Similar but different : Differences in comprehension diagnosis on the Neale Analysis of Reading Ability and the York Assessment of Reading for Comprehension

Colenbrander, Danielle, Nickels, Lyndsey and Kohnen, Saskia

Similar but different: Differences in comprehension diagnosis on the Neale Analysis of Reading Ability and the York Assessment of Reading for Comprehension

Danielle Colenbrander, Lyndsey Nickels and Saskia Kohnen

Department of Cognitive Science and ARC Centre of Excellence in Cognition and its Disorders (CCD), Macquarie University, Australia

Corresponding author: Dr Danielle Colenbrander, School of Experimental Psychology, 12a Priory Road, University of Bristol, BS8 1TU, United Kingdom
(e-mail: d.colenbrander@bristol.ac.uk).

Acknowledgements: The authors wish to thank the staff, parents and students of the participating school. During this research, Danielle Colenbrander was funded by a Macquarie University Research Excellence (MQRES) PhD scholarship, Saskia Kohnen was funded by a Macquarie University Research Fellowship, and Lyndsey Nickels was funded by an Australian Research Council Future Fellowship (FT120100102).

Abstract

Background: Identifying reading comprehension difficulties is challenging. There are many comprehension tests to choose from, and a child's diagnosis can be influenced by various factors such as a test's format and content and the choice of diagnostic criteria. We investigate these issues with reference to the Neale Analysis of Reading Ability (NARA) and the York Assessment of Reading for Comprehension (YARC).

Methods: Ninety-five children were assessed on both tests. Test characteristics were compared using Principal Components and Regression analyses as well as an analysis of passage content.

Results: NARA comprehension scores were more dependent on decoding skills than YARC scores, but children answered more comprehension questions on the NARA and passages spanned a wider range of difficulty. Consequently, 15-34% of children received different diagnoses across tests, depending on diagnostic criteria.

Conclusion: Knowledge of the strengths and weaknesses of comprehension tests is essential when attempting to diagnose reading comprehension difficulties.

Introduction

The aim of reading instruction is to ensure that children can understand what they read. Nevertheless, some children struggle to do so, and this can impact on academic achievement (Cain & Oakhill, 2006a; McLaughlin, Speirs, & Shenassa, 2012; Ricketts, Sperring & Nation, 2014).

According to the Simple View of Reading (Gough & Tunmer, 1986; Hoover & Gough, 1990), there are two main reasons why a child may fail to comprehend written text: children may have difficulties sounding out and/or recognising written words (decoding), or they may have difficulties comprehending oral language. Either difficulty can cause poor reading comprehension, and some children may have deficits in both areas.

Children with different reading profiles require different intervention approaches (Garcia & Cain, 2014; Hulme & Snowling, 2011). For example, a child who has decoding difficulties is likely to benefit from phonics intervention and/or sight word training (depending on the subtype of decoding problem; McArthur et al., 2015), whereas a child who has difficulties with oral language may benefit from oral language intervention (Clarke, Snowling, Truelove & Hulme, 2010; Hulme & Snowling, 2011). Therefore, it is crucial to be able to determine not only *which* children have poor reading comprehension, but also *why* this is the case.

The first step in this process is generally to use a standardized assessment of reading comprehension. However, while many different reading comprehension assessments are available, correlations between scores on these assessments can be surprisingly low (Keenan & Meenan, 2014). One reason for this is that reading comprehension assessments differ along a variety of dimensions, such as passage length and response format (Francis, Fletcher, Catts & Tomblin, 2005; Keenan et al., 2008). Linguistic factors such as vocabulary level and syntactic complexity can affect text difficulty, as will topic familiarity (Duke, 2005; Graesser,

McNamara & Kulikowich, 2011). Therefore, in order to interpret results appropriately, it is crucial to understand the features and characteristics of widely used reading comprehension assessments (Cain & Oakhill, 2006a; Cutting & Scarborough, 2006; Keenan, Betjemann, & Olson, 2008; Keenan & Meenan, 2014; Nation & Snowling, 1997).

In this paper, we explore these issues in relation to the Neale Analysis of Reading Ability Third Edition (NARA; Neale, 1999) and the Australian edition of the York Assessment of Reading for Comprehension (YARC; Snowling et al., 2012). Until recently, successive editions of the NARA were widely used in Australia and the UK, in both educational and research contexts. However, the YARC is increasingly replacing the NARA in both these countries (GL Assessment, 2014; Howe, 2013; Ricketts, 2014).

In our comparison of these assessments, we will consider different profiles of reading impairment. The Simple View of Reading predicts four profiles of reading ability: successful readers; generally poor readers with both poor decoding and poor comprehension skills; poor decoders with poor decoding skills, but intact oral language skills; and poor comprehenders with poor oral language skills, but intact decoding skills (Gough & Tunmer, 1986; Hoover & Gough, 1990). We will explore the extent to which diagnosis of such profiles is consistent across tests.

The issue of diagnosis is complicated by the fact that different diagnostic criteria are used across research and clinical settings (Clarke, Henderson & Truelove, 2010). For example, some studies of poor comprehenders have selected participants with comprehension scores 6 or 12 months below chronological age (Cain & Oakhill, 1999; 2006b) and others those whose scores fall one standard deviation below the mean (Ricketts et al., 2014) or below the 25th percentile (Catts, Adlof & Weismer, 2006). However in clinical practice, much more stringent cut-offs are likely to be used. For example, the Diagnostic and Statistical Manual of Mental Disorders (DSM-5; American Psychiatric Association, 2013)

suggests that a standard score of at least 1.5 standard deviations below the mean is necessary for "the greatest diagnostic certainty" on a particular assessment (American Psychiatric Association, 2013).

Diagnosis of the poor comprehender profile is particularly inconsistent. Some studies consider only reading comprehension scores, and make no distinction on the basis of decoding ability. Other studies require children to have age appropriate decoding scores, or require a set discrepancy between reading comprehension and decoding scores (Keenan, Hua, Meenan, Pennington, Willcut & Olson, 2014). The choice of decoding assessment is also controversial, with some studies advocating the use of tests of word reading accuracy (Garcia & Cain, 2014) and others the use of tests of nonword reading accuracy (Hoover & Gough, 1990) or reading fluency (Silverman, Speece, Harring & Ritchey, 2013).

All of these factors will influence the number of children falling into each profile category, and may also influence the consistency of diagnosis across tests. Therefore, with specific reference to the NARA and the YARC, we will explore how diagnostic criteria interact with test format and content in reading comprehension diagnosis.

Method

Participants

Ninety-five children (55 female) aged 8 to 12 participated in the study. Children were in Australian school Grades 3 to 6 (equivalent to four to seven years of schooling, as the first year of schooling is a Kindergarten year). They spoke English as their primary language and had been attending school in Australia since Kindergarten.

Sixty-five participants came from a Catholic primary school in a middle-class area of Sydney. These children were assessed during the screening phase of another study, where teachers of Grades 3, 4 and 5 had been asked to nominate children who had average decoding

abilities, and either good or poor reading comprehension skills. Nominated children were assessed if they gave verbal consent and had returned parental permission slips.

A further 30 participants were recruited via a university-based Brain Science club. Parents contacted researcher directly if interested in participating in experiments which were advertised in a newsletter. The 30 participants were initially recruited as controls for another study. These children were in Grades 4, 5 and 6 of both government and private schools in Sydney. All participants returned parental permission slips and gave verbal consent.

Assessment

Participants were assessed individually by the first author in a single session of approximately 45 minutes, administered either in a quiet room at the child's school or in a testing laboratory at the University. Participants were administered the NARA, followed by the Test of Word Reading Efficiency (TOWRE; Torgesen, Wagner, & Rashotte, 1999), the Castles and Coltheart Reading Test 2 (CC2; Castles et al., 2009) and the YARC.

Form 1 of the NARA (Third Edition; Neale, 1999) was administered according to the standard instructions. Children read aloud a series of short passages and provided verbal responses to several open-ended comprehension questions. Following standardized administration procedures, passage reading was timed and children's reading errors were corrected by the tester. If a child made a specified number of errors on a passage (usually 16 or more), no comprehension questions were asked for that passage and the test was discontinued. If fewer than the specified number of errors were made, the comprehension questions were asked and the child progressed to the next passage. Reliability and validity information is reported in the manual and the test has Australian norms (Neale, 1999).

The test returns three scores, a comprehension score, a reading accuracy score and a reading rate score. For this study, we did not use the rate scores. This was because accuracy errors are prompted by the tester, so a child's reading rate score will be affected by the speed

of prompting. Additionally, text reading fluency is likely to be influenced by comprehension abilities – children who can comprehend well may be able to use contextual information to help them read aloud (Nation & Snowling, 1997). TOWRE scores are free from these biases and therefore were the only measure of reading fluency used in our analysis (see below).

There are three versions of the YARC, each catering to a different age level – YARC Early Years (Hulme et al., 2009), YARC Passage Reading (Snowling et al., 2012), and YARC Secondary (Snowling et al., 2010). This allows children to be tested on comparable assessments throughout their schooling. We focused on YARC Passage Reading because like the NARA, this version was designed to be used with primary school children. We report data using the Australian standardization of the YARC.

Form A of the YARC Passage Reading Australian Edition (Snowling et al., 2012) was administered according to standard instructions. Children read passages aloud and answered open-ended comprehension questions, similar to the NARA. Accuracy errors were corrected. However, the discontinuation rule was different. On the YARC, children read aloud an initial passage matched to their grade level[1]. If a child made less than a specified number of accuracy errors (usually 20 errors) on a passage, s/he was asked a series of open-ended questions about the passage. If 5 or more of these questions were answered correctly, s/he was asked to read a passage one level higher than the starting passage, and asked comprehension questions about this passage. If accuracy or comprehension cut-offs were not met, s/he was administered a passage the level below the starting passage, until s/he had read 2 passages with at least 2 comprehension questions correct for at least one level. The test contained both fiction and non-fiction passages and children read at least one passage of each type. Like the NARA, the test returned a comprehension score, an accuracy score and a rate

[1] A child's initial passage level can also be determined by administering the Single Word Reading Test (Foster, 2007).

score, but we used only the comprehension and accuracy scores. Reliability and validity information is available in the YARC manual (Snowling et al., 2012).

We also assessed children's word reading accuracy using the CC2 (Castles et al., 2009) and fluency using the TOWRE (Torgesen et al., 1999). The CC2 returns scores for reading of regular words, irregular words, and nonwords. We report only the irregular word and nonword scores, which index lexical and sublexical reading skills respectively. Forty nonwords and 40 irregular words were presented for reading aloud interspersed with each other, and in order of increasing difficulty. A stopping rule of five consecutive errors applied to each item type.

The TOWRE has two subtests: A Sight Word Efficiency subtest where children were asked to read lists of words as fast as they could, and a Phonemic Decoding Efficiency subtest, where the stimuli were nonwords. The child's score was the number of stimuli read correctly within 45 seconds. Form B was presented and Australian norms were used (Marinus, Kohnen & McArthur, 2013).

Results

Descriptive statistics

Standard scores (mean of 100 and standard deviation of 15) were used to compare the results of our sample to that of the published normative samples for each test.

[Insert Table 1 about here]

Reading accuracy and fluency.

Teachers had been asked to nominate participants with no known history of reading difficulties. Consequently, the majority of the reading accuracy and fluency scores were close to the standardization sample mean (One-sample z-tests: NARA Accuracy, $z = 0.80$, $p = 0.42$, NARA Rate, $z = 0.77$, $p = 0.44$, YARC Rate, $z = -1.50$, $p = 0.13$, TOWRE Sight Word Efficiency, $z = 0.01$, $p = 0.42$, TOWRE Phonemic Decoding Efficiency, $z = -1.24$, $p = 0.21$;

CC2 Irregular words, $z = -0.06$, $p = 0.88$). However, YARC Accuracy scores were significantly lower than the test's standardization sample, $z = -3.60$, $p < 0.001$, as were CC2 Nonword Reading scores, $z = -4.56$, $p < 0.001$. Indeed, a substantial proportion of individual children were below average on at least one accuracy or fluency score. For example, 25% of the sample were classified as below average on CC2 nonwords (see Table 1).

Reading comprehension.

Given that screening was intended to identify children with specific reading comprehension difficulties (as well as age-matched controls), mean standard scores for reading comprehension were significantly lower than the standardization means (i.e., < 100) for both the NARA, $z = -7.32$, $p < 0.001$, and the YARC, $z = -2.33$, $p = 0.02$. In addition, participants' standard scores on the NARA were significantly lower than their standard scores on the YARC, $t(94) = -7.22$, $p < 0.001$.

To what extent do diagnoses of reading difficulties differ across tests?

We calculated the number of children falling into each of the four Simple View categories (successful readers, generally poor readers, poor decoders, poor comprehenders) on each test. Because we did not assess children's oral language abilities, discrepancies between reading and reading comprehension scores were used as an indication of oral language skills. Thus, successful readers were children whose decoding accuracy or fluency scores and comprehension scores were within the average range for their age. Generally poor readers were those whose decoding accuracy/fluency and reading comprehension scores fell below the average range, while poor comprehenders were those with age-appropriate accuracy/fluency but below average reading comprehension, and poor decoders were those with below average accuracy/fluency and average reading comprehension.

We calculated the prevalence of each profile within our sample using the different measures of comprehension and decoding (see Table 2). Because diagnosis of the poor

comprehender profile is particularly controversial, we calculated poor comprehender prevalence figures using both a cut-off criterion (decoding standard score above 85, comprehension score below 85) and a discrepancy criterion (decoding standard score above 85, comprehension score below 85, comprehension score at least 15 standard score points below decoding).

We then calculated the consistency of diagnosis – the proportion of children who received the *same* diagnosis on both the NARA and the YARC, when different measures of decoding were used[2]. In order to determine whether consistency levels changed when different diagnostic criteria were used, we compared consistency levels using cut-offs of one standard deviation below the mean (standard scores of 85 or below), and 1.5 standard deviations below the mean (standard scores of 78 or below, consistent with DSM-5 guidelines).

Prevalence of different reading profiles

[Insert Table 2 about here]

More children were diagnosed as successful readers and poor decoders on the YARC. However, a greater proportion of children were diagnosed as generally poor readers and poor comprehenders on the NARA (see Table 2).

In terms of poor comprehender diagnosis, fewer children were diagnosed when the more stringent discrepancy criterion was used. Nonetheless, more children were diagnosed as poor comprehenders on the NARA than on the YARC, regardless of which diagnostic criteria were used.

The choice of decoding assessment made relatively little difference to prevalence figures. Slightly more children were diagnosed as poor comprehenders or successful readers when word or irregular word reading measures were used and conversely, slightly more were

[2] For the sake of simplicity, we used a cut-off definition of the poor comprehender profile for this analysis.

diagnosed as generally poor readers or poor decoders when nonword measures were used, but differences were small.

Consistency of diagnosis

Previous studies have revealed that on average, between 35-65% of children receive the same diagnosis across different reading comprehension tests (Keenan et al., 2014; Keenan & Meenan, 2014). Table 3 shows that the overall consistency between the NARA and YARC in our sample varied from 66% to 85%, depending on which decoding assessments and diagnostic criteria were used. This is higher than previous research, and may be due to the fact that the NARA and YARC are relatively similar in format. Consistency was higher when a more stringent cut off (1.5 standard deviations) was used.

[Insert Table 3 about here]

Consistency across tests was lower when test-internal text reading accuracy measures were used (66%) rather than test-independent measures of word or nonword reading (76%), most likely because text reading can be influenced by comprehension skill – children can use contextual information from the text to help them predict unfamiliar words (Nation & Snowling, 1997). Thus, use of test-internal text accuracy measures may have magnified the differences between the two reading comprehension assessments. However, the choice of decoding measure seemed to make little difference to consistency when independent, word-level measures were used.

What factors contribute to differences in reading profile diagnosis?

Dependence on decoding abilities.

Previous research has shown that a child's comprehension score on the NARA is dependent to some extent on their decoding abilities (Spooner, Baddeley, & Gathercole, 2004). Children are discontinued on the NARA if they make a specified number of accuracy errors, therefore poor decoders are likely to read fewer passages on the NARA than good

readers of the same age and answer fewer comprehension questions, which may limit their comprehension score (Spooner et al., 2004). However, on the YARC, children's comprehension scores are always calculated based on the same number of questions, regardless of their level of reading accuracy. This may be why more children are diagnosed as generally poor readers on the NARA, and poor decoders on the YARC. We explore this possibility below.

Principal components analysis.

We used principal components analysis with oblique rotation to determine to what extent NARA or YARC comprehension scores loaded on a reading accuracy or fluency factor, using the percentile ranks for NARA and YARC comprehension and accuracy scores, the TOWRE Sight Word and Phonemic Decoding Efficiency scores, and the CC2 Irregular Word and Nonword scores.

Our sample size was adequate (Kaiser-Meyer-Olkin = 0.85; Field, 2009) and correlations between items sufficiently large (Bartlett's test of sphericity, $X^2(28) = 464.38$, $p < 0.001$) to justify the use of principal components analysis.

Two factors emerged with eigenvalues greater than 1, that together accounted for 71.43% of the variance (see Table 4). Given multiple comparisons and a sample size of approximately 100, a factor loading should be greater than 0.512 to be considered significant at $\alpha = 0.01$ (Stevens, 2002). Using this criterion, CC2 Nonword and Irregular word scores, YARC and NARA accuracy scores, and TOWRE Sight Word and Phonemic Decoding Efficiency scores formed a single factor, which we refer to as the decoding factor. The NARA and YARC comprehension scores formed the second factor, the comprehension factor. Neither comprehension score loaded significantly on the decoding factor. However, it is worth noting that the loading on the decoding factor was greater for the NARA (0.16) than for the YARC (-0.07).

[Insert Table 4 about here]

Regression analysis.

We ran a simple linear regression analysis to determine the extent to which reading comprehension scores were predicted by decoding scores. Given that all our decoding accuracy and fluency measures loaded onto a single factor, we created a composite decoding score by averaging all these measures.

[Insert Table 5 about here]

The decoding composite accounted for 21% of the variance in NARA comprehension scores, but accounted for only 9% of the variance in YARC comprehension scores (see Table 5). Therefore, NARA comprehension scores do appear to rely to a greater extent on a child's decoding abilities. However, this does not explain why more children are diagnosed as poor comprehenders on the NARA than on the YARC. A possible explanation is that children may read more passages and answer more comprehension questions on the NARA than on the YARC.

Differences in number of comprehension questions answered on the NARA and the YARC by readers of all ability levels.

At all grade levels, children answered on average at least 27 questions on the NARA (see Table 6). A large proportion of children in each grade answered 32 or more questions (46% in Grade 3, 79% in Grade 4 and 83% in Grades 5 and 6), double the number of questions they answered on the YARC. It is possible that the larger sampling of comprehension ability on the NARA resulted in greater sensitivity to small differences in comprehension skill, leading to more children being diagnosed as poor comprehenders on the NARA than on the YARC[3].

[3] It is worth noting that when children are very poor readers, they may answer *fewer* questions on the NARA than on the YARC. Two participants whose reading accuracy and fluency scores were all at least one standard deviation below the mean (one in Grade 5 and one in Grade 6) each read only a single passage and answered only 8 questions before discontinuing on the NARA. However, the vast majority of children answered many more questions on the NARA.

[Insert Table 6 about here]

Relative difficulty of passages.

It is also possible that the NARA passages themselves are more difficult to comprehend than the YARC passages. To explore this possibility, we examined the difficulty levels of NARA and YARC passages using CohMetrix version 3.0 (McNamara, Louwerse, Cai, & Graesser, 2011), a computational tool which analyzes texts on numerous different variables. Based on previous research, we report a number of key variables in our analyses. We report passage length in words and measures of word length and frequency. Passages containing shorter and more frequent words are easier to decode and comprehend (Compton, Appleton, & Hosp, 2004; Freebody & Anderson, 1983; Graesser, McNamara, & Kulikowich, 2011; Ozuru, Rowe, O'Reilly, & McNamara, 2008).

We report measures of syntactic complexity, including sentence length in words, left embeddedness (the mean number of words before the main verb), and number of modifiers per noun phrase (e.g. adjectives). Higher values of all of these measures indicate higher levels of sentence complexity (Graesser, McNamara & Kulikowuch, 2011). We also report Flesch-Kincaid grade levels. This widely used index of text difficulty is calculated using word length in syllables and sentence length in words. In addition, we report measures of referential cohesion, and the mean ratio of given information to new information within sentences. Both of these variables have been shown to affect passage difficulty (Graesser et al., 2011; Keenan et al., 2008; McNamara, Graesser, & Louwerse, 2012). Results are shown in Table 7.

[Insert Table 7 about here]

Because different children complete different passages on the tests, it is difficult to generalize about minimum and maximum difficulty levels for each test. However, YARC passages are generally longer in terms of number of words and sentences. In contrast, levels 3

to 6 of the NARA contain on average longer and less frequent words, and the referential cohesion scores of these NARA passages are well below those of the YARC. Furthermore, levels 3 to 5 of the NARA contain more modifiers per noun phrase than the equivalent YARC levels.

Texts which are lower in referential cohesion may be more difficult to understand because readers are required to carry out inferences to connect relevant elements of the text (Britton & Gulgoz, 1991). This process is likely to be particularly challenging if readers are also faced with unfamiliar vocabulary items or more complex sentences (Graesser et al., 2011; McNamara et al., 2012). This suggests that the higher-level NARA passages may be more challenging than the higher-level YARC passages. When we consider that children read more passages on the NARA than the YARC, this suggests that children may have been exposed to a wider range of difficulty levels across passages on the NARA than the YARC.

Discussion and conclusions

Reading comprehension assessments are used to make decisions about intervention and service delivery. They are also used by researchers to select participants and draw theoretical conclusions. However, reading comprehension is a complex construct, and children's reading comprehension scores can vary not only because of individual differences in underlying skills, but also because of differences between the reading comprehension assessments themselves. Therefore, the choice of reading comprehension assessment has important practical and theoretical implications. In this study, we aimed to explore these issues by assessing a sample of children in the upper primary years on two commonly used comprehension assessments: the NARA (Neale, 1999) and the YARC (Snowling et al., 2012)

Firstly, we found that NARA comprehension scores were more dependent on decoding skills than YARC comprehension scores. We suggest that this is because the number of comprehension questions answered on the NARA is dependent on a child's

accuracy score – the test is discontinued when a child makes 16 or more accuracy errors. By contrast, a child's comprehension score on the YARC is always calculated on the basis of the same number of questions and passages. Thus, for children who are poor decoders, the YARC may be a more accurate estimate of reading comprehension skills than the NARA.

While the format of the YARC is an advantage when it comes to assessing the reading comprehension skills of children with poor decoding abilities, it may be a disadvantage for assessing the comprehension of children with age-appropriate decoding skills. We found that in general, children read more passages and answered up to twice as many comprehension questions on the NARA than they did on the YARC. In addition, the higher-level NARA passages appeared to be more linguistically challenging than the equivalent YARC passages. This raises the possibility that the NARA is more sensitive to subtle differences in reading comprehension skill.

Of course, we cannot rule out the possibility that the NARA is in fact over-diagnosing reading comprehension difficulties. Some children who perform well on developmentally appropriate reading comprehension passages may perform poorly on the more challenging passages of the NARA, which may be taxing skills beyond those expected for a child of that age. Further research is required to determine how well children's comprehension scores on the NARA and YARC reflect their academic performance in the classroom, and longitudinal research exploring the predictive abilities of the NARA and the YARC is also warranted.

When making reading comprehension diagnoses, it is important to consider not only the comprehension test used, but also the measure of decoding. Our results show that it is preferable to use separate word or nonword reading measures, rather than the text reading measures associated with the comprehension tests themselves.

The choice of diagnostic criteria will also influence comprehension diagnosis. Our findings indicated that a stricter diagnostic criterion (1.5 standard deviations below the mean)

led to greater consistency between the NARA and the YARC. However, other studies have found that consistency between tests can be low even when a relatively stringent cut-off point is used (Keenan et al., 2014; Keenan & Meenan, 2014).

Whatever the criterion, the choice of a cut-off point is ultimately arbitrary. Resources for intervention and research are limited, and cut-off points allow decisions about service delivery and study inclusion to be made in an objective manner. However, there may be little difference in skill between a child who falls just below a particular cut-off point, and a child who falls just above it. Therefore, our findings reinforce recommendations that diagnosis of a reading comprehension difficulty should not be made on the basis of a single assessment (Bowyer-Crane & Snowling, 2005; Cain & Oakhill, 2006a; Keenan & Meenan, 2014).

We do not suggest that a diagnosis of comprehension difficulty should only be made if a child scores poorly on more than one comprehension assessment. Rather, we suggest that multiple comprehension assessments can provide additional qualitative information about a child's comprehension skills. If a child receives different diagnoses across tests, the clinician or researcher should interpret this information based on their knowledge of the child, the child's reading accuracy or fluency skills, the testing circumstances, and the strengths or weaknesses of the assessments used, such as we have provided here. Ideally, initial diagnosis should be followed by detailed assessment of underlying skills (for example, oral language skills), to ensure that a child receives appropriate intervention (Keenan & Meenan, 2014).

Nonetheless, it is important to note some limitations of this study. Firstly, this research was based on a sample of children recruited for another study and it was therefore not representative or random. Specifically, children with poor reading comprehension may be over-represented compared to the wider population. Nonetheless, our sample represents a relatively wide range of ability levels, in both decoding and reading comprehension.

Secondly, assessments were always administered in the same order (NARA, TOWRE, CC2, YARC). It is possible that performance on the NARA may have been affected by anxiety or poor attention associated with adjusting to an unfamiliar testing situation. It is equally possible that fatigue or boredom could have affected YARC performance, as it was always the final assessment. Future research should counterbalance the order of assessment presentation in order to rule out these possibilities.

Thirdly, we focused on children in the upper primary grades, so we do not know how results might differ for the lower primary grades. We suggest that differences between the tests may be even larger, because for younger children, most of the variance in reading comprehension ability is explained by decoding ability (Elwer et al., 2013; Keenan et al 2008) – and as we have shown, NARA scores are more dependent on decoding ability. This would be an interesting focus for future research.

Fourthly, we were not able to explore the effects of question type and question difficulty in our sample, because children of different ages answered different questions, and therefore sample sizes differed widely from question to question. However, the type and difficulty of comprehension questions is likely to affect consistency between tests (Bowyer-Crane & Snowling, 2005). Future studies should explore these issues in relation to the NARA and the YARC.

Finally, it is important to note that our results reflect the Australian editions of the NARA and the YARC. Differences between the Australian and UK versions of the test are minimal (three small wording changes were made; Snowling et al., 2010), but the tests were standardised on different samples, so care should be taken when generalising these results beyond the Australian context.

Furthermore, our analyses do not take into account the role played by different standardization samples[4]. The third edition of the NARA was normed in 1997, while the Australian edition of the YARC was normed in 2011. It is possible, for example, that children in the NARA standardisation sample may (for whatever reason) have had superior comprehension skills to the children in the YARC sample. Therefore, an individual child's comprehension performance could appear comparatively worse on the NARA (though this would interact with other test design factors).

We cannot rule out the fact that differences between the standardisation samples may have contributed to differences in reading comprehension diagnosis, because our analyses were carried out on standard scores. However, our motivation for using these scores was the fact that in practice, they are used to make diagnoses in clinical or educational settings. Our results therefore represent the way that the tests are generally used. Nonetheless, it is important for future studies to tease apart any possible interactions between standardization sample and test format and content.

Despite these limitations, our data clearly show that there are substantial differences in diagnosis between reading comprehension tests, even when these tests are similar in format. A single comprehension assessment can never capture the complexity associated with reading comprehension ability. However, if researchers and clinicians are aware of the characteristics of different tests, they are better equipped to interpret assessments results, and make recommendations for intervention.

[4] We thank Dr Julia Carroll and an anonymous reviewer for drawing this to our attention.

References

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

GL Assessment. (2014). York Assessment of Reading for Comprehension: Improving literacy outcomes (City of Edinburgh Council). Retrieved 25 August 2014, from http://www.gl-assessment.co.uk/research-and-articles/york-assessment-reading-comprehension-improving-literacy-outcomes-city

Bowyer-Crane, C., & Snowling, M. J. (2005). Assessing children's inference generation: what do tests of reading comprehension measure? *British Journal of Educational Psychology, 75*, 189-201. doi:10.1348/000709904X22674

Britton, B. K., & Gulgoz, S. (1991). Using Kintsch's computational model to improve instructional text: Effects of repairing inference calls on recall and cognitive structures. *Journal of Educational Psychology, 83*, 329-345. doi:10.1037/0022-0663.83.3.329

Cain, K., & Oakhill, J. (1999). Inference making ability and its relation to comprehension failure in young children. *Reading and Writing, 11*, 489-503. doi:10.1023/A:1008084120205

Cain, K., & Oakhill, J. (2006a). Assessment matters: issues in the measurement of reading comprehension. *British Journal of Educational Psychology, 76*, 697-708. doi:10.1348/000709905X69807

Cain, K., & Oakhill, J. (2006b). Profiles of children with specific reading comprehension difficulties. *British Journal of Educational Psychology, 76*, 683-696. doi:10.1348/000709905X67610

Castles, A., Coltheart, M., Larsen, L., Jones, P., Saunders, S., & McArthur, G. (2009). Assessing the basic components of reading: A revision of the Castles and Coltheart

test with new norms. *Australian Journal of Learning Difficulties, 14*, 67-88. doi:10.1080/19404150902783435

Catts, H. W., Adlof, S. M., & Weismer, S. E. (2006). Language deficits in poor comprehenders: A case for the Simple View of Reading. *Journal of Language, Speech and Hearing Research, 49*, 278-293. doi:10.1044/1092-4388(2006/023)

Clarke, P. J., Henderson, L. M., & Truelove, E. (2010). The poor comprehender profile: Understanding and supporting individuals who have difficulties extracting meaning from text. In J. Holmes (Ed.), *Advances in Child Development and Behaviour* (Vol. 39, pp. 79-129). New York, Boston: Academic Press.

Clarke, P. J., Snowling, M. J., Truelove, E., & Hulme, C. (2010). Ameliorating children's reading-comprehension difficulties: a randomized controlled trial. *Psychological Science, 21*, 1106-1116. doi:10.1177/0956797610375449

Compton, D. L., Appleton, A. C., & Hosp, M. K. (2004). Exploring the relationship between text-leveling systems and reading accuracy and fluency in second-grade students who are average and poor decoders. *Learning Disabilities Research and Practice, 19*, 176-184. doi:10.1111/j.1540-5826.2004.00102.x

Cutting, L. E., & Scarborough, H. S. (2006). Prediction of reading comprehension: Relative contributions of word recognition, language proficiency, and other cognitive skills can depend on how comprehension is measured. *Scientific Studies of Reading, 10*, 277-299. doi:10.1207/s1532799xssr1003_5

Duke, N. K. (2005). Comprehension of what for what: Comprehension as a nonunitary construct. In S. G. Paris and S. A. Stahl (Eds.), *Children's Reading Comprehension and Assessment* (pp. 93-104). Mahwah, NJ: Lawrence Erlbaum.

Elwer, S., Keenan, J. M., Olson, R. K., Byrne, B., & Samuelsson, S. (2013). Longitudinal stability and predictors of poor oral comprehenders and poor decoders. *Journal of Experimental Child Psychology, 115*, 497-516. doi:10.1016/j.jecp.2012.12.001

Field, A. (2009). *Discovering Statistics Using SPSS* (3rd ed.). London, UK: Sage.

Foster, H. (2007). *Single Word Reading Test 6-16.* London: GL Assessment.

Francis, D. J., Fletcher, J.M., Catts, H. W., & Tomblin, J. B. (2005). Dimensions affecting the assessment of reading comprehension. In S. G. Paris and S. A. Stahl (Eds.), *Children's Reading Comprehension and Assessment* (pp. 93-104). Mahwah, NJ: Lawrence Erlbaum.

Freebody, P., & Anderson, R. C. (1983). Effects of vocabulary difficulty, text cohesion and schema availability on reading comprehension. *Reading Research Quarterly, 18*, 277-294. doi:10.2307/747389

Garcia, J. R., & Cain, K. (2014). Decoding and reading comprehension: A meta-analysis to identify which reader and assessment characteristics influence the strength of the relationship in English. *Review of Educational Research, 84*, 74-111. doi:10.3102/0034654313499616

Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education, 7*, 6-10. doi:10.1177/074193258600700104

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher, 40*, 223-234. doi:10.3102/0013189x11413260

Hoover, W. A., & Gough, P. B. (1990). The Simple View of Reading. *Reading and Writing, 2*, 127-160. doi:10.1007/BF00401799

Howe, W. (2013). The use of assessments in the inclusive classroom. Paper presented at the Pedagogy in Practice Conference, Brisbane, Australia.

Hulme, C., & Snowling, M. J. (2011). Children's Reading Comprehension Difficulties: Nature, Causes, and Treatments. *Current Directions in Psychological Science, 20,* 139-142. doi:10.1177/0963721411408673

Hulme, C., Stothard, S. E., Clarke, P., Bowyer-Crane, C., Harrington, A., Truelove, E., & Snowling, M. J. (2009). *York Assessment of Reading for Comprehension (YARC): Early Reading*. London, UK: GL Assessment.

Keenan, J. M., Betjemann, R. S., & Olson, R. K. (2008). Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading, 12*, 281-300. doi:10.1080/10888430802132279

Keenan, J. M., Hua, A. N., Meenan, C. E., Pennington, B. F., Willcut, E., & Olson, R. K. (2014). Issues in identifying poor comprehenders. *L'Annee Psychologique, 114,* 753-777. doi: 10.4074/S0003503314004072

Keenan, J. M., & Meenan, C. E. (2014). Test differences in diagnosing reading comprehension deficits. *Journal of Learning Disabilities, 47*, 125-135. doi:10.1177/0022219412439326

Marinus, E., Kohnen, S., & McArthur, G. (2013). Australian comparison data for the Test of Word Reading Efficiency (TOWRE). *Australian Journal of Learning Difficulties, 18,* 199-212. doi:10.1080/19404158.2013.852981

McArthur, G., Castles, A., Kohnen, S., Larsen, L., Jones, K., Anandakumar, T., & Banales, E. (2015). Sight word and phonics training in children with dyslexia. *Journal of Learning Disabilities, 48,* 391-407. doi: 10.1177/0022219413504996

McLaughlin, M. J., Speirs, K. E., & Shenassa, E. D. (2012). Reading disability and adult attained education and income: Evidence from a 30-year longitudinal study of a

population-based sample. *Journal of Learning Disabilities, 47,* 374-386. doi:10.1177/0022219412458323

McNamara, D. S., Graesser, A. C., & Louwerse, M. M. (2012). Sources of text difficulty: Across the ages and genres. In J. P. Sabatini & E. Albo (Eds.), *Assessing reading in the 21st Century: Aligning and applying advances in the reading and measurement sciences.* Lanham, MD: R & L Education.

McNamara, D. S., Louwerse, M. M., Cai, Z., & Graesser, A. C. (2011). Coh-Metrix version 3.0 [Computer software]. Retrieved 20 December 2012, from http://cohmetrix.com

Nation, K., & Snowling, M. J. (1997). Assessing reading difficulties: the validity and utility of current measures of reading skill. *British Journal of Educational Psychology, 67,* 359-370. doi:10.1111/j.2044-8279.1997.tb01250.x

Neale, M. D. (1999). *The Neale Analysis of Reading Ability Third Edition*. Melbourne, Australia: ACER.

Ozuru, Y., Rowe, M., O'Reilly, T., & McNamara, D. S. (2008). Where's the difficulty in standardized reading tests: The passage or the question? *Behaviour Research Methods, 40*, 1001-1015. doi:10.3758/BRM.40.4.1001

Ricketts, J. (2014). *Reading: Research and assessment*. Retrieved 4 September 2014, from http://www.gl-assessment.co.uk/research-and-articles/reading-research-and-assessment-jessie-ricketts

Ricketts, J., Sperring, R., & Nation, K. (2014). Educational attainment in poor comprehenders. *Frontiers in Psychology, 5,* 445. doi:10.3389/fpsyg.2014.00445

Silverman, R. D., Speece, D. L., Harring, J. R., & Ritchey, K. D. (2013). Fluency has a role in the Simple View of Reading. *Scientific Studies of Reading, 17,* 108-133. doi: 10.1080/10888438.2011.618153

Snowling, M. J., Stothard, S. E., Clarke, P., Bowyer-Crane, C., Harrington, A., Truelove, E., . . . Hulme, C. (2010). *York Assessment of Reading for Comprehension (YARC) Passage Reading Secondary Version*. London, UK: GL Assessment.

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 Edition*. London, UK: GL Assessment.

Spooner, A. L. R., Baddeley, A. D., & Gathercole, S. E. (2004). Can reading accuracy and comprehension be separated in the Neale Analysis of Reading Ability? *British Journal of Educational Psychology, 74*, 187-204. doi:10.1348/00070990477383983

Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Mahwah, NJ: Lawrence Erlbaum.

Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (1999). *Test of Word Reading Efficiency*. Austin, TX: Pro-Ed.

Table 1

Participant Standard Score Means and Standard Deviations

	Standard Scores		
Measure	Mean	SD	Range
NARA Comprehension	88.75	9.96	69 - 113
YARC Comprehension	96.42	10.86	75 - 125
NARA Rate	101.18	11.7	71 - 130
YARC Rate	97.69	12.39	70 - 126
NARA Accuracy	101.23	11.44	73 - 129
YARC Accuracy	94.46	10.8	70 - 117
TOWRE Sight Word Efficiency	100.01	12.84	69 - 127
TOWRE Phonemic Decoding Efficiency	98.09	14	65 - 139
CC2 Irregular Words	99.76	11.62	69 - 124
CC2 Nonwords	92.98	12.1	58 - 126

Note. n = 95. NARA = Neale Analysis of Reading for Comprehension. YARC = York Assessment of Reading for Comprehension. TOWRE = Test of Word Reading Efficiency. CC2 = Castles and Coltheart Reading Test 2. The CC2 does not provide standard scores, therefore standard scores were converted from z-scores. Standard scores have a mean of 100 and a standard deviation of 15.

Table 2
Prevalence of Reading Profiles Using Different Measures of Comprehension and Decoding

	Percentage of children		
Diagnostic category	NARA	YARC	Both tests
Poor comprehender – Cut-off only			
Comprehension standard score <85 and:			
text reading accuracy standard score ≥ 85[a]	25	12	8
nonword reading accuracy standard score ≥ 85[b]	22	12	8
irregular word reading accuracy standard score ≥ 85 [c]	24	15	11
nonword reading fluency standard score ≥ 85 [d]	20	12	7
word reading fluency standard score ≥ 85[e]	24	13	8
Poor comprehender – Cut-off and discrepancy			
Comprehension standard score <85 and 15 points below reading score, and:			
text reading accuracy standard score ≥ 85[a]	18	4	3
nonword reading accuracy standard score ≥ 85[b]	14	4	3
irregular word reading accuracy standard score ≥ 85 [c]	17	9	6
nonword reading fluency standard score ≥ 85 [d]	13	7	4
word reading fluency standard score ≥ 85[e]	15	4	2
Generally poor reader			
Comprehension score <85 and:			
text reading accuracy standard score < 85[a]	8	6	4
nonword reading accuracy standard score < 85[b]	12	6	5
irregular word reading accuracy standard score < 85 [c]	9	3	3
nonword reading fluency standard score < 85 [d]	14	6	6
word reading fluency standard score < 85[e]	9	5	5
Poor decoder			
Comprehension score ≥ 85 and:			
text reading accuracy standard score < 85[a]	0	13	0
nonword reading accuracy standard score < 85[b]	15	20	14
irregular word reading accuracy standard score < 85 [c]	3	9	3
nonword reading fluency standard score < 85 [d]	5	13	5
word reading fluency standard score < 85[e]	5	9	5
Successful reader			
Comprehension score ≥ 85 and:			
text reading accuracy standard score ≥ 85[a]	66	69	54
nonword reading accuracy standard score ≥ 85[b]	52	62	48
irregular word reading accuracy standard score ≥ 85 [c]	63	73	59
nonword reading fluency standard score ≥ 85 [d]	61	69	57
word reading fluency standard score ≥ 85[e]	61	73	57

Note. n = 95. SD = standard deviation. NARA = Neale Analysis of Reading for Comprehension. YARC = York Assessment of Reading for Comprehension. a. NARA accuracy score when comprehension measured on the NARA, YARC accuracy score when comprehension measured on the YARC. b. Castles and Coltheart Reading Test 2 (CC2) Nonword Reading score. c. CC2 Irregular Word Reading score. d. Test of Word Reading Efficiency (TOWRE) Phonemic Decoding Efficiency score. e. TOWRE Sight Word Efficiency score.

Table 3

Consistency of Diagnosis Across Comprehension Assessments Using Different Measures of Comprehension and Decoding

	Consistency (percentage)	
Test Score	1 SD cut-off	1.5 SD cut-off
Text reading accuracy[a]	66	79
Nonword reading accuracy[b]	76	85
Irregular word reading accuracy[c]	76	85
Nonword reading fluency[d]	76	85
Word reading fluency[e]	76	85

Note. n = 95. SD = Standard deviation. NARA = Neale Analysis of Reading for Comprehension. YARC = York Assessment of Reading for Comprehension. a. NARA accuracy score when comprehension measured on the NARA, YARC accuracy score when comprehension measured on the YARC. b. Castles and Coltheart Reading Test 2 (CC2) Nonword Reading score. c. CC2 Irregular Word Reading score. d. Test of Word Reading Efficiency (TOWRE) Phonemic Decoding Efficiency score. e. TOWRE Sight Word Efficiency score.

Table 4

Pattern Matrix Showing Factor Loadings of NARA, YARC, CC2 and TOWRE scores

	Factor	
Test Score	Comprehension	Accuracy
NARA Comprehension	0.78*	0.16
YARC Comprehension	0.89*	-0.07
NARA Accuracy	0.17	0.82*
YARC Accuracy	0.18	0.79*
CC2 Nonwords	-0.08	0.89*
CC2 Irregular Words	0.11	0.68*
TOWRE Phonemic Decoding	-0.02	0.93*
TOWRE Sight Words	-0.15	0.77*

Note. NARA = Neale Analysis of Reading for Comprehension. YARC = York Assessment of Reading for Comprehension. TOWRE = Test of Word Reading Efficiency. CC2 = Castles and Coltheart Reading Test 2. * Significant at $\alpha = 0.01$

Table 5

Regression Analyses Predicting Reading Comprehension Score From Reading Composite Score

	NARA				YARC			
	B	SE *B*	β	R^2	*B*	SE *B*	β	R^2
Constant	7.23	4.25			27.00	5.49		
Reading Composite	0.41	0.09	0.45**	.21	0.33	0.11	0.30**	.09

Note. NARA = Neale Analysis of Reading for Comprehension. YARC = York Assessment of Reading for Comprehension. * $p < .05$ ** $p < .01$

Table 6

Number of Passages Read and Questions Attempted on the Neale Analysis of Reading

Variable	Mean	SD	Mode	Maximum	Minimum
Number of passages read					
Grade 3	3.58	1.04	3	5	2
Grade 4	4.38	0.87	5	5	2
Grades 5/6	3.71	0.70	4	4	1
Number of questions attempted					
Grade 3	27.85	8.91	40	40	16
Grade 4	35.06	7.00	40	40	16
Grades 5/6	29.60	5.75	32	32	8

Note. According to manual instructions, children aged 9 years of age may begin the Neale Analysis of Reading (NARA) on Passage 2, and children aged 10 and above may begin on Passage 3, provided they do not make more than 2 errors on their first passage. Children are awarded credit for preceding, unread passages. Our analyses do not take into account these unread passages, therefore older students sometimes read fewer passages and answered fewer questions than younger students.

Table 7

Passage Analyses

	Passages											
Measures	NARA						YARC					
Passage level	1	2	3	4	5	6	1	2	3	4	5	6
Sentences per passage	4	8	8	8	8	8	8	8	19	9	13	11
Words per passage	26	52	73	97	117	141	66	98	155	182	192	222
Mean syllables per word	1.2 (0.4)	1.4 (0.7)	1.4 (0.6)	1.6 (0.8)	1.7 (1.0)	1.8 (1.1)	1.3 (0.5)	1.2 (0.4)	1.3 (0.6)	1.4 (0.7)	1.4 (0.7)	1.6 (1.0)
Mean letters per word	3.4 (1.7)	4.2 (2.1)	4.5 (2.3)	5.4 (2.5)	5.1 (3.0)	5.7 (2.8)	4.0 (1.7)	3.9 (1.6)	4.3 (2.3)	4.7 (2.2)	4.6 (2.5)	4.9 (2.5)
Mean log frequency (all words)	3.0	3.1	2.9	2.5	2.7	2.7	3.0	2.9	3.1	2.7	2.8	2.9
Mean content word frequency	2.4	2.2	2.0	1.6	1.7	1.7	2.2	2.1	2.1	1.8	1.9	2.0
Flesch-Kincaid grade level	0.6	3.1	4.6	8.1	10.5	12.5	2.7	3.2	3.3	8.8	6.3	11.1
Mean words per sentence	6.5 (1.0)	6.5 (2.3)	9.1 (3.0)	12.1 (4.6)	14.6 (7.9)	17.6 (9.8)	8.3 (3.0)	12.3 (5.1)	8.2 (3.5)	20.2 (10.0)	14.8 (7.8)	20.2 (8.5)
Mean words before main verb	1.5	1.0	3.3	2.3	4.0	6.6	1.5	1.5	2.8	3.3	2.5	8.4
Mean modifiers per noun phrase	0.7	0.7	0.8	1.0	1.0	1.1	0.7	1.3	0.7	0.7	0.8	1.1
Referential cohesion percentile	36.6	4.9	9.5	7.9	2.2	0.9	10.2	21.2	34.5	37.1	29.1	18.1
Mean sentence given/new ratio	0.2 (0.2)	0.2 (0.2)	0.2 (0.1)	0.3 (0.2)	0.2 (0.1)	0.2 (0.1)	0.2 (0.2)	0.3 (0.1)	0.3 (0.1)	0.3 (0.2)	0.2 (0.1)	0.3 (0.1)

Note. NARA = Neale Analysis of Reading (Neale, 1999). YARC = York Assessment of Reading for Comprehension (YARC). Standard deviations are in parentheses, however note that for some measures, standard deviations were not computed by CohMetrix. A higher percentile of referential cohesion indicates a higher level of connection between words and ideas in the text (Graesser, McNamara and Kulikowich, 2011).