"Development of Nonword and Irregular Word Lists for Australian Grade 3 Students using Rasch Analysis"

Sarah Callinan PhD

Everarda Cunningham PhD

Stephen Theiler PhD

Swinburne University of Technology

Abstract

Many tests used in educational settings to identify learning difficulties endeavour to pick up only the lowest performers. Yet these tests are generally developed within a classical test theory (CTT) paradigm that assumes that data does not have significant skew. Rasch analysis is more tolerant of skew and was used to validate two newly-developed tests of phonological and orthographic processing. These tests were administered to 172 Victorian Grade 3 students so that preliminary chronological and reading age norms could be generated. Despite a level of skew that would cause difficulties in a CTT paradigm, the tests were found to be acceptable as per Rasch analysis and, as such, provide additional and important information on poor-performing students. Preliminary norms for the test are provided.

Despite the recognised adverse outcomes of literacy difficulties (Australian Council for Educational Research (ACER), 2000; McGee, Prior, Williams, Smart & Sanson, 2002), particularly Learning Difficulties (LD) (Kortering, Braziel & McClannon, 2010; Undheim & Sund, 2008), there is no widely-accepted method for early identification of LD. This may be because the integration of relevant school psychology and neuropsychology research is slow (Decker, 2008), despite its importance (Cleary & Scott, 2011). It is important to ensure that new methods of identification of LD are both theoretically sound and well suited to administration in the classroom.

While literacy levels in Australian schools dropped slightly from 1975 to 1998, the expenditure per student rose significantly in the same time (Leigh & Ryan, 2008). This could, in part, be attributed to an increase in spending in special education. In the U.S., money put towards special education students is often spent on screening or diagnosing students, rather than actually assisting them (Fuchs & Fuchs, 1995). By developing freely available tests that evaluate relevant cognitive processes implicated in low school achievement, these costs could be reduced. Two such processes that are important in literacy and LD research and require wide scale testing are phonological and orthographic processing.

Phonological processing is “the ability to perceive and deliberately manipulate the sound components of words” (Kirby, Roth, Desrochers & Lai, 2008, p.105) and is the most commonly discussed cognitive deficit in LD literature. Longitudinal studies have consistently demonstrated that early phonological processing can predict later reading achievement or the presence of LD (Carovalas, et al., 2012; Hulme, Bowyer-Crane, Carroll, Duff & Snowling, 2012; Linklater, O'Connor & Palardy, 2009; Muter & Snowling, 2009; Storch & Whitehurst, 2002). Event related potential studies have been used to demonstrate qualitative neurological differences between LD and non-LD students when processing phonological information (Fosker & Thierry, 2005).

Nonword reading is thought to be the "primary index of phonological coding ability" (Stanovich & Siegel, 1994, p. 28), and is considered to be the best single test of the presence of LD (Siegel, 1989; Stanovich, 1999). Nonword reading at Grade 3 predicts a student's performance on a Year 12 Wordchain test better than the performance on the same Wordchain test in Grade 3 (Svensson & Jacobson, 2006). Furthermore, nonword reading at age 11 is also predictive of reading ability at age 14 (Conti-Ramsden & Durkin, 2007). The reciprocal relationship between phonological processing ability and reading ability has lead to the suggestion that phonological processing scores should be compared to a reading age, rather than chronological age, based norm (Siegel & Ryan, 1988). While development of such norms is out of the scope of the current study, any new test of phonological processing should allow for the possibility of such norms being developed in the future.

While phonological processing is the most commonly studied cognitive deficit inherent to LD, the role of orthographic processing deficits have also received attention (Ho, Chan, Tsang & Lee, 2002; Stage, Abbott, Jenkins & Berninger, 2003). Orthographic processing is "the ability to form, store, and access orthographic representations" (Stanovich & West, 1989, p. 404). One way to view the role of orthographic processing in LD is through a model incorporating both surface and phonological dyslexia as subgroups of LD. People with surface dyslexia do not have phonological processing deficits, but have marked orthographic processing deficits (Friedmann & Lukov, 2008).

It is possible that phonological processing deficits are indicative of a neurological disorder while orthographic processing deficits are attributable to lack of exposure to print (Griffiths & Snowling, 2002). Multiple studies have attributed orthographic deficits to environmental factors (e.g., Braten, Lie, Andreassen & Olaussen, 1999; Castles Datta, Gayan & Olson 1999), and findings from longitudinal studies suggest that in the long term issues stemming from orthographic deficits may be less serious and long-lasting than those stemming from phonological deficits (Byrne, Freebody & Gates, 1992).

If recent research into the role of orthographic deficits in LD gains support (de Jong & Messbauer, 2011; O'Brien et al., 2011) then this test will still be a valuable tool. Irregular word reading is considered a good measure of orthographic processing (Kirby et al., 2008) as it requires readers to recognise the whole word rather than reading it phonetically.

Over a decade ago, the need for freely available word lists that measure both orthographic and phonological processing was recognised and three word-lists of regular, irregular and nonwords were developed (Edwards & Hogben, 1999). Despite considerable potential application in schools, these tests were not widely used, possibly because the scale was aimed toward seven to twelve year olds; as such it was not specifically targeted to one population. Therefore tests were skewed for some age groups despite psychometric validation through Classical Test Theory (CTT). Similar to the measure of a nonword reading test developed by Martin and Pratt (2001), this means that the scale had to be targeted to a wide range of both age groups and reading proficiencies within those age groups. In the meantime, the need for freely available screening tests of the cognitive deficits that are indicative of LD that can be used in the classroom is still being voiced (Skues & Cunningham, 2011)

Screening tools cannot be used for definitive diagnosis; they are developed to identify those at-risk of having the targeted disorder. This is done in the hope that the disorder can be identified before it develops beyond the point of feasible remediation (Morrison, 1992). The distribution of scores from the wider population from screening tools that aim to detect specific issues are often strongly skewed because the scale targets those with particularly low (or high) functioning in the tested construct. For example, the Psychological Consequences Questionnaire (Rijnsburger, Essink-Bot, van As, Cockburn & de Koning, 2006), the Self Harm Inventory (Latimer, Covic, Cumming & Tennant, 2009), The Depression Anxiety

Stress Scales (Lovibond & Lovibond, 1995; Shea, Tennant & Pallant, 2009) and the Kessler measure of psychological distress (Kessler et al., 2002) all have skewed person distributions for the general population so that maximum information can be extracted from these scales on the targeted population. All these tests are popular and widely used, despite the assumption of normality in CTT that assumes that data will be normally distributed with little if any skew.

The basis of CTT is that all test scores are the sum of a participant's true score for the test plus error (Novick, 1966). One of the primary issues with the application of CTT with a screening tool is that the assumption of normal distribution can be easily violated in a test designed to only separate those respondents at one end of a spectrum. Ceiling effects can occur when a test designed to pick up poor performances is given to a normal sample (Alexander & Martin, 2000). However, it could also be argued that any screening test that is not skewed towards the target population, especially if a small minority is targeted, is not effectively doing its job. Rasch models provide an alternative to CTT that is more accommodating of issues such as these. This Rasch model has been successfully used to provide psychometric support for some of the skewed scales outlined above (Latimer et al., 2009; Tennant & Pallant, 2009)

The aim of a Rasch model is to measure a latent trait or variable and is based on the relationship between the difficulty of each item and each participant's level of proficiency on the measure (Pallant & Tennant, 2007). An important assumption of a Rasch model is that the likelihood of a given item being endorsed by a given person is a function of both the level of the latent trait in the person and the level of the latent trait being assessed in the item. Both items and respondents are assessed on the same scale and it is possible to avoid person distribution assumptions, dependent on the likelihood estimation technique selected. Difficult items have logit scores akin to high-performing respondents and easy items have similar logit scores as poor-performing respondents. Therefore, if a respondent and an item have the same logit score, there is a 50% chance the respondent will get that item correct. As part of this model, items and people are assigned fit statistics to indicate fit to the overall model, as well as assessments of unidimenstionality. Differential item fit analysis also provides an assessment of the how different groups respond to different items with an important assumption being that two people from different groups at the same level of the latent trait would answer the same item in the same way.

Rasch has been successfully applied to screening tools in other areas (Fink et al., 2004; Shea et al., 2009) as well as the LD field (Chan, Ho, Chung, Tsang & Lee, 2012), and the theory behind using CTT in conjunction with IRT has been addressed both theoretically and mathematically (Bechger, Maris, Verstralen & Beguin, 2003). Rasch analysis is thought to be superior to CTT in many areas of scale development; however most advocates suggest that Rasch be used in conjunction with CTT rather than in its place (e.g., Erhart et al., 2009).

The aim of the current study is to develop two word lists through the tenets of both CTT and IRT. The first, consisting of irregular words, will measure orthographic processing, and second, consisting of nonwords, will measure phonological processing. These tests will be targeted towards Grade 3 students. This is so that the population of interest, poorly performing Grade 3 students, can be targeted accurately. Grade 3 was selected as it is considered a pivotal year in reading acquisition (Jarmulowicz, Taran & Hay, 2007) and is often the year that students with LD start to noticeably fall behind their peers as reading proficiency starts to be assumed (Woolley, 2007). The lists will be refined through the use of Rasch analysis as well as exploratory factor analysis (EFA) and Cronbach's alpha as per Erhart and colleagues' (2009) recommendation. Once these lists are finalised preliminary norms, including reading age norms for the nonword list based on the pilot study, will be presented.

Method

Participants

Ethics

Ethical clearance was initially obtained from Swinburne University Human Research Ethics Committee. Clearance was then gained from the Victorian Department of Education and Early Childhood Development and the Catholic Education Office in the Archdiocese of Melbourne, to conduct the study in both State and Catholic schools.

The principals of the five participating schools all consented to having the study conducted at their school and it was at this point that students were sent home with information sheets and consent forms.

Sampling method

At both the school and student level, the sample was essentially one of convenience. Seventeen State and Catholic schools within Victoria were sent a letter inviting them to participate in the study and five agreed. Four of these schools were State Schools in the outer eastern suburbs of Melbourne and one was a Catholic Primary School in the south-eastern suburbs. In each school, all Grade 3 students were invited to participate as well as all Grade 1 and 2 students in one of the State schools.

Students

One hundred and seventy five Grade 3 students participated in the study. Three students were not suited to the sample because of disability or having English as a second language, however in order to ensure they did not feel excluded they were still asked to participate in the testing and their results were removed from the final sample. The number of participants, mean age and response rates from each school, in chronological order of testing, are shown in Table 1. Students from School 4 and 5 were also asked to be tested again a month after the first round of testing, to establish the word lists' test-retest reliability. Of these 35 students, 86% (i.e., 30) of those asked agreed to participate in the second round of testing.

[Insert Table 1 approximately here]

Grade 1 and 2 students were administered the nonword and irregular word list in order to develop the reading age norms. Therefore 58 Grade 1 students (*M* = 7.05 years, *SD* = 0.32 years) and 57 Grade 2 students (*M* = 8.09 years, *SD* = 0.32 years) from School 3 also participated in the study. The response rates for the Grade 1 and 2 classes was 63% percent in both grade levels.

Measures

Word Lists

Preliminary lists were drawn up and administered to a sample of twelve children from the first school to ensure that both word lists were appropriately targeted. From this small pilot it was gathered that the original test was too easy, so it was amended before administration to the entire sample.

Irregular word reading list

Irregular word reading lists were developed to test orthographic processing. Irregular words are words that are not pronounced the way they are spelt, with no strict spelling-to-sound correspondence. Examples of irregular words include *yacht* and *eye*. The original list of 36 words was formulated to ensure that all currently used English phonemes and letters were present in the list and the number of letters, syllables, phonemes and consonant clusters in each word were all monitored to ensure a good combinations of these factors.

Consonant clusters are the largest number of consecutive consonants in a word that are not broken up by a vowel. Similar to the other three factors, the number of consonant clusters is positively correlated with difficulty in reading a word (Martens & de Jong, 2006; Rispens &

Parigger, 2010; Ziegler & Goswami, 2005; Ziegler, Perry & Coltheart 2000). In order to ensure that the words on the list were easily recognised, the list was compared to a list of the 3041 most commonly used words by Australian Grade 2 children (Lo Bianco, Scull & Ives, 2008) and as all words were on this list, they were retained.

Nonword reading list

Nonword reading lists were developed to test phonological processing. Nonwords are made-up words that are pronounceable but have no meaning. For instance, *blurk* is a nonword while *thdfaxc* is not. Some of the words had more than one correct pronunciation based on either letter sounds or precedent in commonly-used English words. For instance *kour* could be pronounced to rhyme with *four, hour* or *tour*.

The number of syllables, letters, phonemes and consonant clusters were taken into account when developing the original 36 word nonword list and the original list of 36 was put together to ensure good coverage of phonemes and letters. Similar to irregular word reading, all of these factors are positively correlated with the difficulty in reading a word (Pammer, Lavis, Hansen & Cornelissen, 2004; Ziegler & Goswami, 2005; Ziegler et al., 2000). All letters and phonemes were used at least once in the test. Although these factors were a major part in the selection of the full nonword list, the refined word list was selected on psychometric grounds.

Progressive Achievement Test in Reading

The Progressive Achievement Test in Reading (PAT-R), 4th Edition (ACER, 2008), is a widely used and accepted Australian test of reading comprehension. Students are given 40 minutes to complete the test as well as approximately 15 minutes administration time.

Each grade-level has its own norms for each test, and each possible score for each grade-level has a corresponding percentile score. These scores are comparable across grades (Australian

Council for Educational Research, 2004). Therefore a Grade 3 student who obtained a score in the 75th percentile on the Grade 3 PAT-R is equivalent to a Grade 3 student who got a score in the 75th percentile on the Grade 2 PAT-R, as long as both are compared to their grade appropriate norms within each test.

Statistical programs

The Rasch Unidimensional Model Management (RUMM) version 2020 (Andrich, Sheridan & Luo, 2004) using pairwise maximum likelihood estimation technique was used for the Rasch analysis. Secondly, the Predictive Analytics SoftWare (PASW; formally SPSS) version 17 (PASW, 2009) and a freeware Monte Carlo Parallel Analysis eigenvalue generator (Watkins, 2008) was utilised for the EFA.

Procedure

Administration

Students were administered the two word lists and the PAT-R as part of a larger study. The PAT-R was administered to each class as a group. Students were also administered the two word lists, and two short cognitive processing tests not relevant to this study as part of individual testing. Students were removed individually from the classroom for approximately 15 minutes to complete the testing and the word lists were the last of the three tasks administered. There were no time limits for the word lists.

Data Analysis

The final word lists were analysed using both Rasch and CTT based analyses. First of all item fit statistics were examined to ensure that all items fit well to their respective scales as well as both scales having good item fit overall. Similar to item fit, person fit is measured on two fronts, namely overall person fit and individual person fit. Both item and person fit statistics are examined to ensure that no items had a significant χ^2 values or an absolute residual value over 2.5. DIF analysis was conducted to ensure that no items were unduly influenced by either gender or school. Targeting and reliability was measured with both Cronbach's alpha and the Rasch measure of reliability, the Person Separation Index (PSI). Furthermore the mean person location of each scale, where zero indicates a normally distributed scale and scores above an absolute value of one indicate skew, will also be examined. Unidimensionality was also assessed through both Rasch and CTT paradigms. Firstly principal axis factor-analysis was used in conjunction with a Monte Carlo parallel analysis to assess the appropriate number of factors for each of the word lists, and a *t*-test of items identified through the principal components residuals within the Rasch model will also ensure that there are no significant differences between those with the highest negative and positive residuals, indicating multidimensionality. Finally test-retest reliability, using a small sub-sample of students who were retested, will be assessed with a Pearson's correlation.

Each list started with 36 items and was refined as part of an iterative process using all of the above analyses. All analyses were first conducted on each original 36 item list. No decisions about item deletion or rescoring were made based on the results on any one test. Instead, the analyses from all the tests were considered as a whole. In determining item removals, unidimensionality, reliability, DIF analysis, individual item analysis and skew were taken into account. Items were removed one by one with analyses repeated after each deletion.

Redundant items were the first removed. Items with poor misfit, significant DIF analyses or those that did not fit well in an unidimensional model were the next examined for removal. As there was an issue with difficulty in this list, more difficult items were retained for more iterations than easier items to ensure that the misfit issues remained once more seriously misfitting or redundant items were removed. Finally any items with misfit, redundancy, significant DIF analysis or dimensionality removed, items that were not providing much extra information due to performing similarly were removed in order to improve parsimony. This resulted in a 15 item irregular list and a 20 item nonword list.

Results

All 172 participants from the Grade 3 cohort completed both word lists, so there were no missing data in this study. All data were screened to check it met all assumptions of the analyses used in the current study. Univariate outliers were identified through extreme *z*-scores and multivariate outliers were captured by examining scatter plots and calculating Mahalanobis distances. All analyses were run with and without outliers to ensure that they were not affecting results, as this was not the case, they were retained.

A Note on Scoring Options

In the original full word lists three scoring options were noted, two points for a correct answer, zero point for an incorrect answer and one point for the middle category for answers that were primarily correct but with an incorrect suffix or plural. This three category scoring method was only required for twelve items in the irregular word list and fourteen items in the nonword list as the other items did not lend themselves to such response options. The threshold map for the irregular word list showed that all items with the two categories had ordered thresholds and all of those with three categories had disordered thresholds. This was the same for the nonword list, albeit five items with three categories had ordered thresholds. After an item by item analysis of this scoring system using category probability curves it was decided that this middle category was not providing enough information to justify its inclusion in the scale. These category probability curves display the likelihood of a response category being endorsed as a function of the total score. An example of a typical category probability curve for one of these three point scoring items is shown in Figure 1.

[Insert Figure 1 approximately here]

As demonstrated in Figure 1, the problem is not that participants receiving the one point option were on average getting higher scores than those who got the answer correct, or conversely, less than those who got the item incorrect. Instead, it was that this option was never the most likely response for any total score. This is considered an indicator that the category is not informative enough to be retained (Andrich, 2005). As this was the case for all 12 items in the irregular word list, and 9 of the 14 items in the nonword list, the middle category was removed. Consequently, all one point answers were recoded to incorrect answers and all correct answers were given one, instead of two points.

Word Lists

The final irregular word list consisted of 15 items with a possible score range of zero to 15. The mean score was 10.53 (*SD* = 3.45). Rasch analysis of the scale resulted in good model fit, $\chi^2(30) = 30.9$, $p = .421$, for the scale. The means and standard deviations on the re-scored and refined irregular word list in order of difficulty, along with item fit statistics, are shown in Table 2.

[Insert Table 2 approximately here]

The nonword list had more variation and was slightly less skewed to accommodate either type of norm reference being used. The final word list consisted of 20 items with a scoring range of zero to 20. The mean score for the scale was 13.09 (*SD* = 4.97). Rasch analysis on the final 20 item list indicated good model fit, $\chi^2(40) = 35.49$, $p =.674$. The re-scored means and standard deviations of the items, in order of difficulty, are shown in Table 3.

[Insert Table 3 approximately here]

Individual item analysis

The item fit analysis was examined to ensure that all items fit well to each scale as well as both scales having good item fit overall. The overall irregular word list item fit residual standard deviation of 0.74 and 0.59 for the nonword list were clearly lower than the upper limit of 1.5 and indicated that overall item fit was good. This was further supported by the individual item fit analysis displayed in Table 2 for the irregular word list and Table 3 for the nonword list. Individual item fit statistics were examined and no items had significant χ^2 values or fit residuals over an absolute value of 2.5, either of which would indicate item misfit. It is worth noting that some items, such as *strimpex*, had a significant χ^2 value in the full list but had good model fit once misfitting items had been removed from the scale. This highlights the importance of removing items iteratively. The threshold map for both scales was examined and after the rescoring outlined above there were no disordered thresholds.

To further the analysis at the individual item level, DIF analysis was conducted to ensure that no items were unduly influenced by either gender or school. None of the items in either list showed significant variation as a function of either gender or school after Bonferroni corrections were applied. Although no items had significant item misfit, one item, *exhausted*, was significant before a Bonferroni correction. The item characteristic curve for this word is shown in Figure 2. However, the low fit residual of -0.84, well below 2.5, combined with the strong overall model fit indicated that the item was appropriate for the scale. Therefore, it can be assumed that students of different gender or schools of the same word reading proficiency were answering the items in similar ways.

[Please insert Figure 2 approximately here]

Person fit

Similar to item fit, person fit is measured on two fronts, namely overall person fit and individual person fit. The overall person fit residual standard deviation for the refined irregular word list of 0.45 and 0.59 for the nonword list indicated good person fit. There were no individuals with person fit residuals over 2.5 in either list, indicating that all participants fit the models well. Six participants for the nonwords and 37 participants for the irregular word list got all the answers correct, and one student answered all the nonwords incorrectly. All of these students were considered extreme in the Rasch model but given the expectation of negative skew in this scale, these students were retained in the analysis.

Targeting and reliability

The Cronbach's alpha for the scales were good for the irregular word list (α = .88) and the nonword list (α = .89). Furthermore the Rasch measure of reliability, the person separation index (PSI), was also good for the irregular (PSI=.88) and nonword (PSI = .85) lists.

The mean person location for the irregular word list of 1.92, although high, is not unreasonable given the aim of the scale. The reason for this high mean person location is illustrated in the person-item distribution in Figure 3. The items on the lower portion of the figure are separating all students quite well, with the exception of the highest performing students. Despite the ceiling effect, it can be argued that the scale is doing exactly what is required. This scale differentiates between the low performing students quite effectively because of, rather than despite, the skew. Furthermore, the percentage of students correctly answering all questions (16%) is not so high that poor irregular word readers would not be separated from their higher performing peers.

[Please insert Figure 3 approximately here]

The mean person value location of 1.05 for the nonword list, where zero represents no skew, indicated that the scale was well targeted given the reason for its development. This can be seen in the person- item interaction graph in Figure 4. Once again there were few items targeted at the highest performers and the majority of items were targeted towards low to average performers.

[Please insert Figure 4 approximately here]

Tests of local independence

In order to ensure that the two word lists meet the psychometric requirements of both CTT and Rasch analysis, unidimensionality was assessed through both paradigms. Within CTT this was done using a principal axis factor analysis in conjunction with a Monte Carlo parallel analysis, as well as a *t*-test of items identified through the principal component residuals within the Rasch model. The KMO measure of sampling adequacy (Kaiser, 1974) indicated that the data for the irregular word list (KMO = .91) and nonword list (KMO = .87) were appropriate for EFA. Eignenvalues generated through the EFA were compared to those generated through a Monte Carlo parallel anlaysis and in the case of both lists unidimensionality was supported.

To further the initial EFA analysis for unidimensionality, the principal components residuals were tested through RUMM. *t*-tests were conducted to compare the results of the items with the highest positive residuals to the items with the highest negative residuals for both lists. In the case of the irregular word lists, *cooperation, certificates and somersaults* were compared to *tortoise, zero* and *stomach* and these *t*-tests found significant differences in less than 5% of simulations and hence unidimensionality for both word lists was supported. In the case of the nonword list *vun*, *reng*, and *olloy* were compared to *aip*, *joid*, and *poslip* and in less than 1% of simulations the t-tests were significant, thus unidimensionality for both lists was further supported.

Test-Retest Reliability

Thirty students were retested on the both lists one month after the original testing. The test-retest correlations for the irregular word list, *r*(29) = .88, *p* < .001, and the nonword list the nonword word list, *r*(29) = .86, *p* < .001, indicated that the test was stable and suitable for use in an applied setting.

Preliminary Norms for Both Word Lists

The irregular word list was refined from 36 items to 15 items and the Nonword list from 36 to 20 items within the guidelines of both CTT and IRT. Preliminary norms for the irregular word and nonword lists for Grade 3 students are shown in Table 4.

[Please insert Table 4 approximately here]

Discussion

The aim of the current study was to develop and psychometrically evaluate two word lists of nonwords and irregular words. For the purpose of psychometric validation, the word lists were tested within the Rasch paradigm. The irregular word list had good model fit and person separation index. Furthermore, the test was shown to be unidimensional both through exploratory factor analysis and Rasch analysis. Regardless of its eventual use as a predictor of LD or as an alternative explanation for poor reading in the case of non LD students, the 15 word irregular word list should be useful in identifying students with poor orthographic processing.

When compared to the irregular word list the 20 item nonword list was less skewed than the irregular word list. This list was longer to allow for the possibility of reading age norms being developed, as this has been shown to be important when assessing phonological processing (Siegel & Ryan, 1988). In the irregular word list there was only one cut-off score, regardless of reading proficiency. Consequently the cut-off points varied between reading-ages and sufficient spread at a higher level of the nonword list is important. The nonword list demonstrated unidimensionality and reliability, as well as meeting the psychometric requirements of Rasch analysis.

Similar to many scales that have been developed with the aim of screening for abnormal behaviour in a normal sample (Rijnsburger et al., 2006; Latimer et al., 2009; Lovibond &

Lovibond, 1995; Shea et al., 2009), skew was required if the word lists were appropriately targeted to the population of interest. A normally-distributed scale ensures that sufficient questions separate those students at a higher functioning level, as well as those at the lower end of the continuum. However, these extra items would not be beneficial as the test was only developed to identify the lower performing students. Nevertheless, as the scales were used in analyses based on the tenets of CTT, it is important to ensure that extreme skew is unlikely to affect these analyses by violating statistical assumptions. Ultimately, both lists gained psychometric support as per the tenets of Rasch analysis and, in the case of the nonword list, as per CTT as well.

An interesting result to come out of the Rasch analysis was that the three-level scoring structure was not appropriate. Originally, the idea was to award students who managed to get the word stem correct with an error in the suffix. However, the analyses indicated that this extra level of scoring provided no information that the dichotomous system did not already provide. It is possible that differentiating between these kinds of suffix-based errors and other errors may provide more information in a younger cohort. The ability to examine a scale at this level of detail highlights the importance of developing new tests through an IRT paradigm.

A limitation of the current study is that the norms for the refined word lists were based on students that were administered the full test. Furthermore, the sample was neither large nor sufficiently representative enough to be suitable for final norms. The norms presented here are merely preliminary and a larger study would need to be conducted to gain more reliable norms.

Similar to the work of Edwards and Hogben (1999), parsimonious irregular and nonword lists were found to be psychometrically sound measures. However, in contrast to the Edwards and Hogben scales, the scales developed in the current study were specifically targeted to Grade 3 students with poor reading ability, and resulted in shorter lists and more differentiation between targeted students. Overall, the results provided good support for the word lists' psychometric validity, with the caveat that the ceiling effect of the irregular word list needed to be taken into account in CTT analyses. If the irregular word list is used in an analysis that is not robust in regards to the violation of normality, then it may not be suitable for use.

Given the research on the importance of phonological processing (Fosker & Thierry, 2005; Goswami, 2011; Linklater et al., 2009; Muter & Snowling, 2009;) and orthographic processing (de Jong & Messbauer, 2011; O'Brien et al., 2011) in LD research, readily available tests of these processes are needed in Australian schools. It is anticipated that by providing the norms for these tests and making them freely available that this will help to reduce the research to practice gap in educational psychology.

References

Alexander, J. R. M., & Martin, F. (2000). Norming tests of basic reading skills. *Australian Journal of Psychology, 52*(3), 139-148. doi:10.1080/00049530008255381

Andrich, D., Sheridan, B. S., & Luo, G. (2004). RUMM2020: Rasch Unidimensional Measurement Models, Perth: RUMM Laboratory.

Andrich, D. (2005). The Rasch Model Explained. In S. Alagumalai, D. D. Curtis & N. Hungi (Eds.), Applied Rasch Measurement: A book of exemplars. Dordrecht, The Netherlands: Springer Kluwer.

Australian Council for Educational Research (ACER). (2000). *Labour market experiences of Australian youth. LSAY Briefing Reports* (LSAY Briefing; n.1), Retrieved from the ACER website: http://research.acer.edu.au/lsay_briefs/2

Australian Council for Education Research (ACER). (2008). *Progressive Achievement Tests in Reading: Comprehension and Reading* (4th ed.). Camberwell: ACER Press.

Ball, E. W. (1996). Phonological awareness and learning disabilities: Using research to inform our practice. *Advances in Learning and Behavioural Disabilities, 10*A 77-100.

Bechger, T. M., Maris, G., Verstralen, H. H. F. M., & Beguin, A. A. (2003). Using classical test theory in combination with item response theory. *Applied Psychological Measurement, 27*(5), 319-334. doi:10.1177/0146621603257518

Braten, I., Lie, A., Andreassen, R., & Olaussen, B. S. (1999). Leisure time reading and orthographic processes in word recognition among Norwegian third-and fourth-grade students. *Reading and Writing: An Interdisciplinary Journal, 11*(1), 65-88. doi:10.1023/A:1007976521114

Byrne, B., Freebody, P., & Gates, A. (1992). Longitudinal data on the relations of word-reading strategies to comprehension, reading time, and phonemic awareness. *Reading Research Quarterly, 27*(2), 140-151. Retrieved from http://www.jstor.org/pss/747683

Caravolas, M., Lervag, A., Mousikou, P., Efrim, C., Litavsky, M., Onochi-Quintanilla, E., . . . Hulme, C. (2012). Common Patterns of Prediction of Literacy Development in Different Alphabetic Orthographies. *Psychological Science, 23*(6), 678-686. doi: 10.1177/0956797611434536

Castles, A., Datta, H., Gayan, J., & Olson, R. K. (1999). Varieties of developmental reading disorder: Genetic and environmental influences. *Journal of Experimental Child Psychology, 72*, 73-94. doi:10.1006/jecp.1998.2482

Chan, D. W., Ho, C. S-H., Chung, K. K. H., Tsang S-M., Lee, S-H. (2012). The Hong Kong Behaviour Checklist for Primary Students: Developing a Brief Screening Measure. *International Journal of Disability Development and Education, 59*(2), 173-196. doi: 10.1080/1034912X.2012.676437

Cleary, M. J., & Scott, A. J. (2011). Developments in clinical neuropsychology: Implications for school psychological services. *Journal of School Health, 81*(1), 1-7. doi:10.1111/j.1746-1561.2010.00550.x

Conti-Ramsden, G., & Durkin, K. (2007). Phonological short-term memory, language and literacy: developmental relationships in early adolescence in young people with SLI. *Journal of Child Psychology & Psychiatry, 48*(2), 147-156. doi:10.1111/j.1469-7610.2006.01703.x

Decker, S. L. (2008). School neuropsychology consultation in neurodevelopmental disorders. *Psychology in the Schools, 45*(9), 799-811. doi:10.1002/pits.20327 de Jong, P. F. & Messbauer, V. C. (2011). Orthographic context and the acquisition of orthographic knowledge in normal and dyslexia readers. *Dyslexia, 17*(2), 107-122.

Edelen, M. O., Jaycox, L. H., McCaffrey, D. F., & Marshall, G. N. (2006). *Improving the measurement of socially unacceptable attitudes and behaviours with item response theory*. RAND Working Paper WR-383 Retrieved from http://www.rand.org/pubs/working_papers/WR383/

Edelen, M. O., & Reeve, B. B. (2007). Applying Item Response Theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research, 16*, 5-18. doi:10.1007/s11136-007-9198-0

Edwards, V. T., & Hogben, J. H. (1999). New norms for comparing children's lexical and non-lexical reading: A further look at sub-typing dyslexia. *Australian Journal of Psychology, 51*(1), 37-49. doi:10.1080/00049539908255333

Erhart, M., Hasquist, C., Auquier, P., Rajmil, L., Power, M., Ravens-Sieberer, U., et al. (2009). A comparison of Rasch item-fit and Cronbach's alpha item reduction analysis for the development of Quality of Life scale for children and adolescents. *Child: Care, Health and Development, 36*(4), 473-484. doi:10.1111/j.1365-2214.2009.00998.x

Fink, P., Ornbol, E., Huyse, F. J., De Jonge, P., Lobo, A., Herzog, T., . . . Hansen M. S. (2004). A brief diagnostic screening instrument for mental disturbances in general medical wards. *Journal of Psychosomatic Research, 57*(1), 17-24. doi:10.1016/S0022-3999(03)00374-X

Fosker, T., & Thierry, G. (2005). Phonological oddballs in the focus of attention elicit a normal P3b in dyslexic adults. *Cognitive Brain Research, 24*(3), 467-475. doi:10.1016/j.cogbrainres.2005.02.019

Friedmann, N., & Lukov, L. (2008). Developmental surface dyslexias. *Cortex, 44*(9), 1146-1160. doi:10.1016/j.cortex.2007.09.005

Fuchs, D., & Fuchs, L. S. (1995). What's `special' about special education? *Phi Delta Kappan, 76*(7), 522. Retrieved from http://goo.gl/N3kLC

Goswami, U. (2011). A temporal sampling framework for developmental dyslexia. *Trends in Cognitive Sciences, 15*(1), 3-10. doi:10.1016/j.tics.2010.10.001

Griffiths, Y. M., & Snowling, M. J. (2002). Predictors of exception word and nonword reading in dyslexic children: The severity hypothesis. *Journal of Educational Psychology, 94*(1), 34-43. doi:10.1037/0022-0663.94.1.34

Ho, C. S. H., Chan, D. W., Tsang, S. M. & Lee, S. (2002). The cognitive profile and multiple deficit hypothesis in Chinese developmental dyslexia. *Developmental Psychology, 38*(4), 543-553. 0.1007/s11145-007-9084-8

Hulme, C., Bowyer-Crane, C., Carroll, J. M., Duff, F. J. & Snowling, M. J. (2012). The Causal Role of Phoneme Awareness and Letter-Sound Knowledge in Learning to Read: Combining Intervention Studies with Mediation Analyses. *Psychological Science, 23*(6), 572-577. doi: 10.1177/0956797611435921

Jarmulowicz, L., Taran, V. L., & Hay, S. E. (2007). Third graders' metalinguistic skills, reading skills, and stress production in derived English words. *Journal of Speech, Language, and Hearing Research, 50,* 1593-1605. doi:10.1044/1092-4388(2007/107)

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika, 30*(1), 31-35. 10.1007/BF02291575

Kessler, R. C., Andrews, G., Colpe, L. J., Hiripi, E., Mroczek, D. K., Normand, S. L. T., . . . Zaslavsky A. M. (2002). Short screening scales to monitor population prevalences and trends in non-specific psychological distress. *Psychological Medicine, 32*, 959-976. doi:10.1017/S0033291702006074

Kirby, J. R., Roth, L., Desrochers, A., & Lai, S. S. V. (2008). Longitudinal predictors of word reading development. *Canadian Psychology, 49*(2), 103-110. doi:10.1037/0708-5591.49.2.103

Kortering, L. J., Braziel, P. M., & McClannon, T. W. (2010). Career ambitions: A comparison of youth with and without SLD. *Remedial & Special Education, 31*(4), 230-240. doi:10.1177/0741932508324404

Kovelman, I., Norton, E. S., Christodoulou, J. A., Gaab, N., Lieberman, D. A., Triantafyllou, C., Wolf, M., Whitfield-Gabrieli, S. & Gabrieli, J. D. E. (2012). Brain Basis of Phonological Awareness for Spoken Language in Children and Its Disruption in Dyslexia. *Cerebral Cortex, 22,* 754-764. Doi:10.1093/cercor/bhr094

Latimer, S., Covic, T., Cumming, S. R., & Tennant, A. (2009). Psychometric analysis of the Self-Harm Inventory using Rasch modelling. *BMC Psychiatry, 9,* 53. doi:10.1186/1471-244X-9-53.

Leigh, A., & Ryan, C. (2008). How has school productivity changed in Australia? Retrieved March 15, 2009 from http://econrsss.anu.edu.au/~aleigh/pdf/SchoolProductivity.pdf

Linklater, D. L., O'Connor, R., & Palardy, G. J. (2009). Kindergarten literacy assessment of English Only and English language learner students: An examination of the predictive validity of three phonemic awareness measures. *Journal of School Psychology, 47*, 369-394. doi:10.1016/j.jsp.2009.08.001

Lo Bianco, J., Scull, J., & Ives, D. (2008). Oxford Word List Plus. Oxford University Press. Retrieved 1/10/2008 from http://www.oup.com.au/primary/learning/thesuccessfulteacher

Lovibond, P. F., & Lovibond, S. H. (1995). The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. *Behaviour Research and Therapy, 33,* 335-343. doi:10.1016/0005-7967(94)00075-U

Martens, V. E. G., & de Jong, P. F. (2006). The effect of word length on lexical decision in dyslexic and normal reading children. *Brain and Language, 98,* 140-149. doi:10.1016/j.bandl.2006.04.003

Martin, F., & Pratt, C. (2001). *Martin and Pratt nonword reading test*. Camberwell, VIC: Australian Council of Educational Research.

McGee, R., Prior, M., Williams, S., Smart, D., & Sanson, A. (2002). The long-term significance of teacher-rated hyperactivity and reading ability in childhood: findings from two longitudinal studies. *Journal of Child Psychology & Psychiatry, 43*(8), 1004-1017. doi:10.1111/1469-7610.00228

Morrison, A. S. (1992). *Screening in chronic disease* (2nd ed.). New York: Oxford University Press.

Muter, V., & Snowling, M. (2009). Children at familial risk of dyslexia: Practical implications from an at-risk study. *Child and Adolescent Mental Health, 14*(1), 37-41. doi:10.1111/j.1475-3588.2007.00480.x

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology, 3*(1), 1-18. doi:10.1016/0022-2496(66)90002-2

O'Brien, B. A., Wolf, M., Miller, L. T., Lovett, M. W., & Morris, R. (2011). Orthographic processing efficiency in developmental dyslexia: an investigation of age and treatment factors at the sublexical level. *Annals of Dyslexia, 61*, 111-135.

Pallant, J. F., & Tennant, A. (2007). An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS). *British Journal of Clinical Psychology, 46*, 1-18. doi: 10.1348/014466506X96931

Pammer, K., Lavis, R., Hansen, P., & Cornelissen, P. L. (2004). Symbol-string sensitivity and children's reading. *Brain & Language, 89*, 601-610. doi:10.1016/j.bandl.2004.01.009

PASW (2009). Predictive Analytics SoftWare Statistics (Version 17): PASW.

Perfetti, C. A., & Bolger, D. J. (2004). The brain might read that way. *Scientific Studies of Reading, 8*(3), 293-304. Retrieved from http://goo.gl/vxC7O

Reeve, B.B. 2002. An Introduction to Modern Measurement Theory., National Cancer Inst..

Rijnsburger, A. J., Essink-Bot, M. L., van As, E., Cockburn, J., & de Koning, H. J. (2006). Measuring psychological consequences of screening: Adaption of the Psychological Consequences Questionnaire into Dutch. *Quality of Life Research, 15,* 933-940. doi:10.1007/s11136-005-5093-8

Rispens, J., & Parigger, E. (2010). Nonword repetition in Dutch speaking children with specific language impairment with and without reading problems. *British Journal of Developmental Psychology, 28*(1), 177-188. doi:10.1348/026151009X482633

Shea, T. L., Tennant, A., & Pallant, J. F. (2009). Rasch model analysis of the Depression, Anxiety and Stress Scales (DASS). *BMC Psychiatry, 9,* 21-30. doi:10.1186/1471-244X-9-21.

Siegel, L. S. (1989). IQ is irrelevant to the definition of learning disabilities. *Journal of Learning Disabilities, 22(*8), 469-478, 486. doi:10.1177/002221948902200803

Siegel, L. S., & Ryan, E. B. (1988). Development of grammatical-sensitivity, phonological, and short-term memory skills in normally achieving and learning disabled children. *Developmental Psychology, 24*(1), 28-37. Retrieved from http://goo.gl/zTX2i

Skues, J. L., Cunningham, E. G. (2011). A contemporary review of the definition, prevalence, identification and support of learning disabilities in Australian schools. *Australian Journal of Learning Difficulties, 16*(2), 159-180. doi: 10.1080/19404158.2011.605154.

Stage, S. A., Abbott, R. D., Jenkins, J. R., & Berninger, V. W. (2003). Predicting response to early reading intervention from verbal IQ, reading-related language abilities, attention ratings, and verbal IQ-word reading discrepancy: Failure to validate discrepancy method. *Journal of Learning Disabilities, 36*(1), 24-33. doi:10.1177/00222194030360010401

Stanovich, K. E. (1999). The sociopsychometrics of learning disabilities. *Journal of Learning Disabilities, 32*(4), 350-361. doi:10.1177/002221949903200408

Stanovich, K. E., & Siegel, L. S. (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. Retrieved from http://goo.gl/Zwe8T

Stanovich, K. E., & West, R. F. (1989). Exposure to print and orthographic processing. *Reading Research Quarterly, 24*(4), 402-433. Retrieved from http://www.jstor.org/pss/747605

Svensson, I., & Jacobson, C. (2006). How persistent are phonological difficulties? A longitudinal study of reading retarded children. *Dyslexia, 12*(1), 3-20. doi:10.1002/dys.296

Undheim, A. M., & Sund, A. M. (2008). Psychosocial factors and reading difficulties: Students with reading difficulties drawn from a representative population sample. *Health and Disability, 49*, 377-384. doi:10.1111/j.1467-9450.2008.00661.x

Watkins, M. (2008). Monte Carlo PCA for Parallel Analysis (Version 2.3).

Woolley, G. (2007). A comprehension intervention for children with reading comprehension difficulties. *Australian Journal of Learning Disabilities, 12*(1), 43-50. doi:10.1080/19404150709546829

Ziegler, J. C., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: A psycholinguistic grain size theory. *Psychological Bulletin, 131*(1), 3-29. doi:10.1037/0033-2909.131.1.3

Ziegler, J. C., Perry, C., & Coltheart, M. (2000). The DRC model of visual word recognition and reading aloud: An extension to German. *European Journal of Cognitive Psychology, 12*(3), 413-430. doi:10.1080/09541440050114570