test battery containing only static measures with three alternative batteries containing either 5-week progress monitoring screeners or a single DA of decoding. All three alternative batteries yielded significantly higher classification accuracy than the baseline model, and these results speak to the underlying conceptual similarity of RTI and DA frameworks, both measuring children's response to teaching but on different timescales (Grigorenko, 2009). Secondly, Petersen et al. (2018) found the predictive power of DA to be developmentally constrained: for a 'Caucasian' (hereafter 'White') subgroup, the unique variance accounted for by kindergarten dynamic decoding score when entered after static measures of letter identification and PA decreased over time and the DA was not a statistically significant predictor of RD status after Grade 2. Along similar lines, Gellert and Elbro (2018) found a decreasing role of dynamic decoding score in predicting future reading difficulties among their Danish-speaking sample. In this study, a DA of decoding was administered both before and after the onset of formal reading instruction (the end of kindergarten and November of first grade, respectively); scores from both time points were significantly and uniquely predictive of reading accuracy difficulties in second grade, but the contribution of the DA was stronger when it was administered in kindergarten (3%–5% unique variance) than in the first grade (1%–3% unique variance), as static predictors of word reading, letter knowledge, PA, and RAN accounted for more variance over time.

Phonological Awareness (PA)

Five studies employed dynamic measures of PA as predictors of RD status (see Table 2). Bridges and Catts (2011) evaluated the classification accuracy of a dynamic phoneme deletion task (Dynamic Screening of Phonological Awareness; DSPA) among an unselected sample of kindergarten children (Study 1) and a more diverse kindergarten sample in which half of the children were judged to be at-risk of later RD (Study 2). The DA was administered at the beginning of kindergarten and was used alongside static measures of phoneme deletion (Study 1) or Initial Sound Fluency (Study 2) to predict which children would score below the 25th percentile on static outcome measures of word reading accuracy or nonword reading fluency approximately seven to eight months later. Logistic regression models containing only the DA resulted in higher AUCs than models containing only a static predictor (e.g. in Study 1: AUC = .61 when using only a static phoneme deletion task, and AUC = .69 when using DSPA score). As a result, classification was only improved marginally by the combination of static and dynamic scores, finally reaching AUCs of .69 to .76 when RD was defined in relation to word reading accuracy, and AUCs of .77 to .83 when RD was defined in relation to nonword reading fluency. Models attained relatively lower sensitivity in Study 1 (57%–58%) than in Study 2 (85%–92%) with a more diverse and at-risk sample.

O'Connor and Jenkins (1999) administered static measures of oral language and PA at the beginning of kindergarten and again at the start of the first grade, this time with the addition of a dynamic phonemic segmentation task in which children are taught to segment words into onsets and rimes. Measures administered at this second time point were used to predict RD status by the end of the first grade. In discriminant function analysis, the number of trials required to reach mastery emerged as a significant predictor alongside static tests of rapid letter naming and sound repetition. The authors explicitly contrasted the classification accuracy of different predictors. Relative to a baseline model achieving sensitivity of 29.7%, the substitution of dynamic for static total PA score only improved sensitivity slightly

(30.6%; metrics calculated manually by review authors). In contrast, a much larger increase in sensitivity was observed with the substitution of dynamic total score for the number of learning trials required to reach mastery (52.7%), reducing the number of false positives from 26 to 9. Specificity in all models remained at or near 100%.

Krenca et al. (2020) evaluated the classification accuracy of a lexical specificity training game, administered at the beginning of the first grade, in a sample of English–French emergent bilingual children in Canada. During the game, children were shown plates of four pictures; two unfamiliar minimal pair targets, one unfamiliar control and one familiar control. Children were then asked to "show me the [target]". RD status was determined in spring of the first grade by a score under the 25th percentile on a composite measure of word reading accuracy and fluency in English and in French. In logistic models containing nonverbal reasoning and French static PA predictors, only English and not French dynamic lexical specificity score predicted at-risk status for RD in French. This final model yielded a higher AUC than a static-only model (.87 compared to .83), a 15.3% increase in sensitivity to 53.8%, but a slight reduction of 2.3% in specificity to 93.2%.

Finally, Gellert and Elbro (2017a) readministered incorrectly answered items from a static phoneme identification task using a GP procedure in November of kindergarten in Denmark (children were aged 6;4). Logistic regression models predicted RD status at two subsequent time points (November and May of first grade) using static PA, letter knowledge, and dynamic PA score (number of prompts required), entered last. The predictive power of the DA appeared to be developmentally constrained: in the early stages of first grade, RD status was uniquely predicted by DA score (6%; AUC = .71) but this was not the case six months later, where static measures had become more powerfully predictive.

Working Memory

Two cross-sectional studies used the Swanson-Cognitive Processing Test (S-CPT) to classify RD in samples of older children (10 years 6 months to 10 years 9 months) (Swanson, 1994, 1995 Study 2; see Table 2). The S-CPT is a lengthy information processing DA consisting of 11 subtests of verbal and nonverbal working memory. Aside from a static initial score which represents pretest performance without assistance, this DA yields several indices of processing potential, including: gain score (achievement with assistance); probe score (number of prompts required to achieve the gain score); maintenance score (achievement after probes are no longer available); processing difference score (difference between gain and initial score), and stability score (subtracting initial score from maintenance score).

Both studies employed discriminant function analysis with DA variables as predictors, yielding moderate to poor classification accuracy as indicated by Wilk's lambda (Λ), where $\Lambda = 1$ indicates total lack of classification. In Swanson (1994), using a subset of four DA measures, initial score explained most variance in RD status, while the only significant dynamic predictors were probe score ($\Lambda = .45$) and processing difference score ($\Lambda = .40$), accounting for 21% and 10% of variance, respectively. The three S-CPT scores from this analysis correctly classified only 3.9 to 15.4% of children with RD. In Swanson (1995) using the full DA, only gain score was significantly predictive of RD status, though accounting for little variance (5%) and with very low discrimination ($\Lambda = .95$). While these studies do speak to the uniquely predictive status of certain dynamic working memory scores over and above initial (static) scores, they do not contrast DA variables with traditional static predictors of reading and as a result it is not possible to glean the unique contribution of DA here.

Moderator Variables

We turn now to potential moderating factors in the classification accuracy of DA of reading and reading-related skills including language learning status, DA administration format, and orthography.

Language Learning Status

A small number of studies included bi- or multilingual children in their samples. Three studies recruited diverse samples of children but did not disaggregate analyses by language group (Gellert & Elbro, 2017a, 2017b, 2018), and two studies recruited exclusively bilingual children (Krenca et al., 2020; Petersen & Gillam, 2015). Note that studies only administered DAs in the language of school instruction or in an orthography unfamiliar to all participants, and not in bi-/multilingual children's first or home languages.

Only two studies explicitly compared the prognostic value of DA across both groups of children. Petersen et al. (2016) administered the PEARL, a dynamic measure of decoding, at the beginning of kindergarten to classify children with RD at the end of the first grade according to performance in reading fluency. Separate analyses were run for the entire sample and for a subgroup of 300 Hispanic students, 77% of whom were classed as English language learners. For these latter students, the static reading accuracy autoregressor resulted in considerably lower sensitivity (33-36%) than for the whole sample (50-51%), suggesting that this static measure yielded low classification accuracy for all participants in general, and very poor classification accuracy for English language learners in particular. Although the addition of the DA did not increase specificity over the 80% threshold for the Hispanic group, its increase of 34–40% still represented a meaningful improvement given such a low baseline with the static composite alone. This, together with the finding that the dynamic composite also yielded relatively higher sensitivity in the Hispanic group (87%–100%) than in the whole sample (69%–92%), suggests that the DA did offer higher classification accuracy for CLD students.

These findings were extended longitudinally by Petersen et al. (2018) who followed the same participants from Grades 2 to 5. Results again speak to differential prognostic value of the DA by language group: after accounting for static measures of PA and letter identification

in kindergarten, the DA was significantly and uniquely predictive of RD status only in Grade 2 for White students, but continued to be so for Hispanic students at every time point until Grade 5. Indeed, for White students the unique variance accounted for by the dynamic test decreased from 7% to 2% between Grade 2 and Grade 5, while the opposite pattern was observed for Hispanic students, rising from 5% to 14%. Additionally, similar to findings in Petersen et al. (2016), static tests throughout the grades afforded relatively lower sensitivity to Hispanic students, such that increases provided by dynamic scores were more meaningful, particularly in later grades. While AUCs in the White group consistently passed the acceptable level in static testing-only models, for Hispanic students this level was only ever reached once dynamic scores were included, again indicating the relatively larger classification enhancements offered by DA to linguistically diverse students.

Dynamic Assessment Format

DA assessment formats were fairly equally represented across studies (PTP, $n = 7$; GP, $n = 8$). No particular patterns emerged between the two formats in terms of classification accuracy, although not all studies reported comparable metrics such as AUC values. The PTP format was particularly common in studies examining decoding (5 out of 8 studies), while GP was relatively more common in studies examining phonological awareness (3 out of 5 studies). To some extent, this may be due to task demands, as GPs may be more appropriate for the training of metalinguistic skills; on the other hand, elaborated hints are less necessary for the learning of sound-symbol correspondences often found in decoding studies in which corrective feedback need only be binary in nature. It was also noted that both of the computerised DAs employed a PTP format (Aravena et al., 2018; Krenca et al., 2020).

Orthography

In total, 11 of the 15 studies in the sample conducted DA in English, whereas four studies (representing half of the eight dynamic decoding studies) taught children to read in an

unfamiliar orthography, i.e., different to that of the language of school instruction (Aravena et al., 2018 in Hebrew; Cho et al., 2020 in Mandarin; Gellert & Elbro 2017b, 2018 using novel shapes). These were either real non-alphabetic orthographies (Hebrew and Mandarin) or novel symbols. As a result, these DA procedures may be considered purer measures of learning potential by virtue of the fact that children had no experience with the orthography in which they were taught (i.e., children could not bring pre-existing knowledge to the task as they could with DA tasks using English orthography). Comparison with the remaining four studies using a familiar orthography (all in English) showed no discernible pattern in terms of classification accuracy achieved by statistical models (AUCs ranged from 0.83 to 0.95 and 0.74 to 0.96 in dynamic decoding studies using familiar and unfamiliar orthographies, respectively).

Discussion

Dynamic assessment is an alternative approach to static testing that measures an individual's potential to learn new skills when given assistance; as such, it is proposed to be a more sensitive and less biased approach to the identification of RD, particularly among children from CLD backgrounds. The purpose of this systematic review was to examine the ability of DA to classify children with or at risk of RD, and the factors that may moderate its ability to do so. The present review therefore makes a timely and unique contribution to the literature on DA and reading. Fourteen peer-reviewed papers, representing 15 individual studies written in English, published between 1994 and 2020, and reporting empirical data collected on samples of school-aged children were eligible for inclusion in the current review. The majority of studies were published in the USA and employed DA of early code-based skills (PA and decoding), while PTP and GP administration formats were fairly evenly represented. The majority of the studies in the sample classified RD according to a standard score or percentile cut-off on composite variables capturing performance on static norm-

referenced assessments of reading. Exceptions to this included one study which used an IQ-achievement discrepancy in its classification of RD (Swanson, 1995), and one study which defined RD as a score of –1 *SD* in terms of both growth and final level achieved in word identification fluency (a 'dual-discrepancy' criterion; Cho et al., 2020).

The extent to which DA lives up to its promise depends on its ability to predict RD status correctly, and this constituted our first research question. While variability in the choice of classification criteria across the different studies places limitations on the synthesis of results, some themes did emerge from the review. When considered as a group, studies provided evidence for the ability of DA to achieve good classification accuracy of RD either alone or in addition to traditional static tests. In some cases, models with a single dynamic learning predictor yielded classification accuracy higher than or very similar to models containing only static predictors; this was common in DA of decoding (Cho et al., 2020; Gellert & Elbro, 2018; Petersen et al., 2016, 2018), though was also found for DA of PA (Bridges & Catts, 2011). As well as offering improvements relative to static-only models, some dynamic-only models represented good classification accuracy in an absolute sense, for instance achieving AUCs in the *good* to *excellent* range and/or sensitivity above 80% (Cho et al. 2020; Gellert & Elbro, 2018; Petersen et al. 2018, 2016). However, dynamic-only models were not seen to offer a complete substitution for static predictors, as models combining both types of predictors did achieve marginally higher classification accuracy. This finding supports the value-added nature of static tests for the purposes of classification of later RD (Bridges & Catts, 2011; Gellert & Elbro, 2018; Petersen et al., 2016, 2018).

Our second research question concerned potential moderating factors in the classification accuracy of DA. In terms of the construct targeted by DA, we found studies examining decoding, PA, and working memory. Studies using dynamic decoding tasks tended to produce the highest classification accuracy, with only Aravena et al. (2018)

reporting an AUC below .8 (though two of the eight studies did not report AUC values). Despite this exception, learning potential scores from dynamic decoding tasks were generally shown to be uniquely predictive of RD status over and above the contribution of static reading-related measures, in some cases in the context of rather comprehensive test batteries and after accounting for an autoregressor (Compton et al., 2010; Gellert & Elbro, 2017b, 2018; although see Cho et al., 2020). To some degree, this may be due to the proximity of dynamic decoding tasks to reading outcome measures used to classify children with or at risk of RD. In dynamic decoding procedures, children are required to apply and synthesise sound-symbol correspondences to read words (in the case of DAs which use the language of instruction) or nonwords (in the case of DAs which use a novel orthography), a skill which is similarly required in standardised static assessments of word reading which are used to determine risk for RD, such as the Woodcock Reading Mastery Test and Dynamic Indicators of Basic Early Literacy Skills.

Two dynamic decoding studies hinted at the developmentally constrained nature of static and dynamic predictors of later RD status. In Petersen et al. (2018), a dynamic decoding task administered in kindergarten was not significantly or uniquely predictive of RD status beyond second grade for a White subgroup, and in Gellert and Elbro (2018), a dynamic decoding task administered earlier (in kindergarten) was a stronger unique predictor of RD status in second grade than a dynamic task administered later (after the onset of explicit literacy instruction in first grade). One reason for this may be floor effects commonly found in static measures of reading-related skills when administered before or around the onset of explicit literacy instruction (Catts et al, 2009). This is an issue circumvented by DA of decoding which measures children's potential to learn novel symbol-sound correspondences and decoding rules rather than pre-existing knowledge. Indeed, static screening has been found to be more accurately predictive of future RD status when

administered later rather than earlier (e.g., in first grade rather than kindergarten; Poulsen et al., 2017), an effect which may be attributable to the lessening of floor effects. Another reason may be the tendency of static measures to account for more variance over time: though this appeared to be the case in Gellert and Elbro (2018), this pattern is far less clear-cut in Petersen et al. (2018) in which variance explained by static measures in the Hispanic subgroup either decreased or remained stable over time. Nevertheless, the current review provides tentative support for the relatively stronger predictive power of DA of decoding when administered in the earlier stages of reading instruction (Gellert & Elbro, 2018) or in the prediction of RD status up until the second grade (Petersen et al., 2018). This early predictive power of DA of decoding is likely to be beneficial for the purposes of identifying children who may benefit from early reading intervention.

Dynamic assessments of PA targeted phoneme identification (Gellert & Elbro, 2017a), segmentation (O'Connor & Jenkins, 1999), and deletion (Bridges & Catts, 2011), as well as training in minimal pairs (Krenca et al., 2020). Conclusions regarding the classification accuracy of DA in this context are more challenging given the small number of studies, differing criteria for RD, and relatively small sample sizes (with the exception of O'Connor & Jenkins, 1999). Nevertheless, dynamic measures of PA were found to be uniquely predictive of later RD status in all five studies. Bridges and Catts (2011) reported appreciably higher AUCs for their dynamic phoneme deletion score when RD was defined according to nonword reading fluency: this finding may relate to the availability of compensatory strategies of poor readers such as a reliance on declarative memory in real word reading tasks (Ullman & Pullman, 2015), a strategy unavailable in nonword reading measures.

We found two studies examining the classification accuracy of DA of working memory (Swanson, 1994, 1995). Both provided evidence for the uniquely predictive nature of dynamic processing potential scores (probe and processing difference score in Swanson 1994,

and gain score in Swanson 1995), though analyses revealed only poor to moderate classification accuracy. Although the predictive accuracy of DA of WM was not compared with static measures, results serve to pinpoint the active ingredient of DA in indicating that *change* in performance (when given assistance) significantly predicted at-risk status. Working memory is a known predictor of reading skill (Peng et al., 2018), though the poor classification accuracy of the DA may be due to its relatively more distal relationship with reading than other dynamic tests which target decoding and PA.

Moderating Effects

Dynamic assessment of bilingual children in particular should be considered a useful prognostic tool, as static assessments of language may fail to consider external factors such as children's language exposure and usage in the home, leading to possible misdiagnosis of reading disorders (Petersen & Gillam, 2015). Few studies in this review included bi- or multilingual children in their samples, and only two studies explicitly compared the classification accuracy of DA for these children relative to their monolingual peers. In these studies, DA of decoding did provide higher sensitivity for a mostly bilingual Hispanic subgroup (Peterson et al., 2016) and was uniquely predictive compared to a subgroup of White children (Peterson et al., 2018). That other studies containing linguistically diverse participants did not explicitly compare classification accuracy for mono- and bi-/multilingual children potentially speaks to the practicalities of recruiting the large sample sizes necessary to perform sufficiently powered classification models between groups. This is an area that deserves more attention, and we recommend that future studies recruit diverse samples of children from different language backgrounds to better understand the prognostic capabilities of DA for children from differing linguistic environments.

PTP and GP administration formats were fairly evenly represented among the studies included in this review. Given no particular pattern of classification accuracy, results appear

to suggest that choice of DA administration format may be determined by task demands: in particular, PTP was found more commonly among decoding studies in which children were trained in novel sound-symbol correspondences (though GP was used when more complex decoding strategies were involved; Cho et al., 2020 and Compton et al., 2010). On the other hand, GP was more commonly utilised in studies targeting PA, a metalinguistic skill amenable to a more intensive form of examiner input and feedback.

Lastly, we considered orthography as a moderator variable given that variation in orthographic depth has been shown to affect the rate of reading acquisition (Seymour et al., 2003). For the most part, studies carried out DA in the language and orthography of school instruction (for the most part in English using the English (Latin) alphabet), though four studies taught children to decode in an unfamiliar orthography. No patterns emerged from this review to suggest that DA in an unfamiliar orthography resulted in higher classification accuracy of risk for RD. Future work may seek explicitly to compare the prognostic accuracy of DA carried out in the same orthography of school instruction versus a completely novel orthography, particularly for children from CLD backgrounds.

Study Quality and Cost-Effectiveness

The present review is advantageous in taking account of study quality according to the QATSDD tool (Sirriyeh et al., 2012). While the mean study quality was judged to be 66.3%, one consistently low scoring category was *evidence of sample size considered in terms of analysi*s. Sample size varied considerably across studies, with four studies having fewer than 100 participants (Bridges & Catts, 2011 Studies 1 and 2; Krenca et al., 2020; Petersen & Gillam, 2015). Given that the prevalence of word-level RD in the general population is estimated to be 5-17% (Grigorenko et al., 2020), there may be concern that studies with such small sample sizes are not adequately powered to detect true positive cases of RD. For a minimum sensitivity of 80% and a population RD prevalence rate of 10%, Bujang and Adnan

(2016) recommend a minimum sample size of approximately $n = 200$ (with 20 affected and 180 unaffected cases). We note that 10 of the 15 studies in the present review recruited samples smaller than this. Additionally, larger and more representative samples of unselected children are necessary to test DAs in the general population, and initially promising results should be cross-validated across different samples (Jenkins et al., 2007). Issues related to sample size represent a novel finding in the literature of DA and reading skills, and we recommend that future studies recruit large enough samples in order to detect RD with sufficient statistical power.

A number of studies in the present review raised concerns regarding the cost-effectiveness of DA. Amongst the studies included in this review, DA procedures took between approximately 5 and 30 minutes to administer. It should be noted that, although in some cases DAs added substantially to the length of standardised screening batteries, this still represented an advantage over progress monitoring which typically lasts several weeks or longer. Nevertheless, certain studies questioned whether the time taken by DA justified relatively small increases in predictive accuracy (e.g., O'Connor & Jenkins, 1999). On the other hand, some very short DA procedures appeared to result in significantly improved accuracy particularly for decoding (Gellert & Elbro, 2017b, 2018; Petersen et al., 2016, 2018). Only two studies in this review used computerised DAs (Aravena et al. 2018; Krenca et al. 2020). Given the time-consuming administration and complex scoring procedures required by many DA procedures, computerisation and automated reporting methods may be attractive, particularly for improving usability among educational practitioners and reducing human error. This is a potentially fruitful domain for further dynamic testing research.

Limitations and Directions for Future Research

There are a number of limitations of the present review. Firstly, the inclusion criteria we imposed resulted in a pool of only 15 eligible studies, despite having no restriction on

publication date or country of origin (albeit papers had to be written in English). This suggests the current research literature for DAs to classify RD is small. Additionally, the majority of studies reviewed originated from the USA, imposing limitations on the generalisability of results in terms of different populations, languages, and orthographies. Articles were only screened if they were peer-reviewed publications, however we acknowledge this may have led to a body of literature being excluded, including doctoral theses or dissertations. Unfortunately, due to resource constraints we were unable to implement a comprehensive grey literature search strategy, which future reviews may implement.

The limited sample of research papers did not allow us to answer all of our research questions fully for two reasons: firstly, some studies did not explicitly contrast static with dynamic predictors, meaning it was not always possible to determine unique variance accounted for by DA, and secondly, few studies explicitly compared classification accuracy for bi-/multi- and monolingual children. We also found inconsistency across reporting practices in terms of model classification accuracy: in particular, some studies did not report sensitivity or specificity, and in most cases it was not possible to manually calculate these metrics as studies did not provide confusion matrices indicating the raw number of true positives, false positives, and so on. Relatedly, not all studies reported AUC values, posing limitations on the potential for statistical synthesis across different samples. We therefore recommend that future studies explicitly contrast the classification accuracy provided by static as well as dynamic measures, and whether higher accuracy is achieved with samples of children from CLD backgrounds. Of additional interest is whether DA administered in a bi-/multilingual student's first or home language may accurately predict future risk of RD. Such information may ultimately have implications for educational practitioners, providing guidance as to whether to invest time and resources in DA as a screener for risk of later RD.

Some evidence emerged concerning the developmentally constrained nature of DA. In particular, one Danish study (Gellert & Elbro 2017a) found that a DA of phonological awareness administered at the beginning of kindergarten was a significant predictor of RD risk by the beginning of first grade but not by the end. However, Danish children begin school relatively later than children in many other countries including the USA and the UK, and therefore future studies should seek to validate this finding by administering DA at the beginning and end of the first year of formal instruction in countries where children are younger and less cognitively mature when they start school. Here it is of interest whether DA provides an advantage over the floor effects often produced by static measures, and can ultimately be used to identify RD risk as early as possible.

Finally, studies in the present review focused for the most part on word-level outcomes such as reading accuracy and fluency as opposed to higher-level reading outcomes such as reading comprehension. Indeed, no studies in the review used DA to predict risk status for specific reading comprehension impairment (S-RCI), which is said to have a prevalence rate of around 5-11% in school-aged children (Kelso et al., 2020). To date there has been some work on the dynamic assessment of skills known to be predictors of reading comprehension performance, such as vocabulary learning (Gellert & Elbro, 2013; Nation et al., 2007; Petersen et al., 2020) inference making (Elleman et al., 2011) and sentence integration (Gruhn et al., 2020), though such studies have not employed DA for the purposes of classification of RD (or S-RCI specifically). Indeed, it may be hypothesised that such a passage- or discourse-level deficit would be better predicted by DA targeting the passage level itself, for example through inference making, sentence integration, or oral narrative, although DA of word learning may also yield good classification given the vocabulary weaknesses consistently found in children with S-RCI (Landi & Ryherd, 2017). Furthermore, given developmental changes in the relative importance of decoding and linguistic

comprehension over time, it may be of interest to compare the classification accuracy of such dynamically administered measures at different points in reading development.

Conclusions

We conducted a systematic review to examine the possible benefits of DA of reading and reading-related skills to the identification of children with or at risk of RD. Three particular trends emerged across the 15 studies we reviewed. Firstly, although some dynamic tests achieved classification accuracy similar to or even higher than static tests alone, the best accuracy was achieved by combining the two. In particular, dynamic tasks targeting skills more proximal to reading were better able to identify RD, with the best results seen for DA of decoding. Secondly, there is tentative evidence for the relatively stronger predictive power of DA in the earliest stages of reading acquisition; as a result, DA shows promise for the early identification of children at risk of RD, potentially avoiding the need to delay assessment until static measures become more reliable predictors. Thirdly, there is some evidence that DA of decoding produces relatively larger improvements in the classification accuracy of future RD for CLD children. We recommend further research to build on the promising results reported here, specifically well-powered studies with diverse samples of participants, and the use of dynamic testing of other reading and reading-related constructs to classify at-risk status for other types of reading difficulties such as specific reading comprehension impairment.

References

Asterisks represent studies included in the systematic review.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: Author.

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

Bailey, C. E., Manis, F. R., Pedersen, W. C., & Seidenberg, M. S. (2004). Variation among developmental dyslexics: Evidence from a printed-word-learning task. *Journal of Experimental Child Psychology*, *87*(2), 125–154. https://doi.org/10.1016/j.jecp.2003.10.004

*Bridges, M. S., & 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. https://doi.org/10.1177/0022219411407863

Burns, M. K., & Senesac, B. V. (2005). Comparison of dual discrepancy criteria to assess response to intervention. *Journal of School Psychology*, *43*(5), 393–406. https://doi.org/10.1016/j.jsp.2005.09.003

Bujang, M. A., & Adnan, T. H. (2016). Requirements for minimum sample size for sensitivity and specificity analysis. *Journal of Clinical and Diagnostic Research*, *10*(10), 1–6. https://doi.org/10.7860/JCDR/2016/18129.8744

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

Catts, H., Petscher, Y., Schatschneider, C., Bridges, M., & Mendoza, K. (2009). Floor effects associated with universal screening and their impact on early identification of reading

disabilities. *Journal of Learning Disabilities*, *42*(2), 163–176. https://doi.org/10.1177/0022219408326219

*Cho, E., Compton, D. L., & Josol Cynde Katherine. (2020). Dynamic assessment as a screening tool for early identification of reading disabilities: A latent change score approach. *Reading and Writing*, *33*(3), 719–739. https://doi.org/10.1007/s11145-019-09984-1

*Compton, D. L., Fuchs, D., Fuchs, L. S., Bouton, B., Gilbert, J. K., Barquero, L. A., Cho, E., & Crouch, R. C. (2010). Selecting at-risk first-grade readers for early intervention: Eliminating false positives and exploring the promise of a two-stage gated screening process. *Journal of Educational Psychology*, *102*(2), 327–340. https://doi.org/10.1037/a0018448

Dumas, D., McNeish, D., & Greene, J. A. (2020). Dynamic measurement: A theoretical–psychometric paradigm for modern educational psychology. *Educational Psychologist*, *55*(2), 88–105. https://doi.org/10.1080/00461520.2020.1744150

Elleman, A. M., Compton, D. L., Fuchs, D., Fuchs, L. S., & Bouton, B. (2011). Exploring dynamic assessment as a means of identifying children at risk of developing comprehension difficulties. *Journal of Learning Disabilities*, *44*(4), 348–357. https://doi.org/10.1177/0022219411407865

Fuchs, D., & Fuchs, L. S. (2006). Introduction to Response to Intervention: What, why, and how valid is it? *Reading Research Quarterly*, *41*(1), 93–99. https://doi.org/10.1598/RRQ.41.1.4

Gellert, A. S., & Elbro, C. (2013). Do experimental measures of word learning predict vocabulary development over time? A study of children from grade 3 to 4. *Learning and Individual Differences*, *26*, 1–8. https://doi.org/10.1016/j.lindif.2013.04.006

*Gellert, A. S., & Elbro, C. (2017a). Does a dynamic test of phonological awareness predict early reading difficulties? A longitudinal study from kindergarten through grade 1. *Journal of Learning Disabilities*, *50*(3), 227–237. https://doi.org/10.1177/0022219415609185

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

*Gellert, A. S., & Elbro, C. (2018). Predicting reading disabilities using dynamic assessment of decoding before and after the onset of reading instruction: A longitudinal study from kindergarten through grade 2. *Annals of Dyslexia*, *68*(2), 126–144. https://doi.org/10.1007/s11881-018-0159-9

Glover, T., & Albers, C. (2007). Considerations for evaluating universal screening assessments. *Journal of School Psychology*, *45*(2), 117–135. https://doi.org/10.1016/j.jsp.2006.05.005

Grigorenko, E. L. (2009). Dynamic assessment and response to intervention two sides of one coin. *Journal of Learning Disabilities*, *42*(2), 111–132. https://doi.org/10.1177/0022219408326207

Grigorenko, E. L., Compton, D. L., Fuchs, L. S., Wagner, R. K., Willcutt, E. G., & Fletcher, J. M. (2020). Understanding, educating, and supporting children with specific learning disabilities: 50 years of science and practice. *The American Psychologist*, *75*(1), 37–51. https://doi.org/10.1037/amp0000452

Grigorenko, E. L., & Sternberg, R. J. (1998). Dynamic testing. *Psychological Bulletin*, *124*(1), 75–111. https://doi.org/10.1037/0033-2909.124.1.75

Gruhn, S., Segers, E., Keuning, J., & Verhoeven, L. (2020). Profiling children's reading comprehension: A dynamic approach. *Learning and Individual Differences*, *82*. https://doi.org/10.1016/j.lindif.2020.101923

Haywood, H., C. (1997). Interactive assessment. In R. Taylor (Ed.), *Assessment of individuals with mental retardation*. Singular Publishing Group.

Haywood, H. C., & Lidz, C. S. (2007). *Dynamic assessment in practice: Clinical and educational applications (*pp.100-140*)*. Cambridge University Press.

Hill, J. (2015). How useful is dynamic assessment as an approach to service delivery within educational psychology? *Educational Psychology in Practice*, *31*(2), 127–136. https://doi.org/10.1080/02667363.2014.994737

Hosmer, D., Lemeshow, S., & Sturdivant, R. (2013). *Applied logistic regression* (3rd ed.). John Wiley.

Jenkins, J. R., Hudson, R. F., & Johnson, E. S. (2007). Screening for at-risk readers in a response to intervention framework. *School Psychology Review*, *36*(4), 582–600. https://doi.org/10.1080/02796015.2007.12087919

Kelso, K., Whitworth, A., Parsons, R., & Leitao, S. (2020). Hidden reading difficulties: Identifying children who are poor comprehenders. *Learning Disability Quarterly*. https://doi.org/10.1177/0731948720961766

*Krenca, K., Gottardo, A., Geva, E., & Chen, X. (2020). English phonological specificity predicts early French reading difficulty in emerging bilingual children. *Annals of Dyslexia*, *70*(1), 27–42. https://doi.org/10.1007/s11881-019-00188-4

Landi, N., & Ryherd, K. (2017). Understanding specific reading comprehension deficit: A review. *Language and Linguistics Compass*, *11*(2), 1–24. https://doi.org/10.1111/lnc3.12234

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Melby-Lervåg, M., Lyster, S.-A. H., & Hulme, C. (2012). Phonological skills and their role in learning to read: A meta-analytic review. *Psychological Bulletin*, *138*(2), 322–352. https://doi.org/10.1037/a0026744

Morrisroe, J. (2014). *Literacy Changes Lives 2014: A New Perspective on Health, Employment and Crime*. National Literacy Trust.

Murphy, V., & Unthiah, A. (2015). *A systematic review of intervention research examining English language and literacy development in children with English as an Additional Language (EAL)*. University of Oxford.

Nation, K., Snowling, M. J., & Clarke, P. (2007). Dissecting the relationship between language skills and learning to read: Semantic and phonological contributions to new vocabulary learning in children with poor reading comprehension. *Advances in Speech Language Pathology*, *9*(2), 131–139. https://doi.org/10.1080/14417040601145166

*O'Connor, R. E., & Jenkins, J. R. (1999). Prediction of reading disabilities in kindergarten and first grade. *Scientific Studies of Reading*, *3*(2), 159–197. https://doi.org/10.1207/s1532799xssr0302_4

Peña, E., & Halle, T. (2011). Assessing preschool dual language learners: Traveling a multiforked road. *Child Development Perspectives*, *5*(1), 28–32. https://doi.org/10.1111/j.1750-8606.2010.00143.x

Peng, P., Barnes, M., Wang, C., Wang, W., Li, S., Swanson, H. L., Dardick, W., & Tao, S. (2018). A meta-analysis on the relation between reading and working memory. *Psychological Bulletin*, *144*(1), 48–76. https://doi.org/10.1037/bul0000124

*Petersen, D B, Gragg, S. L., & Spencer, T. D. (2018). Predicting reading problems 6 years into the future: Dynamic assessment reduces bias and increases classification accuracy.

Language, Speech, and Hearing Services in Schools, *49*, 875–888. https://doi.org/10.1044/2018_LSHSS-DYSLC-18-0021

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

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

Petersen, Douglas B, Tonn, P., Spencer, T. D., & Fosterc, M. E. (2020). The classification accuracy of a dynamic assessment of inferential word learning for bilingual English/Spanish-speaking school-age children. *Language, Speech & Hearing Services in Schools*, *51*(1), 144–164. https://doi.org/10.1044/2019_LSHSS-18-0129

Poulsen, M., Nielsen, A., Juul, H., & Elbro, C. (2017). Early identification of reading difficulties: A screening strategy that adjusts the sensitivity to the level of prediction accuracy. *Dyslexia*, *23*(3), 251–267. https://doi.org/10.1002/dys.1560

Save the Children. (2015). *Ready to Read: Closing the gap in early language skills so that every child in England can read well*. https://www.savethechildren.org.uk/content/dam/global/reports/education-and-child-protection/ready-to-read-england.pdf

Schleicher, A. (2019). *PISA 2018: Insights and interpretations*. Organisation for Economic Co-operation and Development. https://www.oecd.org/pisa/PISA%202018%20Insights%20and%20Interpretations%20FINAL%20PDF.pdf

Seymour, P. H. K., Aro, M., & Erskine, J. M. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology*, *94*, 143–174.

Sirriyeh, R., Lawton, R., Gardner, P., & Armitage, G. (2012). Reviewing studies with diverse designs: The development and evaluation of a new tool. *Journal of Evaluation in Clinical Practice*, *18*(4), 746–752. https://doi.org/10.1111/j.1365-2753.2011.01662.x

Stanovich, K., & Siegel, L. (1994). Phenotypic Performance Profile of Children With Reading Disabilities: A regression-based test of the phonological-core variable-difference model. *Journal of Educational Psychology*, *86*(1), 24–53. https://doi.org/10.1037/0022-0663.86.1.24

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

Streiner, D., & Cairney, J. (2007). What's under the ROC? An introduction to Receiver Operating Characteristics Curves. *The Canadian Journal of Psychiatry*. https://doi.org/10.1177/070674370705200210

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

*Swanson, H. L. (1995). Effects of dynamic testing on the classification of learning disabilities: The predictive and discriminant validity of the Swanson-Cognitive Processing Test. *Journal of Psychoeducational Assessment*, *13*(3), 204–229. https://doi.org/10.1177/073428299501300301

Torgesen, J. K. (2000). Individual differences in response to early interventions in reading: The lingering problem of treatment resisters. *Learning Disabilities Research & Practice*, *15*(1), 55–64. https://doi.org/10.1207/SLDRP1501_6

Tzuriel, D. (2000). Dynamic assessment of young children: Educational and intervention perspectives. *Educational Psychology Review*, *12*(4), 385–435. https://doi.org/10.1023/A:1009032414088

Ullman, M. T., & Pullman, M. Y. (2015). A compensatory role for declarative memory in neurodevelopmental disorders. *Neuroscience and Biobehavioral Reviews*, *51*, 205–222. https://doi.org/10.1016/j.neubiorev.2015.01.008

Ziegler, J. C., Bertrand, D., Tóth, D., Csépe, V., Reis, A., Faísca, L., Saine, N., Lyytinen, H., Vaessen, A., & Blomert, L. (2010). Orthographic depth and its impact on universal predictors of reading: A cross-language investigation. *Psychological Science*, 21(4), 551–559. https://doi.org/10.1177/0956797610363406