Published in final edited form as:
J Learn Disabil. 2014 ; 47(2): 125–135. doi:10.1177/0022219412439326.

Test Differences in Diagnosing Reading Comprehension Deficits

Janice M. Keenan, Ph.D.
Department of Psychology, University of Denver, 2155 S. Race, Denver, CO 80208
jkeenan@du.edu

Chelsea E. Meenan, M.A.
Department of Psychology, University of Denver, 2155 S. Race, Denver, CO 80208
chelsea.meenan@du.edu

Abstract

We examined the implications of test differences for defining and diagnosing comprehension deficits using reading comprehension tests. We had 995 children complete the Gray Oral Reading Test-3, the Qualitative Reading Inventory-3, the Woodcock-Johnson Passage Comprehension-3, and the Peabody Individual Achievement Test, and compared which children were identified by each test as being in the lowest 10%. Although a child who performs so poorly might be expected to do poorly on all tests, we found that the average overlap between tests in diagnosing comprehension difficulties was only 43%. Consistency in diagnosis was greater for younger children, when comprehension deficits are due to weaker decoding skills, than for older children. Inconsistencies between tests were just as evident when identifying the top performers. The different children identified as having a comprehension deficit by each test were compared on four profile variables - word decoding skill, IQ, ADHD symptoms, and working memory skill – to understand the nature of the different deficits assessed by each test. Theoretical and practical implications of these test differences in defining and diagnosing comprehension deficits are discussed.

A variety of reading comprehension tests are on the market for use in evaluating children for supplementary services and for research investigating individual differences in comprehension. Selecting from among these choices for a test to use in diagnosis or for inclusion in a research project is often a matter of convenience. Factors such as how long the test takes to administer, how easy it is to score, and whether the test is already available or needs to be purchased tend to dictate a user's choice. Selection on the basis of these factors is justified if all of the tests are basically equivalent measures of the construct of reading comprehension. However, in recent years, data have been accumulating showing that many reading comprehension tests are not interchangeable. Format differences between tests, which might have been thought to have consequences mainly for administering and scoring, have now been shown to create differences between the tests in the underlying comprehension skills that they assess (Cutting & Scarborough, 2006; Francis, Fletcher, Catts, & Tomblin, 2005; Keenan, Betjemann, & Olson, 2008; Nation & Snowling, 1997).

The present article explores the implications of test differences for defining and diagnosing comprehension deficits. Do they imply that a child diagnosed as having a comprehension deficit with one test might not be diagnosed as having a deficit if a different test is used? Or is it the case that when a child performs so poorly as to be in the very low end of the distribution, then that child will perform poorly on any test, and hence be consistently

Correspondence concerning this article should be sent to Janice M. Keenan, Department of Psychology, University of Denver, 2155 S. Race, Denver, Colorado 80208. jkeenan@du.edu..

diagnosed, regardless of the particular differences in tests? As we will show, the odds of being similarly diagnosed are rather low. So we then explore why. We consider the extent to which age affects the consistency – is there more consistency or less consistency across tests in diagnoses for younger versus older children? We also compare cognitive profiles of the different children identified as comprehension deficit on each test to understand what deficits are being assessed on one test that are not being assessed on the other.

Reading Comprehension Test Differences

Most reading comprehension tests were developed long before there were theoretical frameworks for comprehension processes (Pearson & Hamm, 2005). Consequently, test developers typically offer information regarding test format (e.g., passage length, question type), administration (e.g., amount of time required to administer the test), and measurement (e.g., reliability, characteristics of the populations used in norming the instrument), but not analyses of component skills underlying their tests. Analyses of component skills assessed by different tests have, however, been a focus of comprehension researchers in recent years, and these test comparisons have demonstrated important differences between the tests.

The first analysis was done by Nation and Snowling (1997) on two British tests of reading comprehension – the Neale Analysis of Reading Ability and the Suffolk Reading Scale. The same children were given both tests, as well as tests of listening comprehension and decoding skill. Nation and Snowling then used multiple regression to analyze the extent to which performance on the reading comprehension tests was accounted for by performance on the decoding and listening comprehension tasks. They found that although decoding skill explained significant variance on both reading comprehension tests, listening comprehension did not account for any additional variance on the Suffolk, only the Neale. Because the Suffolk involves sentence completion (referred to as a cloze test), the authors concluded that reading comprehension tests involving a cloze format are essentially assessing word decoding skill and not the comprehension skills associated with listening comprehension tasks. A related conclusion regarding cloze tests was later offered by Francis, et al. (2005) who found through latent trait modeling that there was a stronger relationship between decoding and reading comprehension when reading comprehension was assessed with a cloze test than with multiple-choice questions.

Cutting and Scarborough (2006) compared three tests commonly used in the US (the Wechsler Individual Achievement Test reading comprehension subtest, the Gates-MacGinitie Reading Test, and the Gray Oral Reading Test). Although they did not find the striking discrepancy between tests that Nation & Snowling (1997) found, they did find differences between tests in how much of their variance was accounted for by decoding. They also reported that only 25% of children in the sample were diagnosed as having a comprehension deficit by all three tests (Rimrodt, Lightman, Roberts, Denckla, & Cutting, 2005).

Keenan, et al. (2008) reported dramatic differences between four US reading comprehension tests in the degree to which performance is explained by word decoding skill versus listening comprehension skill, and showed that the differences are not just a function of using a cloze-test format. They compared the same children on the Woodcock Johnson Passage Comprehension– 3 (WJPC, Woodcock, McGrew & Mather, 2001), the Peabody Individual Achievement Test (PIAT, Dunn & Markwardt, 1970), the Gray Oral Reading Test– 3 (GORT, Wiederholt & Bryant, 1992) and the Qualitative Reading Inventory– 3 (QRI, Leslie & Caldwell, 2001). Word decoding accounted for far more variance than listening comprehension when the test was either the WJPC-3 or the PIAT, but the reverse pattern was found for the other tests. Although the WJPC-3 is a cloze test, the PIAT is not;

so Keenan et al. proposed that the extent to which individual differences in reading comprehension tests are largely accounted for by word decoding skill is not so much a function of test format as of passage length. Both the PIAT and WJPC-3 use sentence-length passages, whereas the QRI-3 and GORT-3 use longer passages. Keenan et al. contend that longer passages increase dependence on higher-level language skills involved in constructing mental models of situations that dynamically change across sentences of the passage. Also, with longer passages, a reader has more context that can be used to recover from decoding failures (e.g., using text that states “pulled a rabbit out of the hat” to determine that the word was *magician*, not *musician*), so decoding skill accounts for less variance.

It is well known that decoding skill accounts for a larger portion of variance in reading comprehension tests for younger than for older children (Curtis, 1980; Tunmer & Hoover, 1993). However, Keenan et al. (2008) discovered that the specific test used for assessing comprehension greatly influences how dramatic this developmental difference appears to be. The disparity in the degree to which word decoding explained reading comprehension for young vs. old children (or for low vs. high word-reading skill) was quite large for tests assessing comprehension with short texts (PIAT and WJPC), whereas it was less dramatic for the tests with longer passages (GORT and QRI). Thus, detection of developmental differences is influenced by test differences.

Current Study

Diagnoses of Comprehension Deficit

The present study extends the research reviewed above to examine how test differences impact diagnoses of reading comprehension disorders. Reading comprehension tests are used for diagnoses clinically to determine whether a child qualifies for services and in research to define samples with comprehension deficits so as to study issues such as the nature of comprehension deficits (Cain & Oakhill, 2007), their etiology (Betjemann, Keenan, Olson, & DeFries, 2011; Keenan, et al., 2011; Keenan, et al., 2006) and stability (Cain & Oakhill, 2007; Catts, Adlof, & Weismer, 2006; Nation, Cocksey, Taylor, & Bishop, 2010). The question we address is: Does it matter which test the clinician or the researcher uses when identifying performers in the low tail? How often would a child diagnosed with one test be diagnosed if a different test had been used?

We are able to address this question because, as part of an ongoing behavior genetic study of comprehension (Keenan, et al., 2006), we have tested more than 1000 children where each child is given the four different reading comprehension tests examined in Keenan et al. (2008). We can thus determine whether the same children perform poorly on each test, or whether there are differences between tests in who performs in the low tail.

It may be that when children perform so poorly as to be in the extreme low tail that differences between tests in their relative assessment of component skills hardly matter. In other words, test differences that manifest when assessing the full sampling distribution may be dwarfed by the severity of deficits that lead to performance in the low tail. After all, component comprehension skills are interrelated (Cain & Oakhill, 2007), so a severe deficit in one component could affect all related skills such that the child ends up performing poorly on all tests. However, differences between tests in assessment of component skills may result in different children being identified as in the low tail. If that is the case, then we need to know the extent to which that happens and understand how the low performers on each of the tests differ. Because our behavioral genetic study also includes assessments of each child on word decoding, working memory, ADHD symptomology, and IQ, we can

examine whether there are differences in these cognitive profiles between the children defined as low performers by the different tests.

Profile Characteristics of Poor Comprehenders

How might a child who performs poorly on one test end up not performing poorly on another? We know that performance on some reading comprehension tests, such as the PIAT and WJ Passage Comprehension, is mainly dependent on word decoding skill, especially for younger readers (Keenan, et al., 2008). Thus, we might expect children with poor word decoding skills to comprise the low tail of the distribution on these tests; however, when tests are less weighted toward assessing decoding, deficits in other cognitive skills, such as working memory or attention, may define those in the low tail. Thus, we examined children in the low tail of each reading comprehension test on four cognitive profile characteristics: word decoding ability, working memory capacity, ADHD symptomology, and IQ.

We examined word decoding ability because identifying words is so central to grasping the meaning of a text. The question we had was whether low decoding ability is associated with low performance on all reading comprehension tests, or whether there are differences between the tests in how much low decoding skill defines the poor performers.

We examined working memory skill because it is critical to constructing a coherent text representation and maintaining the gist of what is being read (Daneman & Carpenter, 1980), and deficits in it are associated with comprehension deficits (Swanson, Howard, & Sáez, 2007). Individual differences in working memory have been shown to account for significant variance in reading comprehension even after controlling for attention, basic decoding skills, reading fluency and vocabulary (Sesma, Mahone, Levine, Eason and Cutting, 2009). But are there differences across tests in how much working memory limitations contribute to poor comprehension? It may be that tests using longer passages tax working memory more. Likewise, some administrative features of tests might put more load on working memory; e.g., in the PIAT, even though the text is only a single sentence, it needs to be held in working memory as the child looks at all the pictures to choose which picture best represents the meaning of the sentence.

We examined attention because some studies indicate that individuals with ADHD have a reading comprehension deficit (Brock & Knapp, 1996; Gregg et al., 2002; Keenan, Betjemann & Miller, 2008; Samuelsson, Lundberg, Herkner, 2004), although there are other studies that suggest they do not (Ghelani, Sidhu, Jain, & Tannock, 2004). This mixed picture could be due to the different reading comprehension tests used. If attention demands vary across tests, then those with ADHD symptoms may be more likely to be in the low tail of some tests than others.

Full-Scale IQ assesses a number of skills related to comprehension processes. We therefore expected that children in the low tail of reading comprehension tests would be at the lower end on IQ as well. The main question, however, is whether there are differences between the tests in how much low IQ defines poor performers.

Method

Participants

The sample consisted of 995 children : 888 twins[1] (308 Mz, 580 Dz) and 107 of their siblings recruited for a behavioral genetic study of comprehension skills (Keenan et al., 2006) as part of the Colorado Learning Disabilities Research Center (Olson, 2006). The median age was 11.17 years (range 8–18, with one who just turned 19). All were native

English speakers. Ethnicity of the sample was: 89% Caucasian, 4.3% Hispanic, 2.2% American Indian, 1.4% Asian, 1.3% African American, and 3% not reported or reported as other. The representativeness of our sample can be further gleaned from Table 1, which presents the standard scores for all those measures we used in the study that are nationally normed, and shows overall performance slightly above average.

Reading Comprehension Tests

Test Versions—The individualized testing of our large sample took place over many years. During that time, newer versions of some of our tests came on the market. However, to maintain continuity with earlier data collection on the project, we needed to continue to use the earlier versions. Thus, we used the PIAT, but it should be noted that it is identical in format to the PIAT-R. Similarly, we used GORT-3, but the passages and test items are identical to GORT-4. What typically varies across versions are only the norming data used to compute standard scores. Our test comparisons do not use standard scores, but rather are based on our standardizing the raw scores for each test across our large sample. Thus, our findings are relevant not only to the versions of the tests that we used, but to later versions as well.

Gray Oral Reading Test-3 (GORT)—The GORT-**3** (Wiederholt & Bryant, 1992) asks children to read aloud expository and narrative passages that range in length from 85–150 words. After the child reads each passage, the examiner then reads the multiple-choice comprehension questions to the child. Thus, because the examiner reads the questions, the child's decoding skill does not affect his/her ability to understand the questions.

Qualitative Reading Inventory-3 (QRI)—For the QRI-**3** (Leslie & Caldwell, 2001), children read aloud grade-level narrative and expository passages ranging in length from 250–785 words. Comprehension is assessed both by retelling the passage and answering open-ended comprehension questions. Retellings are scored based on the number of idea units recalled. Comprehension questions are scored as either correct or incorrect based on a scoring template. A composite of retell and comprehension question scores was created to use in the current analyses.

Peabody Individual Achievement Test (PIAT)—In the PIAT (Dunn & Markwardt, 1970) a single sentence is read silently, the sentence is removed, and a set of four pictures is then presented from which the child selects the one that best expresses the meaning of the sentence.

Woodcock Johnson Passage Comprehension-3 (WJPC)—In the WJPC-3 (Woodcock, McGrew & Mather, 2001), children silently read short passages of one or two sentences and provide a missing word (cloze format) to demonstrate their comprehension.

Cognitive Profile Measures

Word Decoding—Word decoding ability was derived using a composite of *z*-scores for the Timed Oral Reading of Single Words (Olson, Forsberg, Wise, & Rack, 1994) and the PIAT word recognition subtest (Dunn & Markwardt, 1970).

IQ—IQ was measured by the Wechsler Intelligence Scale for Children-Revised (WISCR; Wechsler, 1974) or Wechsler Adult Intelligence Scale-Revised (WAIS- R; Wechsler, 1981).

[1]It should be noted that if dependence within twin pairs has any effect, it would bias the results in favor of consistency in diagnosis across tests, not inconsistency. Analyses were performed on the full sample using both members of a twin pair to take advantage of the large sample size.

ADHD symptom count—The Attention Deficit Hyperactivity Disorder Rating Scale-IV (ADHDRS-IV, DuPaul et al., 1998) was used to assess number of symptoms associated with ADHD. Ratings were obtained from both parents and teachers on 18 symptoms. Parent and teacher ratings were combined by positively coding each symptom if it was endorsed by either the parent or the teacher (Lahey et al., 1994; Willcutt, et al., 2005).

Working memory—Working memory was measured using a composite of sentence span (Daneman & Carpenter, 1980), counting span (Case, Kurland & Goldberg, 1982), and digit span using forward and backward digit span from the WISC-R or the WAIS-R.

Procedure

Standardizing Test Scores—Although most of our reading comprehension tests provide standard scores, because the QRI-3 does not, we standardized the raw scores on each test on our sample of 995 children. Because the QRI-3 provides measures of both retellings and comprehension questions, we first standardized each of those, then formed a composite by averaging those standard scores, and then standardized the composite. For all tests, *z*-scores were computed using the standardized residuals that were saved after regressing on children's age and age squared, which were then standardized against the present sample.

Defining Reading Comprehension Deficit—In order to compare diagnoses across tests, a diagnosis of a reading comprehension deficit was defined as scoring in the low tail of the test's score distribution. We used a cutoff of the lowest 100 scorers (lowest 10% of the sample) so that we were comparing the same sample size across all tests.

Results

Correlations

The current study uses the same tests as Keenan, et al. (2008). When they reported that these reading comprehension tests are only modestly correlated (column 1, Table 2), the sample consisted of 512 children; since then it has grown considerably to 995. We therefore first checked to see how the increase in sample size might have affected the intercorrelations among the tests. The correlations from the current sample (column 2, Table 2) are quite similar to those reported by Keenan et al., although those involving the QRI-3 have increased slightly, probably reflecting improved standardization with a bigger sample (recall it was the only nonstandardized test). Most importantly, the correlations between the tests remain modest.

The focus of the present study is on using the tests to diagnose children with comprehension deficits (CD). Thus, our next step was to compare tests on their similarity in assigning a diagnosis of CD. With those in the low tail (10^{th} percentile) representing a positive diagnosis and the others representing a negative diagnosis of CD, we computed dichotomous correlations between the tests. These are shown in column 3 of Table 2, and they are generally low, averaging only .37. The dichotomous correlations are smaller than the continuous correlations between the tests because the variance of continuous data is constrained. Importantly, the pattern is similar across the two types of correlations for comparisons involving the PIAT, QRI-3, and WJPC-3; however, the dichotomous correlations for the GORT-3 are notably lower, suggesting that the GORT differs the most in assigning a CD diagnosis.

Overlap in Diagnoses between Tests

Another way of assessing similarity of diagnosis across tests is to compare the overlap in the specific children identified as being the poorest performers on each test. The lowest 100

scorers on each test (approximately 10% of the sample) were identified in order to determine the degree of overlap across measures in the likelihood that a child diagnosed as having a CD by one test would also be diagnosed by another test. Only 20 children were found to be consistently identified by all four tests. The fourth column of Table 2 displays the number of children that overlap in a CD diagnosis between each test pair. The overlap statistics follow the same pattern as the dichotomous correlations, with relatively low consistency between tests in whom they diagnose as having a CD. The consistency between tests in diagnosis, measured as percent overlap, ranged from a low of 35% between the PIAT and GORT-3 to a high of 56% between the PIAT and the WJPC-3. That is, a child who is among the lowest 100 scorers on the PIAT has only a 35% chance of being found in the low 100 scorers of the GORT-3, but a 56% chance of being in the low scorers of the WJPC-3. The average across all pairwise test comparisons was only 43%, meaning that the odds are less than half that a child diagnosed with a reading comprehension deficit with one test would get that same diagnosis if a different test had been used.

Reliability

One factor that could influence the amount of overlap across tests is the reliability of each test. The published reliability statistics for each test are presented in Table 3. Unfortunately, the procedures that publishers use to calculate reliability vary considerably, so it is probably not appropriate to compare these values. If we are to compare tests, we need estimates of reliability that have been obtained using the same method for each test. Fortunately, we could exploit the twin feature of our sample to obtain such an estimate. Specifically, we computed correlations between the monozygotic (MZ) twins to have comparable, low-bound estimates of reliability across the tests. The correlations between MZ twins can be considered an estimate of test-retest reliability because MZ twins share both their genes and family environment. They are a conservative estimate because even though MZ twins share genes and family environment, nonshared environmental influences and measurement error may reduce the correlation (DeFries, Vandenberg, & McClearn, 1976). But nonshared environmental influences on reading comprehension tend to be small (Keenan, et al., 2006), so the MZ correlations can be considered appropriate estimates of test-retest reliability. They are shown in the second column of Table 2. It is evident that they are considerably lower than the reliabilities reported by the publishers in their test manuals. Because the MZ correlations provide comparable estimates of reliability across all our tests, they suggest that one reason why the GORT-3 may have the lowest overlap with the other tests is because it has the lowest reliability.

Diagnosis Consistency as a Function of Age

Because Keenan et al. (2008) found a significant interaction between age and decoding skill in accounting for individual differences on some tests (PIAT and WJPC-3) but not on others (GORT-3 and QRI-3), we thought it was important to examine whether age affects the consistency of diagnosis across tests. Using a median split, we divided the sample into a younger group (M=9.32, SD=.83) and an older group (M=13.78, SD=1.87). We compared the overlap between tests for these two groups in two ways. As we did above for the full sample, we compared tests on who they defined as the lowest 100; however, because each age group is half of the full sample, the lowest 100 in each of the age groups now represents not the lowest 10% but rather the lowest 20% of scorers. In order to compare the age group results with the full sample results and have the same percentage defining comprehension deficit, we also compared tests on the lowest 10% for each age group, which consisted of the lowest 50 children on each test.

The results are shown in Table 4. The first column repeats the findings from the full sample so that they can be easily compared to the age group findings. The second and third columns

show the overlap among the lowest 100 scorers for the young and older groups respectively. There is considerable similarity in diagnosis consistency between what we found for the full sample and what we observe for the age group samples. Most notable is that the PIAT and the WJPC-3 tests are the most similar in their diagnoses, regardless of whether one examines the full sample, the younger sample or the older sample. The most discrepant diagnoses are obtained between the GORT-3 and the other tests for both the full and the younger sample, but interestingly, those test differences diminish for the older sample.

When we confine the age groups to the lowest 10% of scorers, shown in the fourth and fifth columns, the results are similar to what we observed for the lowest 100. The PIAT and WJPC-3 are the most similar in their diagnoses. The GORT-3 is the most discrepant from the other tests for the younger group, but not the older group. As we will show, these results are reflecting the relative importance of decoding for each test.

Comparing the last two columns allows us to see the same pattern across tests as reported by Keenan et al. (2008) when they noted a developmental interaction with type of reading comprehension test. For the first three comparisons involving the PIAT and the WJPC-3 with each other and with the QRI-3, we find that there is more consistency in diagnoses across tests for the younger 10% than for the older 10%. The overlap of the QRI-3 and WJPC-3 drops from 52 for the younger group to 38 in the older group; similarly, PIAT – WJPC-3 drops from 64 to 50 and PIAT – QRI-3 drops from 52 to 34. In contrast, the next three comparisons involving the GORT-3 with the other tests show similar amounts of test overlap regardless of age group; for the comparison of the younger to the older group, the percent overlaps are 36 vs. 32, 40 vs. 36, 40 vs. 40. Thus, the most similarity across tests in diagnoses of reading comprehension deficits occurs when children are younger and variance on the test is explained more by decoding skill.

Test Overlap in the High Tails

To determine if the inconsistency in diagnoses of comprehension deficits that we observed across reading comprehension tests was unique to the low tail of the distribution, we assessed the amount of overlap for the highest 10 percent of scorers for each test (those in the 90^{th} percentile and above). Those identified by each test as being in the high end of the distribution also showed inconsistency across tests, demonstrating that inconsistency across tests is not limited to the low end of the distribution. In fact, there was even less consistency across tests in identification of the top performers (M = 33%) than in identification of the poorest performers (M = 44%; $t(10) = 4.37$, $p < .05$), with overlaps ranging from 26% to 42%. As in the age group analyses, where we found less consistency between tests for older children, it appears that when decoding skills are high (top performers and older children) and account for less of the variance in reading comprehension performance, then there are greater differences between tests in identification of both who has a comprehension deficit and who is a top performer.

Profile Characteristics as a Function of the Reading Comprehension Test

Table 5 displays correlations between each reading comprehension test and each of the four profile characteristics – word decoding, ADHD symptom count, IQ, and working memory. All of the profile variables are significantly correlated with each of the tests, and all correlations are positive except those involving ADHD, where higher numbers of symptoms are associated with lower performance on each test. The strongest correlations are between IQ and each of the tests and between word decoding skill and the WJPC-3 and the PIAT tests.

To determine if there are profile differences between those identified as CD (lowest 10 percent) by each test, we performed a series of analyses of variance, one for each profile variable, in which the independent variable was type of reading comprehension test, which had four levels, and the dependent variable was the profile variable. The partial overlap in diagnoses between tests meant that we did not have a true between-subjects design because the same person sometimes appeared in the low tail of more than one test; but because the overlap was only partial, it meant that we also did not have a within-subjects design. Our options were to: 1) treat the design as if it were truly between-subjects, which overestimates the expected variance between groups and thus is a conservative test, or 2) exclude those participants that appeared in the low tail of more than one group in order to make it a true between-subjects design, although this means loss of power due to eliminating those participants with overlapping diagnoses. We did both analyses and found a similar pattern of results for each of the four profile variables whether including or excluding overlap cases. Results from both analyses are reported below.

Table 6 presents the average *z*-score on each of the profile variables of those uniquely identified as CD by each test and the number of unique cases. As we reported above, IQ was the most highly correlated profile variable with the tests, but as Table 5 shows there were no significant differences between tests in the IQs of those diagnosed as having a CD (including overlapping cases ($F(3, 396) < 1$; excluding overlaps $F(3, 130) = 1.30$). There were also no significant differences between tests in the ADHD symptom count of those in the low tail (including overlapping cases $F(3, 395) = 1.92$; excluding overlaps ($F(3, 129) < 1$). However, there were significant differences between tests in whether poor decoding and poor working memory defined the lowest performers. For decoding skill, the significant differences that occurred across tests (including overlapping cases $F(3, 396) = 11.98$, $p<.01$; excluding overlaps $F(3, 130) = 8.4$, $p < .01$) stemmed from the poorest performers on the WJPC-3 and the PIAT having significantly lower word decoding than poor performers on the QRI-3 and GORT-3 (WJ*vs*QRI: $t(53.85)=3.84$, $p<.01$; WJ*vs*GORT: $t(65.25)=3.42$, $p<.01$; PIAT*vs*QRI: $t(63)=4.14$, $p<.01$); PIAT*vs*GORT: $t(75)=3.59$, $p<.01$). For working memory, the significant differences between tests (including overlapping cases $F(3,396) = 3.06$, $p= .03$; excluding overlaps $F(3,130) = 3.48$, $p = .02$) were largely due to the PIAT, where children need to remember the sentence they just read as they examine a set of pictures to select the best depiction of the sentence's meaning. Post-hoc Tukey HSD tests showed that average working memory was significantly lower for those in the low tail of the PIAT than those in the QRI-3 ($t(63)=3.19$, $p<.01$).

Discussion

The present research asked whether a child diagnosed as having a comprehension deficit with one test would be similarly diagnosed if a different test were used. Recent research has shown that format differences between tests affect not just administration and scoring, but also the specific comprehension skills assessed (Cutting & Scarborough, 2006; Francis, et al., 2005; Keenan, et al., 2008; Nation & Snowling, 1997). Our work thus extends this research to its implications for defining and diagnosing comprehension deficits.

We had thought it possible that when a child performs so poorly as to be in the very low end of the distribution, then that child might perform poorly on any test, and hence be consistently diagnosed, regardless of the particular differences between the tests. However, our results showed that was not the case; there was considerable inconsistency across tests in diagnoses (defined as being in the lowest 10%). The average correlation between the four reading comprehension tests in assigning a diagnosis was only .37. Similar inconsistency between tests was observed in whom they identified as the top scoring 10%, showing that inconsistency in diagnosis is not to due to floor effects on performance.

We found that when children are younger there is greater similarity across tests in diagnoses of comprehension deficits than when they are older. The one exception was the GORT-3. Its consistency did not vary with age. It was also the least consistent with the other tests in the full sample analyses; an inconsistency that was then shown to be evident only when the children were younger. The difference between the GORT-3 and the other tests is that children with poor decoding skills can do well on the comprehension questions because they do not have to read them (the examiner reads them to the child) and because research has shown that a majority of these multiple-choice questions can be answered correctly even without reading the passage (Keenan & Betjemann, 2006). The other tests, however, require decoding and more of their variance is explained by decoding skill when children are younger (Keenan, et al., 2008). Thus, to the extent these tests agree in diagnosing comprehension deficits in young children, it is because they are identifying children whose poor comprehension is largely due to poor decoding.

In sum, we have gained an important insight into when reading comprehension tests will yield similar diagnoses. Consistency will be observed when performance is mostly explained by decoding skill, either because of the particulars of the test or because of the skill level of the children. When decoding skills are high, however, as they typically are for older children, or as they were for those scoring in the top 10% of the distribution, then decoding accounts for less of the variance in reading comprehension performance and there are greater differences between tests, both in their identification of who has a comprehension deficit and who is a top performer.

How do we explain the occurrence of the inconsistencies in diagnosis? One reason is the reliabilities of the tests. The GORT-3 had the lowest reliability and the lowest consistency with the other tests. But another reason is the differences between tests in the skills that they assess. Our assessment of profile differences between the children that were identified as low performers across the tests showed that tests differ in the relative importance of some, but not all, component comprehension skills. Although IQ was highly correlated with all of the tests, tests did not differ in how important IQ was to identifying poor performers. Nor did they differ in how important symptoms associated with ADHD were to identifying poor performers. The two skills that did differ between tests in who was identified as poor performers were word decoding skill and working memory.

That poor performers would differ across tests in decoding skill was expected given previous research showing that decoding explains more variance on the PIAT and WJPC-3 than on the GORT-3 and the QRI-3 when individual differences are examined across the full range (Keenan et al., 2008). As noted earlier, decoding skill explains so little variance on the GORT-3 because one does not need to read the passage in order to answer the questions (Keenan & Betjemann, 2006), and it is less important on the QRI-3 because its longer passages provide lots of context to compensate weak decoding.

What was less expected were the findings on working memory differences among the different poor performers identified by each test. We found that working memory was more important on tests employing short texts (PIAT and WJPC-3) than those with longer texts. Although this seems counterintuitive, it may well be explained by the memory load that is created by format features of the tests. The PIAT requires that the sentence be held in working memory as the child looks at four pictures to choose which best represents the meaning of the sentence; the WJPC-3 is a cloze test where the missing word is sometimes at the beginning of a 2-sentence passage, requiring the child to hold all the words in memory while reading the rest of the passage and considering various word choices. Further research will be needed to determine what aspects of test format and passage characteristics control

the working memory load of tests. But our findings provide an intriguing rationale for such research.

Conclusion

When people learn that we have compared the same child on different reading comprehension tests, they typically ask: “So, which is the best test?” Our results show that there is not a single best test to rely on as a measure of reading comprehension. Different tests yield different pictures of how proficient a student is on the various skills that comprise comprehension. Thus, we cannot recommend a specific test as best; rather, our results suggest that people use more than one test to achieve a broad assessment of the component skills of reading comprehension. Our results also suggest that the field might consider taking a broader, more multi-faceted approach to assessing comprehension, one that goes beyond reading comprehension tests alone and directly assesses component skills such as listening comprehension, vocabulary, and working memory to identify the source of deficits.

The present findings on test differences in assessing comprehension difficulties have relevance for the revision of diagnoses of reading problems proposed in the working draft of the revised Diagnostic and Statistical Manual of Mental Disorders (DSM-5; http://www.dsm5.org/), which acknowledges that word reading deficits and comprehension deficits are partially separate disorders. The present study shows that diagnosing these comprehension deficits will not be a straightforward matter since the odds are less than half that any two tests would yield the same diagnosis. Although there is more diagnostic consistency when these tests are used with younger children, it is because they are identifying children whose poor comprehension is largely due to poor word decoding. People use reading comprehension tests to assess comprehension skills other than word decoding, because decoding can be assessed more simply by having the child just read word lists. If the focus of interest is on assessing comprehension skills other than word decoding, we think clinicians and researchers may want to consider using measures of listening comprehension so as to bypass word reading (Fletcher, Lyon, Fuchs, Barnes, 2007; Keenan, et al., 2010).

Our findings underscore the complexity of comprehension and how tests differ in their assessment of its component skills. But most importantly, they convey a new message: the inconsistency in diagnosis tells us is that there is some separability in these skills, and that very few children are poor in all these skills. It thus supports a multicomponent view of comprehension deficits (c.f., Cain & Oakhill, 2007), and it suggests that practitioners and researchers should take a more nuanced view of comprehension deficits and use multiple tests to establish the nature of a child's reading comprehension deficit.

Acknowledgments

This research was supported by a grant from NIH HD27802 to the Colorado Learning Disabilities Research Center, for which J. Keenan is a co-PI. We thank the other co-PIs (John DeFries, Richard Olson, Bruce Pennington, and Erik Willcutt) for sharing their data for the profile analysis. Portions of these data were presented at the Society for the Scientific Study of Reading, 2009 and 2011. We thank Amanda Miller and Sarah Priebe for assistance with the data, Richard Olson for comments on the manuscript, all the participants and their families, and all the testers and scorers.

References

Betjemann RS, Keenan JM, Olson RK, DeFries JR. Choice of reading comprehension test influences the outcomes of genetic analyses. Scientific Studies of Reading. 2011; 15:363–382. [PubMed: 21804757]

Brock SE, Knapp P. Reading comprehension abilities of children with attention-deficit/hyperactivity disorder. Journal of Attention Disorders. 1996; 1(3):173–185.

Cain, K.; Oakhill, J. Reading comprehension difficulties: Correlates, causes, and consequences. In: Cain, K.; Oakhill, J., editors. Children's comprehension problems in oral and written language. Guilford Press; New York: 2007. p. 41-75.

Case R, Kurland M, Goldberg J. Operational efficiency and the growth of short-term memory span. Journal of Experimental Child Psychology. 1982; 33:386–404.

Catts HW, Adlof SM, Weismer SE. Language deficits in poor comprehenders: A case for the simple view of reading. Journal of Speech, Language and Hearing Research. 2006; 49:278–293.

Curtis ME. Development of components of reading skill. Journal of Educational Psychology. 1980; 72(5):656–669.

Cutting L, Scarborough H. 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. 2006; 10(3):277–299.

Daneman M, Carpenter PA. Individual differences in working memory and reading. Journal of Verbal Learning & Verbal Behavior. 1980; 19(4):450–466.

DeFries JC, Vandenberg SG, McClearn GE. Genetics of specific cognitive abilities. Annual Review of Genetics. 1976; 10:179–207.

Dunn, LM.; Markwardt, FC. Examiner's manual: Peabody individual achievement test. American Guidance Service; Circle Pines, MN: 1970.

DuPaul GJ, Power TJ, McGoey KE, Ikeda MJ, Anastopoulos AD. Reliability and validity of parent and teacher ratings of attention-deficit/hyperactivity disorder symptoms. Journal of Psychoeducational Assessment. 1998; 16(1):55–68.

Fletcher, JM.; Lyon, GR.; Fuchs, LS.; Barnes, MA. Learning disabilities: From identification to intervention. Guilford Press; New York: 2007.

Francis, DJ.; Fletcher, JM.; Catts, HW.; Tomblin, J. Dimensions Affecting the Assessment of Reading Comprehension. In: Paris, SG.; Stahl, SA.; Paris, SG.; Stahl, SA., editors. Children's reading comprehension and assessment. Lawrence Erlbaum Associates Publishers; Mahwah, NJ US: 2005. p. 369-394.

Ghelani K, Sidhu R, Jain U, Tannock R. Reading Comprehension and Reading Related Abilities in Adolescents with Reading Disabilities and Attention-Deficit/Hyperactivity Disorder. Dyslexia: An International Journal of Research and Practice. 2004; 10(4):364–384.

Gregg N, Coleman C, Stennett RB, Davis M. Discourse complexity of college writers with and without disabilities: A multidimensional analysis. Journal of Learning Disabilities. 2002; 35(1): 23–38. 56. [PubMed: 15490898]

Keenan, JM.; Betjemann, RS.; Miller, AC. Reading and Listening Comprehension in Children with ADHD. Invited talk for the Symposium on Attention and Reading Skills presented at the Annual Meeting of the Society for the Scientific Study of Reading; Asheville, North Carolina. Jul. 2008

Keenan J, Betjemann R, Olson R. Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. Scientific Studies of Reading. 2008; 12(3):281–300.

Keenan J, Betjemann R, Wadsworth S, DeFries J, Olson R. Genetic and environmental influences on reading and listening comprehension. Journal of Research in Reading. 2006; 29(1):75–91.

Keenan, JM.; Priebe, SJ.; Miller, AC.; Meenan, C.; Hua, A.; Olson, RK. Speaking up for listening comprehension. Presented at the Annual Meeting of the Society for the Scientific Study of Reading; Berlin, Germany. Jul. 2010

Keenan, JM.; Olson, RK.; Byrne, B.; Samuelsson, S. Preschool Predictors of Grade 4 Reading Comprehension, Listening Comprehension, and Decoding. Invited talk for the Symposium on Continuities between Listening Comprehension and Reading Comprehension presented at the Annual Meeting of the Society for the Scientific Study of Reading; St. Petersburg, FL. Jul. 2011

Leslie, L.; Caldwell, J. Qualitative Reading Inventory-3. Addison Wesley Longman; New York: 2001.

Nation K, Cocksey J, Taylor J, Bishop D. A longitudinal investigation of early reading and language skills in children with poor reading comprehension. Journal of Child Psychology and Psychiatry. 2010; 51:1031–1039. [PubMed: 20456536]

Nation K, Snowling M. Assessing reading difficulties: The validity and utility of current measures of reading skill. British Journal of Educational Psychology. 1997; 67(3):359–370. [PubMed: 9376312]

Olson RK. Genes, environment, and dyslexia: The 2005 Norman Geschwind memorial lecture. Annals of Dyslexia. 2006; 56(2):205–238. [PubMed: 17849199]

Olson, R.; Forsberg, H.; Wise, B.; Rack, J. Measurement of word recognition, orthographic, and phonological skills. In: Lyon, GR., editor. Frames of reference for the assessment of learning disabilities: New views on measurement issues. 1994. p. 243-277.

Pearson, P.; Hamm, DN. The Assessment of reading comprehension: A review of practices-past, present, and future. In: Paris, SG.; Stahl, SA.; Paris, SG.; Stahl, SA., editors. Children's reading comprehension and assessment. Lawrence Erlbaum Associates Publishers; Mahwah, NJ US: 2005. p. 13-69.

Rimrodt, S.; Lightman, A.; Roberts, L.; Denckla, MB.; Cutting, LE. Are all tests of reading comprehension the same?. Poster presented at the annual meeting of the International Neuropsychological Society; St. Louis, MO. Feb. 2005

Samuelsson S, Lundberg I, Herkner B. ADHD and reading disability in male adults: Is there a connection? Journal of Learning Disabilities. 2004; 37(2):155–168. [PubMed: 15493237]

Sesma H, Mahone E, Levine T, Eason S, Cutting L. The contribution of executive skills to reading comprehension. Child Neuropsychology. 2009; 15(3):232–246. [PubMed: 18629674]

Swanson, H.; Howard, CB.; Sáez, L. Reading comprehension and working memory in children with learning disabilities in reading. In: Cain, K.; Oakhill, J., editors. Children's comprehension problems in oral and written language: A cognitive perspective. Guilford Press; New York, NY US: 2007. p. 157-185.

Tunmer, WE.; Hoover, WA. Language-related factors as sources of individual differences in the development of word recognition skills. In: Thompson, G.; Tunmer, WE.; Nicholson, T.; Thompson, G.; Tunmer, WE.; Nicholson, T., editors. Reading acquisition processes. Multilingual Matters; Clevedon England: 1993. p. 123-147.

Wechsler, D. Manual for the Wechsler intelligence scale for children. revised edition. The Psychological Corporation; San Antonio, TX: 1974.

Wechsler, D. Manual for the Wechsler adult intelligence scale. revised edition. The Psychological Corporation; San Antonio, TX: 1981.

Wiederholt, L.; Bryant, B. Examiner's manual: Gray Oral Reading Test-3. Pro-Ed; Austin, TX: 1992.

Willcutt E, Pennington B, Olson R, Chhabildas N, Hulslander J. Neuropsychological analyses of comorbidity between reading disability and attention deficit hyperactivity disorder: In search of the common deficit. Developmental Neuropsychology. 2005; 27(1):35–78. [PubMed: 15737942]

Woodcock, RW.; McGrew, KS.; Mather, N. Woodcock-Johnson III tests of Achievement. 2001.

Table 1

Standard Scores for Measures Used in the Study That Are Nationally Normed

Measure	Mean	Standard Deviation
Full-Scale IQ	109.05	13.17
Verbal IQ	110.61	14.36
Performance IQ	105.55	12.93
Reading Comprehension Tests		
GORT-3	11.20	3.14
PIAT	107.39	12.52
WJPC-3	102.60	10.31
Word Decoding		
PIAT Test of Word Reading	105.36	12.03

Table 2

Correlations between Pairs of Reading Comprehension Tests and Number of Overlapping Cases of Diagnoses of Comprehension Deficit (lowest 10 percent)

Test Pair	Bivariate Correlations Keenan, et al. (2008) (N=512)	Bivariate Correlations Current Sample (N=995)	Dichotomous Correlations Current Sample (Phi) (N=995)	Number of Overlapping Diagnoses In Low Tail (N=100)
QRI|WJPC	.47	.56	.44	50
PIAT|WJPC	.70	.68	.51	56
PIAT|QRI	.45	.50	.39	45
PIAT|GORT	.51	.54	.28	35
QRI|GORT	.35	.45	.29	36
WJPC|GORT	.54	.53	.32	39

Table 3

Monozygotic Correlations as Estimates of Test-Retest Reliability and Published Reliabilities for Each Reading Comprehension Test

Test	MZ Correlation (N=152 pairs)	Published Reliability
QRI-3	.54	.94 – .98
WJPC-3	.71	.92
PIAT	.77	.64
GORT-3	.44	.75

Table 4

Consistency of Comprehension Deficit Diagnosis (% of Children Diagnosed by Both Tests) for Full Sample, and for Younger and Older Age Groups

	Full Sample	Younger	Older	Younger	Older
Test Pair	**Overlap Low 100**	**Overlap Low 100 (20%)**	**Overlap Low 100 (20%)**	**Overlap Low 50 (10%)**	**Overlap Low 50 (10 %)**
QRI|WJPC	50	54	46	52	38
PIAT|WJPC	56	67	55	64	50
PIAT|QRI	45	51	47	52	34
PIAT|GORT	35	43	51	36	32
QRI|GORT	36	41	47	40	36
WJPC|GORT	39	49	44	40	40

J Learn Disabil. Author manuscript; available in PMC 2015 March 01.

Table 5

Correlations between Reading Comprehension Tests and the Profile Variables - ADHD Symptomology, Word Decoding Skill, IQ, and Working Memory Capacity

Test	ADHD Sx	Decoding	IQ	Working Memory
QRI-3	−.208	.442	.556	.333
WJPC-3	−.222	.694	.668	.462
PIAT	−.165	.735	.644	.482
GORT-3	−.134	.463	.559	.364

*All values are significant at $p < .001$.

Table 6

Mean z-score (standard deviation) for Each Profile Characteristic for Children Uniquely Identified in the 10th Percentile of Each Test

Test	N	ADHD	Decoding	IQ	Working Memory
QRI-3	34	.10 (1.22)	−.17 (.73)	−.44 (.79)	−.04 (.60)
WJPC-3	23	.34 (1.08)	−.87 (.57)	−.68 (.89)	−.32 (.50)
PIAT	31	.21 (.94)	−.99 (.80)	−.69 (.75)	−.48 (.55)
GORT-3	46	.06 (1.17)	−.22 (.98)	−.40 (.84)	−.15 (.70)