Use of early word-reading fluency measures to predict outcomes on the Phonics Screening Check

Nicola Bell[1], Max Farrell-Whelan[1], Kevin Wheldall[1 2]

[1] *MultiLit Research Unit, MultiLit Pty Ltd, Macquarie Park, Australia.*

[2] *Emeritus Professor, Department of Educational Studies, Macquarie University, Macquarie Park, Australia.*

Disclosure statement: The authors hereby declare a financial interest in the outcomes of this study. All authors work for MultiLit Pty Ltd, which publishes the WARL and WARN assessments and the InitiaLit program. This was made clear in the Ethics application for the project and in the consent forms signed by school principals and parents.

The following is the post-print version of an article that has been peer reviewed and accepted for publication. It was published by SAGE Journals on 15 June 2020. Full reference details are below.

Bell, N., Farrell-Whelan, M., & Wheldall, K. (2020). Use of early word-reading fluency measures to predict outcomes on the Phonics Screening Check. *Australian Journal of Education, 64*(2), 161-176. https://doi.org/10.1177/0004944120931146

Abstract

Teachers in England and South Australia annually administer the Phonics Screening Check (PSC) to Year 1 students, with the purpose of identifying struggling readers. Students who do not meet the score threshold have not met the expected standard of word-decoding ability, meaning further support may be warranted. We sought to quantify the extent to which other early reading measures, such as the Wheldall Assessment of Reading Nonwords (WARN) and Wheldall Assessment of Reading Lists (WARL), predicted students' likelihood of not meeting PSC expected standards. Predicting PSC outcomes, and thereby identifying struggling readers in advance, has important implications for possible intervention strategies. Logistic regression and receiver operating characteristic analyses were conducted to examine the longitudinal relationships between real-word and pseudoword predictors as measured by the WARL and WARN and PSC pass/fail outcomes. Students who scored lower on predictors were less likely to meet the PSC expected standards. Results indicate that the WARL and WARN could be used to identify students who will not meet PSC expected standards, facilitating earlier intervention where it is most critically required.

Keywords

Struggling readers, reading difficulties, literacy, phonics, screening tests, early intervention

Use of early word-reading fluency measures to predict outcomes on the Phonics Screening Check

In early literacy development, comprehension of written text is highly constrained by limited word-level decoding abilities (Nation, 2019). Indeed, reading at any age cannot truly be said to take place without accurate word recognition. Evidence suggests that beginning readers with weak decoding skills may be difficult for classroom teachers to identify (Graham et al., 2020). Thus, to assist in identifying those students in need of additional literacy support, a Phonics Screening Check (PSC) was rolled out in England in 2012 (Standards & Testing Agency, 2012). No similar assessment measure has yet been implemented nationwide in Australia, although the South Australian government mandated the PSC in 2018, and the New South Wales government has announced a similar state-wide rollout in 2020. The purpose of the PSC, both in South Australia and in England, has been to screen and subsequently support children in Year 1 who have not met expected word-level decoding standards (Department for Education, 2018; Government of South Australia, 2019).

In the present study, two separate measures of real-word and pseudoword reading fluency were administered at the beginning of Year 1, when students were between 6 and 7 years old and had completed one year of formal schooling (i.e., Kindergarten). Here, reading fluency was defined narrowly, after Torgesen and Hudson (2006), as rate and accuracy in oral reading (see Wolf & Katzir-Cohen, 2001, for discussion). The predominant goal of this study was to determine whether the quick reading fluency measures could successfully predict which children would not meet the 'expected standard' score on the PSC, administered towards the end of the school year. While other Australian reading tests are already commonly used in Year 1 classrooms (e.g., Progressive Achievement Tests in Reading [Stephanou et al., 2008], York Assessment of Reading for Comprehension [Snowling et al., 2012]), these do not contain reading isolated real word and pseudoword stimuli, as does the

PSC. Ultimately, if the PSC is implemented nationwide in Australia, Year 1 teachers may consider it useful to know ahead of time which students are likely not to reach the PSC score threshold and thereby decide whether to deliver extra support in the intervening months.

Given the consistency with which reading outcomes are predicted by literacy precursor skills (Duff et al., 2015; Thompson et al., 2015) and demographic factors like socio-economic status (Buckingham et al., 2013), it would seem – at a group level – quite possible to predict the general trend of children's reading development. As such, one may question whether there is any additional information to be obtained from administering a pre-PSC reading assessment measure, beyond what is obtained from evaluating broad-stroke risk factors. However, many of these factors are not obvious. For example, some developmental patterns that foreshadow word-level decoding difficulties, such as delayed receptive language skills, are very hard to detect (Speech Pathology Australia, 2017, p. 19). In addition, there is limited information available regarding the specificity or sensitivity of school-entry speech and language screening tools, and the forms that such assessments take are not consistent across Australian states. Thus, from the perspective of a classroom teacher, predicting which students in Year 1 will struggle with learning to read is difficult, when based solely on knowledge of risk factors.

With particular respect to literacy development, there is also evidence that teachers frequently overestimate reading ability in their students (Bates & Nettelbeck, 2001; Graham et al., 2020). In the context of the PSC, South Australian (SA) teachers and school leaders participating in the 2017 trial observed that students performed more poorly than was expected (Hordacre et al., 2017). This feedback suggests that those students for whom performance was unexpectedly poor were not identified as such *before* the test was administered. Accordingly, teachers' estimations could achieve greater accuracy if influenced by reading performance results collected six months prior.

On the other hand, there is also likely to be a considerable amount of reading performance variability that emerges after the start of Year 1. Beyond the population-based positive and negative factors that influence reading development, the quality of literacy instruction to which students are exposed will likely vary, as will their individual responsiveness to such instruction. Hence, it is not clear whether – regardless of demography or literacy precursor skills – students who show reading difficulties at the beginning of Year 1 will still show these same difficulties at the end. This was the question addressed in the present study. In theoretical terms, a strong relationship between an individual's single-item reading fluency in Term 1 and his or her PSC performance in Term 4 may be taken as evidence of stability in word-level reading development during the intervening months. Alternatively, a weak relationship indicates that either the predictive assessment measures were not sensitive to the same underlying factors, or that students' literacy development was too much in flux to adequately predict development early on.

In 2011, a pilot trial of the PSC was conducted in England. The measure comprised 20 real words and 20 'pseudowords' (i.e., decodable and orthographically legal nonsense words) (Standards & Testing Agency, 2011). All items were phonically regular, meaning they could be decoded using the reader's knowledge of phoneme-grapheme correspondence conventions (Coltheart et al., 2001). The rationale behind including pseudowords in the PSC (and in other assessments with pseudoword stimuli) is that these items are entirely new to children undertaking the test. Performance is therefore representative of how well a reader decodes unfamiliar text, which is distinct from – but which underpins – the process by which a reader recognises familiar words by sight (Share, 1995).

In 2012, when the PSC was rolled out nationally in England, the threshold score denoting the expected standard was set at 32 points out of 40. This score was achieved or exceeded by approximately 58% of Year 1 English school students (Standards & Testing

Agency, 2012). Although the items in the English PSC have changed in the years following its pilot implementation, the structure of the test has not. Nevertheless, the percentage of children achieving or exceeding 32 points has increased substantially, with 82% of students meeting the expected decoding standard in 2018 (Department for Education, 2018). Improved performance over time is not the goal of PSC administration, although it does reflect a progressively greater emphasis on systematic phonics instruction in England more broadly (Stainthorp, 2020). Indeed, this was the main reason the test was introduced (Standards & Testing Agency, 2011, pp. 5-6).

Implementation of the PSC in England has not met with universal approval (e.g., Clark & Glazzard, 2018; UK Literacy Association). According to the majority of teachers surveyed by Clark and Glazzard, the results did not provide additional information beyond what could be gleaned from classroom observation. Interestingly, this finding clashes with the positive responses of most teachers and school leaders in the South Australian PSC trial (Hordacre et al., 2017). This difference of opinion reported by the self-selected samples in Clark and Glazzard (2018) and Hordacre et al. (2017) surveys may speak to the difference in educational policies between England and Australia. Irrespective of the political controversy surrounding the PSC, a large-scale study by Double et al. (2019) has found that students who 'pass' the screen achieve better reading comprehension in subsequent years, relative to those who do not 'pass'. Hence, the PSC may be used to identify those children whose decoding skills – which are foundational to reading comprehension – would benefit from remediation.

In South Australia, the PSC threshold was set at 28 points out of 40, both when the measure was trialled in 2017 and when it was rolled out in 2018. The difference in threshold scores between South Australia and England is due to the earlier timing of test administration (Buckingham & Wheldall, 2020). While children in England are generally administered the PSC approximately five or six weeks before the end of Year 1 (Standards & Testing Agency,

2018a), children in South Australia were administered the PSC with 15 or 16 weeks left in the school year (Government of South Australia, 2019). In 2018, 43% of Year 1 students achieved or exceeded 28 points out of 40 (Government of South Australia, 2019).

On a smaller scale than what has previously been published in government-funded reports from South Australia (Hordacre et al., 2017; Government of South Australia, 2019) and England (Department for Education, 2018), Wheldall et al. (2019) investigated the PSC results of a Year 1 cohort who received whole-class systematic synthetic phonics instruction. The majority of children in the Wheldall et al. study were also included in the present study, this time with the aim of determining whether PSC results could be predicted in advance. At the time of PSC administration, students in the present study were, on average, 6.9 years old. For clarity, this means they were similar in age to South Australian students who have previously participated in the PSC, but with exposure to between five and nine weeks more classroom instruction. In comparison to Year 1 students from England, students in the present study were approximately six months older, but with exposure to between three and four weeks less classroom instruction.

The predominant research question addressed in the present study was whether those students who do not reach the PSC expected standard could be identified (with sufficient sensitivity and specificity), based on real-word and/or pseudoword reading fluency skills. Given that the reading fluency measures were administered concurrently with PSC administration, as well as one and two terms in advance of PSC administration, we also sought to address whether the power of real-word and pseudoword reading predictors increased or remained stable over time.

Methods

Participants

A total of 137 children (64 females) from Year 1 classes in New South Wales participated in the present study. An in-depth analysis of the original sample's PSC results (n = 151) is available from Wheldall et al. (2019). The difference in sample sizes between the present article and that by Wheldall et al. is due to student absences on the dates of Term 1 and Term 3 testing (see 'Procedure'). For completeness, we only included those who were present at all three time points (i.e., n = 137). The mean age of participants at the time of testing in Term 4 was 6.9 years (SD = 4.7 months). All students were recruited from one of three school sites – two of which were separate campuses of the same college. According to information sourced from MySchool (Australian Curriculum, Assessment and Reporting Authority [ACARA], 2019), the first of the two schools had an Index of Community Socio-Educational Advantage (ICSEA) score of 1076, and the majority of students (84%) spoke a language other than English. For reference, an average ICSEA score is 1000, whereas 500 indicates extreme disadvantage and 1300 indicates extreme advantage (ACARA, 2011). The second of the two schools had an ICSEA score of 1182 and a minority of students (9%) spoke a language other than English. Overall, the participants in this study may be expected to have mixed language backgrounds and average or above-average levels of socio-educational advantage.

Ethics statement

Ethics approval was obtained from the Human Research and Ethical Committee at Macquarie University. Written consent was obtained from the principals of schools where testing took place. All parents also gave written informed consent for their children to participate in the research.

Tests and materials

Phonics Screening Check

A summary of the PSC is provided below (see Wheldall et al., 2019, for a complete description). The 2018 English PSC (Standards & Testing Agency, 2018b) instructs examinees to read aloud 20 real words and 20 pseudowords from a stimulus booklet. Students are given a score of '1' for an item if it is correctly pronounced with all phonemes blended together within a 10-second timeframe. Self-corrections are accepted as correct. All plausible pronunciation variants of pseudowords are accepted, while only the correct pronunciations of real words are accepted (with consideration given to children's speech impediments and accents).

Wheldall Assessment of Reading Lists (WARL)

The Wheldall Assessment of Reading Lists (WARL; Wheldall et al., 2015) is a curriculum-based measure (CBM), in which examinees are instructed to read aloud from a list of 100 real words as quickly and carefully as possible, within one minute. The WARL comprises three Initial Assessment lists ('A', 'B' and 'C') and 10 Progress Monitoring lists. For the present study, only the Initial Assessment lists were administered, and the raw score generated for each student was calculated as the average number of words (across 'A', 'B', and 'C') read correctly within the one-minute time-frame.

Wheldall Assessment of Reading Nonwords (WARN)

The Wheldall Assessment of Reading Nonwords (WARN) is a newly designed CBM, in which examinees are instructed to read from a list of 50 pseudowords as quickly and carefully as possible, within 30 seconds. Similar to the WARL, words are scored as 'incorrect' if mispronounced, omitted, or read after a pause of three or more seconds. Self-corrections are acceptable, and word-order reversals are counted as only one error. The WARN comprises three Initial Assessment lists ('A', 'B' and 'C') and 10 Progress Monitoring lists. For the present study, only the Initial Assessment lists were administered,

and the raw score generated for each student was calculated as the average number of words (across 'A', 'B', and 'C') read correctly within the 30-second timeframe.

Procedure

Students were withdrawn from their normal school class and taken to a quiet room on the school campus for testing. Testing took place at three separate timepoints. The WARL and WARN were administered at the end of the first school term (hereafter 'WARL T1' and 'WARN T1'), at the start of the third school term (hereafter 'WARL T3' and 'WARN T3') and at the start of the fourth school term (hereafter 'WARL T4' and 'WARN T4'). The PSC was administered (together with the WARL T4 and WARN T4) at the start of the fourth term (approximately 28 weeks after the administration of WARL T1 and WARN T1). All testing sessions lasted between 10 and 15 minutes per child. Examiners were research assistants, who were trained beforehand on the tasks to be administered. The tests were scored at the time of administration. On the basis of what was written down, the tests were then subsequently double-scored by a different (similarly trained) person to ensure accuracy of results. Scorers and double-scorers were unaware of PSC threshold values at the time of administration and scoring.

Whole-class literacy program

All participants received 'InitiaLit-F' (MultiLit, 2017) instruction in their Foundation school year, and 'InitiaLit-1' (MultiLit, 2018) instruction in Year 1. By the time they were administered the PSC, they had completed one year and three school terms of the whole-class InitiaLit program. Both InitiaLit-F and InitiaLit-1 programs comprise instruction in synthetic phonics, morphology, grammar, oral language comprehension and vocabulary. (See Wheldall et al., 2019, for more information.) Students who have participated in the InitiaLit program tend to show advanced pseudoword decoding skills relative to standardised norms (MultiLit, 2017; 2018).

Data analysis

The question of whether the WARL and/or WARN could accurately identify students who would not meet the PSC expected standards was first addressed using binary logistic regression analysis, being the most appropriate given the binary pass/fail decision rule of the screening test.

Receiver operating characteristic (ROC) analyses were also conducted to evaluate sensitivity, specificity and area under the curve (AUC). For clarity of reporting, students were classified as 'failing' the PSC if their score did not reach the PSC threshold, or 'passing' the PSC if their score reached or exceeded the PSC threshold. 'Sensitivity' was defined as the proportion of children correctly predicted to fail the PSC, while 'specificity' was defined as the proportion of children correctly predicted to pass the PSC. Sensitivity and specificity are inversely related (Greiner, Pfeiffer, & Smith, 2000), since capturing a higher proportion of 'true positives' (i.e. children correctly predicted to fail the PSC) necessitates capturing a higher proportion of 'false positives' (i.e. children incorrectly predicted to fail the PSC). The ROC plots the rate of true positives (sensitivity) against that of false positives (1-specificity) across all possible cut-off scores. The resultant AUC provides the overall classification accuracy, or discrimination ability, of the predictor variable in each model. Specifically, the AUC represents the probability that a randomly selected student from the group that fails the PSC will have a lower score on the predictor test (WARL/WARN) than a randomly selected student from the group that passes. According to Hosmer and Lemeshow (2000), AUC may denote a model's discriminating ability status as non-informative (AUC ≈ .5), acceptable (.7 ≤ AUC < .8), excellent (.8 ≤ AUC < .9) or outstanding (AUC ≥ .9). Given that the present study's predominant research question asked whether the WARL and/or WARN could be used to predict which students would *not* meet the PSC score threshold, *sensitivity* was considered the most important criterion. In other words, the cost of missing a 'true positive'

was considered greater than the cost of obtaining a 'false positive'. As per the protocol in Schäfer (1989), sensitivity was pre-selected, and corresponding specificity and predictor score cut-off values were determined therefrom. Ninety percent sensitivity was considered suitable. Although somewhat arbitrary, this sensitivity level was based on that of screening measures used in prior studies (Johnson et al., 2009; Thomson et al., 2015). The aim was thus to generate and evaluate regression models in which 90% of children who were predicted to fail the PSC were correctly classified as such.

In the present study's main analyses, the threshold for passing the PSC was set at 28 points. This decision was based on the precedent set in 2017 and 2018 when the PSC was implemented in SA schools. (See Appendices A through C for results from the same analyses conducted with the English PSC threshold score of 32 points.)

Results

As summarised in Table 1, the sample obtained a mean PSC score of 33.15 (SD = 6.45). Due to the appearance of non-normal data distributions, non-parametric analyses (i.e., Kruskal-Wallis tests) were conducted (using a Bonferroni-adjustment of $\alpha = .017$ to account for multiple comparisons) and confirmed that there were no significant differences between school sites on PSC, WARL or WARN scores. Students from all schools were therefore combined for subsequent analyses.

[Tables 1 and 2 about here.]

Logistic regression analyses

The WARL and WARN were strongly correlated, both within and across time points (see Table 2). Thus, to avoid multicollinearity, separate regression analyses were computed using the WARL and WARN as predictor variables at each time point (Term 1, Term 3 and Term 4). The base rate of correct classifications (i.e., accuracy), before any other variable was included in analyses, was 82.5%. In other words, because a far greater proportion of

children in this sample passed (n = 113) than failed (n = 24) the PSC, one could classify all children into the 'pass' group and still obtain 82.5% accuracy (i.e., 0% accuracy in correctly classifying those who failed; 100% accuracy in classifying those who passed). To be of statistical (and practical) significance, the respective regression models generated with WARL and WARN scores would need to be sensitive enough to detect a high proportion of students who will fail the PSC, while also demonstrating high (ideally more than 82.5%) overall classification accuracy.

The WARL was examined in the first set of analyses (see Table 3). First, a binary logistic regression model was generated using WARL T1 scores (from Term 1) as the predictor variable. The dependent categorical variable was whether children performed at (/above) versus below the SA PSC threshold of 28 points. A test of the full model, which includes the predictor variable, versus a model with intercept only, was statistically significant, $\chi^2(1) = 38.59$, $p < .001$, and correctly classified 86.1% of total students (again, compared with a base accuracy rate of 82.5%). A second logistic regression model was generated using WARL T3 scores (from Term 3) and the SA PSC threshold of 28 points. The model was significant, $\chi^2(1) = 46.33$, $p < .001$, and correctly classified 87.6% of total students. Finally, the WARL T4 regression model was significant, $\chi^2(1) = 46.83$, $p < .001$, and correctly classified 85.4% of students.

The WARN was examined in the second set of analyses (see Table 4). The model with WARN T1 included as the predictor variable was statistically significant, $\chi^2(1) = 53.61$, $p < .001$, correctly classifying 87.6% of students. Similarly, the WARN T3 model was significant, $\chi^2(1) = 65.79$, $p < .001$, and correctly classified 89.8% of total students. The third regression model in Table 4 was generated using WARN T4 scores. Again, the model was significant, $\chi^2(1) = 52.35$, $p < .001$, and correctly classified 90.5% of total students.

[Tables 3 and 4 about here.]

Overall, the results from binary logistic regression analyses indicated that WARL and WARN scores measured at any time point fitted well (Hosmer & Lemeshow *p*s > .5) and contributed statistically significant value to predicting which students would fail the PSC, based on the criterion. In practical terms, the base rate of 82.5% accuracy was increased by at least 3 percentage points for the WARL (85.4-87.6%) and 5 percentage points for the WARN (87.6-90.5%). In even more practical terms, the number of students correctly categorised based on chance alone increased from 113 (out of 137) to between 118 and 120 with inclusion of the WARL, and between 120 and 124 with inclusion of the WARN.

Probability functions. Figures 1 and 2 show the probability of failing the PSC (<28) given any score on the WARL or WARN, respectively. These values have been calculated from the parameters presented in Tables 3 and 4[1]. Observing any individual's WARL or WARN score, not only can we make the categorical prediction that he or she will fail the PSC, but we can also state the probability that he or she will fail (see y-axis on Figures 1 and 2). This information is independent of our 90% sensitivity decision rule, which is arbitrary, and may be of use to those wishing to apply a different criterion. Intervention programs can be costly, and it may be preferable to minimise the number of students falsely identified as requiring intervention (at the cost of missing some who do require intervention). A lower sensitivity (higher specificity) can be applied in such cases. For example, it might be decided that a student with ≥50% probability of failing should be included in an intervention program. As per Figure 2, this would be commensurate with a raw WARN score of 8.7 or below.

The rightward shift of the probability functions from Term 1 through Term 4 timepoints shows that to maintain an equal probability of failure, the WARL and WARN scores must increase at each testing stage. The slopes of the functions reflect each model's

[1] Probabilities are calculated as $P = 1/(1+\exp(-(\beta_0-\beta_1*W)))$, where W is raw score on WARL or WARN.

predictive power, with steeper slopes describing greater discriminating ability. Discriminating ability is reported formally as the AUC values in the following section.

Receiver operating characteristic (ROC) analyses

ROC analyses were conducted to examine sensitivity and specificity – values that are unaffected by base accuracy rates. As per prior literacy screening measure studies (e.g., Johnson et al., 2009; Thomson et al., 2015), a sensitivity level of 90% was held constant for all analyses. Specificity and predictor score cut-off values associated with 90% sensitivity and 28-point PSC thresholds are summarised in Table 5. As per Appendix C, similar results were obtained when the 32-point PSC threshold was used.

A WARL T1 model with 90% sensitivity had a specificity level of 56%. This was associated with a WARL score of 25.0. In other words, 90% of children with WARL T1 scores at or below 25.0 would be expected to score below 28 points on the PSC, while 56% of children with WARL T1 scores at or above 25.0 would be expected to score at or above 28.0 points on the PSC (see Figure 1). The specificity of our WARL T3 model was 58%, which was associated with a WARL score of 34.8. The specificity of our WARL T4 model was 74%, which was associated with a WARL score of 49.8. The same ROC analyses were conducted using WARN scores at each of the three timepoints (see Figure 2). Again, specificity and WARN scores aligning with a 90% model sensitivity were primarily sought. For WARN T1, the specificity was 69%. This was associated with a WARN T1 score of 9.9. For WARN T3, the specificity was 85%, which was associated with a WARN T3 score of 12.2. Finally, the specificity of our WARN T4 model was 74%, which was associated with a WARN T4 score of 16.2.

The results from ROC analyses confirm what was found through logistic regressions. Essentially, AUC values for WARL and WARN scores (.863-.942) denoted ‘excellent’ or ‘outstanding’ overall predictiveness (Hosmer & Lemeshow, 2000). The specificity of each

model associated with 90% sensitivity ranged from 56% (WARL T1) to 85% (WARN T3). In practical terms, this means WARN T3 scores could be used to correctly classify 90% of students who obtained a PSC score of less than 28 (i.e., true positives) and 85% of students who obtained a PSC score of 28 or more (i.e., true negatives). The range of predictor specificity values at 90% sensitivity (see Table 5) is likely due to sampling error, since specificity decreases exponentially towards higher sensitivity values. The AUC values, which are based on the entire range of sensitivities, therefore offer a more robust indication of the level of overall certainty with which PSC outcomes can be predicted.

[Table 5 about here.]

[Figures 1 and 2 about here.]

Discussion

The present study asked whether students who were likely not to meet the PSC score threshold could be identified using real-word (i.e., WARL) and/or pseudoword (i.e., WARN) fluency measures administered in advance of the PSC. The predictor measures were reasonably consistent across time in the degree to which they explained students' likelihood of failing to meet PSC standards, although there was some variation in the precision offered by the respective predictors. The results reported in this article provide evidence that the WARL and WARN help to predict whether or not students reach the PSC score threshold.

Educational implications

According to our findings, real-word and pseudoword reading fluency measures contributed to predicting PSC results, indicating that it may be useful for classroom teachers to employ similar measures to guide their expectations of students' decoding progress and later PSC performance. Based on results from the present study's dataset, a Year 1 classroom teacher delivering a systematic synthetic phonics program can approximate the probability with which an individual student with a specific WARL or WARN score will not meet the

PSC expected standard. At a group level too, the teacher can identify a cut-off WARL or WARN score below which it is probable students will not reach the threshold score. Generally speaking, the measure with marginally but consistently higher specificity was the WARN. This result might have been expected given that both WARN and PSC stimuli are decodable and therefore rely on the reader's knowledge of letter-sound correspondences, while the WARL includes irregular words that require familiarity.

Equipped with knowledge of which students are struggling to learn to decode words early in Year 1, teachers can then recommend and provide more targeted support. This strategy aligns with a Response to Intervention model of instructional pedagogy, in which students are given increasingly specialised support as they show evidence of requiring it (Coyne et al., 2018). Previous research has indicated that, for students struggling to read in the context of whole-class instruction, intervention delivered in a small-group or individual instructional context is beneficial for literacy learning (Coyne et al., 2018; Grapin et al., 2018). Importantly, the WARL and WARN as employed in the present study are considered a starting point for responding to students' demonstrated difficulties. It is not suggested that they stand to replace the PSC, which has been widely established as a valid and reliable screening tool, but rather to supplement it.

Theoretical implications

Interestingly, approximately half of PSC group-classification variance could not be explained by prior or concurrent reading fluency, and it is worth considering why this might be the case. The obvious difference between PSC and fluency tasks was whether student performance was timed. Hence, a slow and accurate reader may have achieved a low fluency score but a high PSC score. From a cognitive-linguistic perspective, this finding indicates that word-reading automaticity depends on a set of skills that do not entirely overlap with those contributing to word-reading accuracy. Such a conclusion aligns with research into how

reading develops in areas with a shallow orthography (e.g. Italian, Finnish or Greek). There, and in stark contrast with the English orthography, grapheme-phoneme correspondences are consistent, which means literacy difficulties are identified more so on the basis of slow and effortful reading than on inaccuracies in word recognition (Diamanti et al., 2018). Similar fluency-impaired profiles have also been observed in children learning to read English (Wolf & Bowers, 1999). To read decontextualised words fluently, information about an item's orthographic representation must be integrated in a timely fashion with its phonological representation (Wolf & Katzir-Cohen, 2004). Thus, reading fluency – more so than reading accuracy – relies on the timing and coordination of underlying word recognition processes. The contribution of such processes to reading fluency is likely reflected in the finding that WARL and WARN scores did not overlap entirely with PSC scores.

Limitations and future directions

There was a high proportion (82.5%) of Year 1 students who met the 28-point PSC score threshold. Hence, with respect to external validity, the main limitation of the study is that findings will not necessarily generalise to students who do not perform at a similar level of accuracy. The children whose results are reported here received whole-class systematic synthetic phonics (SSP) reading instruction in Foundation and Year 1. Possibly, this factor contributed to their high overall PSC accuracy (see Wheldall et al., 2019, for discussion). In the future, it would be interesting to examine whether the longitudinal relationships between single-item reading fluency and PSC outcomes depended on instructional context. Indeed, the results reported here would have a broader application if replicated with a sample who did not receive InitiaLit. Hypothetically, too, the predictive value of measures like the WARL and WARN should be reduced if literacy instruction changes substantially in the months prior to PSC administration. To test this hypothesis and, more generally, to examine the influence of reading instruction on the longitudinal trajectories of young children's word reading,

further research is needed. A second limitation of the present study was that the models developed to explain the relationship between predictor variables and the PSC were not validated internally. Theoretically, it is desirable for any statistical model to be validated on the basis of a separate confirmatory sample (Chatfield, 1995; Steyerberg et al., 2001). Further research that replicates our findings would therefore be valuable.

The PSC results reported here were collected at the start of the fourth Year 1 school term. Given that this time point is between five and nine weeks later than in South Australia and between three and four weeks earlier than in England, it may be considered a study limitation that PSC administration was not synchronised with other systems. That said, the finding that similarly significant results were found for both SA and English PSC-score thresholds lends validity to the models generated. In addition, the pass rates and associated predictive models for students in the present study were calculated for two related measures (i.e., WARL and WARN) across three time points. Reliability of the statistical models is therefore demonstrated by the replicated probability functions, all six of which have similar slopes and shift over time in a direction and distance as would be expected.

In the present study, assessment data at all three timepoints were collected by research assistants. This step meant they could be individually trained on the test administration and scoring protocols, thereby resulting in fewer errors. However, it may have reduced the generalisability of results, since, in a real-life classroom context, teachers would usually be expected to be administrators. This point of difference between our study and real-life testing conditions may warrant further investigation to determine whether there is indeed a discrepancy in scoring between external administrators and classroom teachers.

Conclusions

In this study, real-word and pseudoword reading fluency measures were administered in advance of – and concurrently with – the PSC. The research question under investigation

was how well performance on the PSC, as measured using the binary outcome measure of meeting the 28-point SA PSC threshold, could be predicted ahead of time by scores on one of the reading fluency measures. At each testing stage, reading fluency statistically predicted the likelihood of not reaching the PSC threshold, although the practical significance of the generated models' predictive values varied somewhat between timepoints and measures. Ultimately, it is hoped the results reported here will assist teachers in identifying and providing targeted intervention to those children early in Year 1 who are struggling with basic word decoding skills.

References

Australian Curriculum, Assessment and Reporting Authority (ACARA). (2011). *Guide to understanding ICSEA.* ACARA.

Australian Curriculum, Assessment and Reporting Authority (ACARA). (2019). *My School.* https://www.myschool.edu.au

Bates, C., & Nettelbeck, T. (2001). Primary school teachers' judgements of reading achievement. *Educational Psychology, 21*(2), 177-187. https://doi.org/10.1080/01443410020043878

Buckingham, J., Wheldall, K., & Beaman-Wheldall, R. (2013). Why poor children are more likely to become poor readers: The school years. *Australian Journal of Education, 57*(3), 190-213. https://doi.org/10.1177/0004944113495500

Buckingham, J., & Wheldall, K. (2020). Why all states and territories should follow South Australia's lead and introduce the Year 1 Phonics Check: An update. *LDA Bulletin, 51*(2-3), 14-16.

Chatfield, C. (1995). Uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society, 158*(3), 419-466. https://www.jstor.org/stable/2983440

Clark, M. M., & Glazzard, J. (2018). *The phonics screening check 2012-2017: An independent enquiry into the views of head teachers, teachers and parents.* https://www.newman.ac.uk/wp-content/uploads/sites/10/2018/09/The-Phonics-Screening-Check-2012-2017-Final-Report.pdf

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review, 108*(1), 204-256. https://doi.org/10.1037/0033-295X.111.1.159

Coyne, M.D., Oldham, A., Dougherty, S.M., Leonard, K., Koriakin, T., Gage, N.A., Burns, D., & Gillis, M. (2018). Evaluating the effects of supplemental reading intervention

within an MTSS or RTI reading reform initiative using a regression discontinuity design. *Exceptional Children, 84*(4), 350-367. https://doi.org/10.1177/0014402918772791

Department for Education. (2018). *Phonics screening check and key stage 1 assessments in England, 2018.* London: UK Department for Education.

Diamanti, V., Goulandris, N., Campbell, R., & Protopapas, A. (2018). Dyslexia profiles across orthographies differing in transparency: An evaluation of theoretical predictions contrasting English and Greek. *Scientific Studies of Reading, 22*(1), 55-69. https://doi.org/10.1080/10888438.2017.1338291

Double, K. S., McGrane, J. A., Stiff, J. C., & Hopfenbeck, T. N. (2019). The importance of early phonics improvements for predicting later reading comprehension. *British Educational Research Journal, 45*(6), 1220-1234. https://doi.org/10.1002/berj.3559

Duff, F.J., Nation, K., Plunkett, K., & Bishop, D.V.M. (2015). Early prediction of language and literacy problems: Is 18 months too early? *PeerJ, 3*(e1098). https://peerj.com/articles/1098/

Government of South Australia. (2019). *2018 Phonics Screening Check.* Retrieved from: https://www.education.sa.gov.au/sites/default/files/2018-phonics-screening-check-fact-sheet.pdf

Graham, L. J., White, S. L. J., Tancredi, H. A., Snow, P. C., & Cologon, K. (2020). A longitudinal analysis of the alignment between children's early word-level reading trajectories, teachers' reported concerns and supports provided. *Reading and Writing*. Advance online publication. https://link.springer.com/article/10.1007/s11145-020-10023-7

Grapin, S. L., Waldron, N., & Joyce-Beaulieu, D. (2018). Longitudinal effects of RtI implementation on reading achievement outcomes. *Psychology in the Schools, 56*(2), 242-254. https://doi.org/10.1002/pits.22222

Greiner, M., Pfeiffer, D., & Smith, R.D. (2000). Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. *Preventative Veterinary Medicine, 45*, 23-41. https://doi.org/10.1016/s0167-5877(00)00115-x

Hosmer, D.W., & Lemeshow, S. (2000). *Applied logistic regression (2nd ed.)*. New York: Wiley.

Hordacre, A., Moretti, C., & Spoehr, J. (2017). *Evaluation of the trial of the UK phonics screening check in South Australian schools.* Australian Industrial Transformation Institute, Flinders University of South Australia.

Johnson, E.S., Jenkins, J.R., Petscher, Y., & Catts, H.W. (2009). How can we improve the accuracy of screening instruments? *Learning Disabilities Research and Practice, 24*(4), 174-185. https://doi.org/10.1037/0022-0663.92.2.248

Language & Reading Consortium. (2018). The Simple View of Reading across development: Prediction of Grade 3 reading comprehension from prekindergarten skills. *Remedial and Special Education, 39*(5), 289-303. https://doi.org/10.1177/0741932518762055

MultiLit. (2017). *InitiaLit Foundation: Whole class initial instruction in literacy.* MultiLit Pty Ltd.

MultiLit. (2018). *InitiaLit Year 1: Whole class initial instruction in literacy.* MultiLit Pty Ltd.

Nation, K. (2019). Children's reading difficulties, language, and reflections on the Simple View of Reading. *Australian Journal of Learning Difficulties.* https://doi.org/10.1080/19404158.2019.1609272

Schäfer, H. (1989). Constructing a cut-off point for a quantitative diagnostic test. *Statistics in Medicine, 8*(11), 1381-1391. https://doi.org/10.1002/sim.4780081110

Share, D.L. (1995). Phonological recoding and self-teaching: Sine qua non of reading acquisition. *Cognition, 55*(2), 151-218. https://doi.org/10.1016/0010-0277(94)00645-2

Snowling, M. J., Stothard, S. E., Clarke, P., Bowyer-Crane, C., Harrington, A., Truelove, E., & Hulme, C. (2012). *York assessment of reading for comprehension – passage reading (Australian ed.).* GL Assessment.

Speech Pathology Australia. (2017). *Speech pathology in schools.* https://speechpathologyaustralia.cld.bz/Speech-Pathology-in-Schools-2017

Stainthorp, R. (2020). *A national intervention in teaching phonics: A case study from England.* [Manuscript submitted for publication]. Institute of Education, University of Reading.

Standards & Testing Agency. (2011). *Year 1 phonics screening check: Pilot 2011: Technical report.* Department for Education.

Standards & Testing Agency. (2012). *Year 1 phonics screening check: 2012 scoring guidance.* Department for Education.

Standards & Testing Agency. (2018a). *Phonics screening check: 2018 administration guidance*. Department for Education.

Standards & Testing Agency. (2018b). *Phonics screening check: 2018 scoring guidance.* Department for Education.

Stephanou, A., Anderson, P., & Urbach, D. (2008). *Progressive achievement tests in reading: Comprehension, vocabulary and spelling (PAT-R).* Australian Council for Educational Research.

Steyerberg, E.W., Harrell, F.E., Borsboom, G.J.J.M., Eijkemans, M.J.C., Vergouwe, Y., & Habbema, J.D.F. (2001). Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. *Journal of Clinical Epidemiology, 54*(8), 774-781. https://doi.org/10.1016/S0895-4356(01)00341-9

Thompson, P.A., Hulme, C., Nash, H.M., Gooch, D., Hayiou-Thomas, E., & Snowling, M.J. (2015). Developmental dyslexia: Predicting individual risk. *Journal of Child Psychology and Psychiatry, 56*(9), 976-987. https://doi.org/10.1111/jcpp.12412

Torgesen, J. K., & Hudson, R. F. (2006). Reading fluency: Critical issues for struggling readers. In S. J. Samuels & A. Farstrup (Eds.), *What research has to say about fluency instruction.* (pp. 130-158). International Reading Association.

UK Literacy Association. (2012, October 26). Phonics screening check fails a generation of able readers. *UK Literacy Association.* https://ukla.org/news/story/phonics_screening_check_fails_a_generation_of_able_readers

Wheldall, K., Bell, N., Wheldall, R., Madelaine, A., & Reynolds, M. (2019). Performance of Australian children on the English Phonics Screening Check following systematic synthetic phonics instruction in the first two years of schooling. *Australian Journal of Learning Difficulties, 24*(2), 131-145. https://doi.org/10.1080/19404158.2019.1635500

Wheldall, K., Reynolds, M., & Madelaine, A. (2015). *The Wheldall Assessment of Reading Lists (WARL).* MultiLit Pty Ltd.

Wolf, M., & Bowers, P.G. (1999). The double-deficit hypothesis of the developmental dyslexias. *Journal of Educational Psychology, 91*(3), 415-438. https://doi.org/10.1037/0022-0663.91.3.415

Wolf, M., & Katzir-Cohen, T. (2001). Reading fluency and its intervention. *Scientific Studies of Reading, 5*(3), 211-239. https://doi.org/10.1207/S1532799XSSR0503_2

Table 1.

Summary of WARL and WARN results and Phonics Screening Check results.

Time point	Measure	Mean	SD	Median
End Term 1	WARL T1	44.80	21.39	41.0
	WARN T1	16.05	7.24	15.0
Start Term 3	WARL T3	55.80	22.21	54.0
	WARN T3	19.28	8.35	18.0
Start Term 4	WARL T4	66.89	21.25	69.0
	WARN T4	23.70	8.38	24.0
	PSC	33.15	6.45	36.0

Note. PSC = Phonics Screening Check; WARL = Wheldall Assessment of Reading Lists; WARN = Wheldall Assessment of Reading Nonwords.

Table 2.

Correlations between WARL and WARN results at each testing time point.

	WARL T3	**WARL T4**	**WARN T1**	**WARN T3**	**WARN T4**
WARL T1	.923**	.797**	.873**	.848**	.772**
WARL T3		.871**	.826**	.882**	.807**
WARL T4			.716**	.768**	.837**
WARN T1				.903**	.792**
WARN T3					.858**

Note. WARL = Wheldall Assessment of Reading Lists; WARN = Wheldall Assessment of Reading Nonwords. ** = $p < .001$.

Table 3.

Logistic regression models for 28-point Phonics Screening Check threshold values when the WARL was administered at three timepoints.

Measure	R^2	Variable	β (SE)	95% CI for Odds Ratio			Wald
				Lower	Odds Ratio	Upper	
WARL T1	**.406	WARL	-0.11 (0.02)	0.86	0.90	0.94	**19.60
		Constant	2.11 (0.73)	-	8.21	-	*8.29
WARL T3	**.475	WARL	-0.10 (0.02)	0.87	0.90	0.94	**23.31
		Constant	3.00 (0.85)	-	20.07	-	**12.52
WARL T4	**.479	WARL	-0.10 (0.02)	0.88	0.91	0.94	**25.01
		Constant	3.88 (1.02)	-	48.49	-	**14.51

Note. Model R^2 = Nagelkerke statistic. WARL = Wheldall Assessment of Reading Lists. * = $p < .01$; ** = $p < .001$.

Table 4.

Logistic regression models for 28-point Phonics Screening Check threshold values when the WARN was administered at three timepoints.

Measure	R^2	Variable	β (SE)	95% CI for Odds Ratio			Wald
				Lower	Odds Ratio	Upper	
WARN T1	**.536	WARN	-0.51 (0.11)	0.48	0.60	0.75	**20.06
		Constant	4.42 (1.19)	-	83.43	-	**13.90
WARN T3	**.631	WARN	-0.50 (0.11)	0.49	0.61	0.75	**21.30
		Constant	5.27 (1.30)	-	194.98	-	**16.59
WARN T4	**.525	WARN	-0.28 (0.05)	0.68	0.76	0.84	**27.55
		Constant	3.83 (0.93)	-	46.01	-	**16.90

Note. Model R^2 = Nagelkerke statistic. WARN = Wheldall Assessment of Reading Nonwords. ** = $p < .001$.

Table 5.

Results from receiver operating characteristic (ROC) analyses with 28-point Phonics Screening Check threshold.

Measure	AUC	Specificity	Score
WARL T1	.863	56%	25.0
WARL T3	.892	58%	34.8
WARL T4	.906	74%	49.8
WARN T1	.909	69%	9.9
WARN T3	.942	85%	12.2
WARN T4	.903	74%	16.2

Note. AUC = Area under the curve; WARL = Wheldall Assessment of Reading Lists; WARN = Wheldall Assessment of Reading Nonwords.

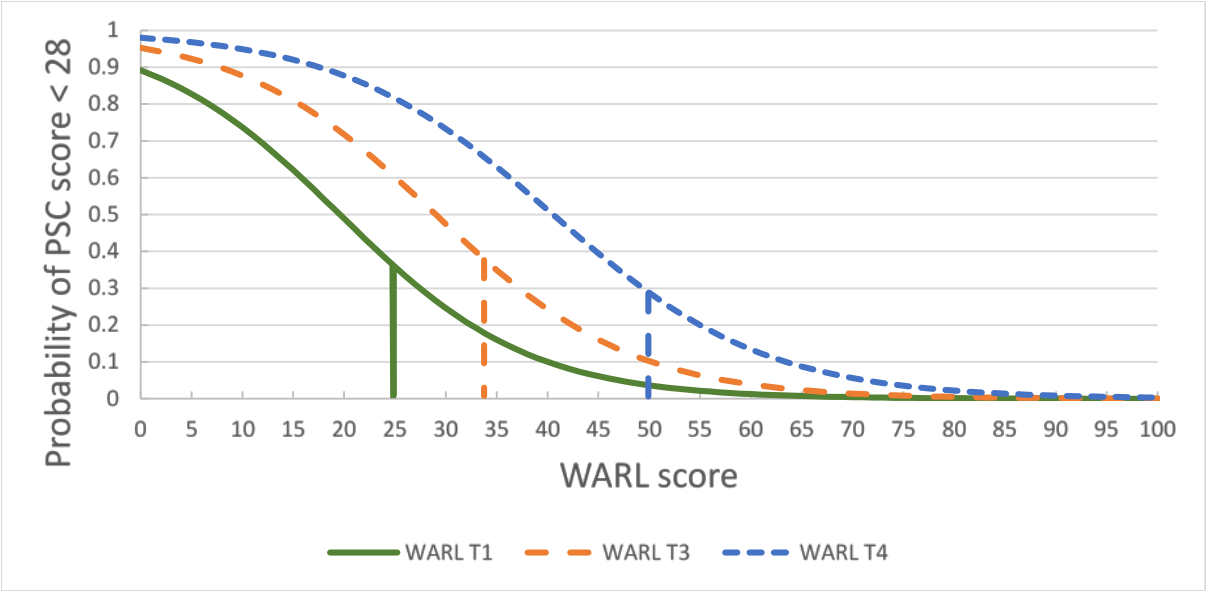

Figure 1. Probability of not achieving the expected standard, as indicated by a pass/fail Phonics Screening Check (PSC) threshold of 28 points. Vertical lines approximate the Wheldall Assessment of Reading Lists (WARL) score below which the student is likely not to meet the PSC expected standard (sensitivity = 90%).

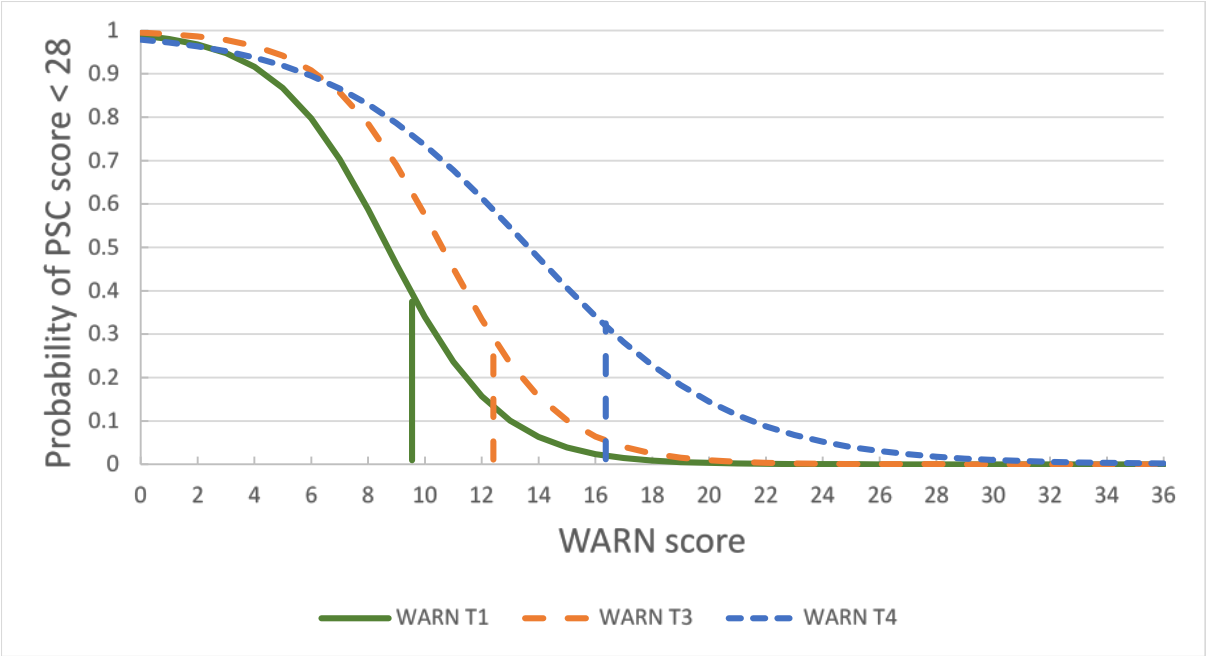

Figure 2. Probability of not achieving the expected standard, as indicated by a pass/fail Phonics Screening Check (PSC) threshold of 28 points. Vertical lines approximate the Wheldall Assessment of Reading Nonwords (WARN) score below which the student is likely not to meet the PSC expected standard (sensitivity = 90%).

Appendix A.

Logistic regression models for 32-point Phonics Screening Check threshold values when the WARL was administered at three timepoints.

Measure	R^2	Variable	β (SE)	95% CI for Odds Ratio			Wald
				Lower	Odds Ratio	Upper	
WARL T1	**.379	WARL	-0.09 (0.02)	0.89	0.92	0.95	**23.80
		Constant	2.19 (0.61)	-	8.92	-	**12.70
WARL T3	**.429	WARL	-0.08 (0.02)	0.89	0.92	0.95	**27.70
		Constant	2.98 (0.72)	-	19.68	-	**17.06
WARL T4	**.399	WARL	-0.08 (0.01)	0.90	0.93	0.90	**27.44
		Constant	3.63 (0.87)	-	37.77	-	**17.55

Note. Model R^2 = Nagelkerke statistic. WARL = Wheldall Assessment of Reading Lists. * = $p < .01$; ** = $p < .001$.

Appendix B.

Logistic regression models for 32-point Phonics Screening Check threshold values when the WARN was administered at three timepoints.

Measure	R^2	Variable	β (SE)	95% CI for Odds Ratio			Wald
				Lower	Odds Ratio	Upper	
WARN T1	**.463	WARN	-0.34 (0.07)	0.62	0.71	0.81	**24.62
		Constant	3.49 (0.84)	-	32.83	-	**17.44
WARN T3	**.519	WARN	-0.30 (0.06)	0.66	0.74	0.83	**28.64
		Constant	3.85 (0.84)	-	46.78	-	**20.86
WARN T4	**.491	WARN	-0.24 (0.04)	0.72	0.79	0.85	**31.94
		Constant	4.06 (0.87)	-	58.16	-	**21.73

Note. Model R^2 = Nagelkerke statistic. WARN = Wheldall Assessment of Reading Nonwords. * = $p < .01$; ** = $p < .001$.

Appendix C.

Results from receiver operating characteristic (ROC) analyses with 32-point Phonics Screening Check threshold.

Measure	AUC	Specificity	Score
WARL T1	.831	61%	27.0
WARL T3	.852	65%	39.5
WARL T4	.842	64%	53.2
WARN T1	.863	55%	10.1
WARN T3	.888	73%	13.5
WARN T4	.872	64%	18.0

Note. AUC = Area under the curve; WARL = Wheldall Assessment of Reading Lists; WARN = Wheldall Assessment of Reading Nonwords.