SCIENTIFIC STUDIES OF READING, *12*(3), 281–300
Copyright © 2008 Taylor & Francis Group, LLC
ISSN: 1088-8438 print / 1532-799X online
DOI: 10.1080/10888430802132279

Reading Comprehension Tests Vary in the Skills They Assess: Differential Dependence on Decoding and Oral Comprehension

Janice M. Keenan
Psychology Department
University of Denver

Rebecca S. Betjemann
Institute for Behavioral Genetics
University of Colorado

Richard K. Olson
Department of Psychology
University of Colorado

Comprehension tests are often used interchangeably, suggesting an implicit assumption that they are all measuring the same thing. We examine the validity of this assumption by comparing some of the most popular reading comprehension measures used in research and clinical practice in the United States: the Gray Oral Reading Test (GORT), the two assessments (retellings and comprehension questions) from the Qualitative Reading Inventory (QRI), the Woodcock–Johnson Passage Comprehension subtest (WJPC), and the Reading Comprehension test from the Peabody Individual Achievement Test (PIAT). Modest intercorrelations among the tests suggested that they were measuring different skills. Regression analyses showed that decoding, not listening comprehension, accounts for most of the variance in both the PIAT and the WJPC; the reverse holds for the GORT and both QRI measures. Large developmental differences in what the tests measure were found for the PIAT and the WJPC,

Correspondence should be sent to Janice M. Keenan, Department of Psychology, University of Denver, 2155 South Race, Denver, CO 80208. E-mail: jkeenan@du.edu

but not the other tests, both when development was measured by chronological age and by word reading ability. We discuss the serious implications for research and clinical practice of having different comprehension tests measure different skills and of having the same test assess different skills depending on developmental level.

Reading a passage involves decoding, comprehension, and the interaction between the two processes. Until recently, however, reading assessment and research on reading disability has focused mainly on word decoding skills. This emphasis was likely because decoding is primary to comprehension, and because decoding failures are more easily defined than comprehension failures, thus rendering understanding decoding a more tractable problem than understanding comprehension. But because comprehension is the ultimate goal of reading and because comprehension failures can lead to school failures, there has been an increased interest in trying to assess and understand comprehension (RAND Reading Study Group, 2002). Much of this work on comprehension has involved the use of standardized tests of reading comprehension. The questions we address in this article are: How comparable are these tests? Are they all measures of the same process?

It is both common and reasonable to assume that the many tests of reading comprehension available on the market are interchangeable because all are presented as measures of the same construct, *comprehension*. However, as Davis (1944) pointed out long ago, and as we have relearned from more recent research on discourse processing, comprehension is not a unitary construct. It consists of multiple cognitive processes. As such, it is possible that different tests might tap into these processes in different ways. This possibility makes it important to know how the tests compare to one another. Are they all measures of the same component skills of comprehension, or do some tests measure different skills than other tests?

Knowing what skills are measured by a specific comprehension test and how comparable the test is to other tests is important both for assessment and for research. In assessment, the amount of time for testing is limited, so typically only one test is used. But to understand a child's comprehension skills, it is important to know how representative that test is, and if a child who shows poor reading comprehension on that test would also perform poorly on a different test. Similarly, researchers who use one reading comprehension test need to know if they are measuring the same thing as other researchers who might have used a different test, so that they know how to ascribe differences between studies in their attempts to replicate each other's findings.

The only information that has been available in the past to allow researchers and clinicians to compare different reading comprehension tests refers not to the types of cognitive processes measured by the tests, but rather to more practical information pertaining to test format (e.g., passage length, question type), test 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). Information about the types of comprehension skills assessed is not offered perhaps because, as Pearson and Hamm (2005) noted, most comprehension tests were developed long before there were theoretical frameworks for understanding comprehension processes.

Although test developers have not offered analyses of the skills underlying their tests, reading researchers whose focus is on comprehension processes are beginning to take on the task and ask exactly what the tests they are using measure. Nation and Snowling (1997) were the first to raise the question. They compared two British tests of reading comprehension—the Neale Analysis of Reading Ability and the Suffolk Reading Scale—in terms of their covariance with measures of both decoding and listening comprehension. Performance on both tests was influenced by decoding skill. However, listening comprehension accounted only for additional variance on the Neale; it did not account for any additional variance on the Suffolk. Because assessment of comprehension on the Suffolk involves sentence completion (referred to as a cloze test), the authors concluded that a cloze-test format essentially measures word recognition skill. A related conclusion regarding cloze tests was recently offered by Francis, Fletcher, Catts, and Tomblin (2005) who found through latent trait modeling that there was a stronger relationship between decoding and comprehension when comprehension was assessed with a cloze test than with multiple-choice questions.

Cutting and Scarborough (2006) examined three tests commonly used in the United States (the Wechsler Individual Achievement Test reading comprehension subtest, the Gates–MacGinitie Reading Test, and the Gray Oral Reading Test) and did not find the striking discrepancy that Nation and Snowling (1997) found in how much variance in reading comprehension was accounted for by oral language. Without further research, it is not clear whether that is because of the specific reading comprehension tests that they analyzed or because of differences in the measures used to define oral language and listening skill. However, even though the patterns of variance accounted for by decoding and listening were similar across Cutting and Scarborough's three tests, they did report a startling inconsistency among the tests in identifying which children were disabled. Specifically, they stated that Rimrodt, Lightman, Roberts, Denckla, and Cutting (2005) used their three tests to classify children as having a comprehension deficit, and Rimrodt et al. reported that even though 43.5% of their sample was identified by at least one of the tests as having a reading comprehension deficit, only 9.4% of the sample was identified as having a reading comprehension deficit by all three tests.

These initial studies comparing reading comprehension tests thus present a mixed picture as to whether one can assume that various tests are comparable. We think it is important therefore to expand the comparison of tests on their component skills so as to provide researchers and clinicians further information about the comparability of tests. In addition, because there is the suggestion from Nation and Snowling (1997) and Francis et al. (2005) that tests using a cloze format may be more influenced by decoding skill than other comprehension tests, it is important to examine additional tests to try to provide further insight into this issue. The present study therefore not only examines different tests, but also expands the range of test formats examined so that we can begin to determine whether it is the cloze format or some other aspect of the test that determines whether it is more a measure of decoding than comprehension skill.

Another question that we address in this study is the extent to which there are developmental differences in what a test measures. Is it possible for the same test to be assessing different skills depending on the age or the decoding ability level of the reader? By examining test performance across different ages and different levels of reading skill, we are able to answer this question. In sum, the goal of this research is to determine if what we are calling reading comprehension varies with the specific test being used, and if what the specific test measures varies with developmental level. If so, then we must begin to face the problems both for research and clinical practice inherent in referring to all of them as measures of the same construct.

OUR TEST SELECTIONS

The reading comprehension tests we compare in this article are all used in our behavioral genetic study of reading comprehension (cf. Keenan, Betjemann, Wadsworth, DeFries, & Olson, 2006) conducted as part of a larger study of learning disabilities (DeFries et al., 1997; Olson, 2006). We have been using multiple measures of reading comprehension in the hope that across all these measures we are capturing most of the variance associated with individual differences in comprehension skill. Because there was no research available about the particular comprehension skills assessed by any of the tests when we were designing our test battery, we attempted to cover a range of comprehension skills by covering a range of test formats, reasoning that different formats involve different task demands and different task demands are likely to tap a broader range of skills.

There are many test formats used in constructing reading comprehension tests. Among them are (a) whether reading is oral or silent, (b) the length of the passage, and (c) the particular type of comprehension assessment. The reading comprehension tests in our battery are the Gray Oral Reading Test–3 (GORT; Wiederholt & Bryant, 1992), the Qualitative Reading Inventory–3 (QRI; Leslie & Caldwell, 2001), the Woodcock–Johnson Passage Comprehension subtest (WJPC) from the Woodcock–Johnson Tests of Achievement–III (Woodcock, McGrew, & Mather, 2001), and the Reading Comprehension subtest from the Peabody Individual Achievement Test (PIAT; Dunn & Markwardt, 1970), which

TABLE 1
How Our Reading Comprehension Tests Instantiated Various Test Format Options

	GORT	*QRI*	*PIAT*	*WJPC*
Reading				
Oral	X	X		
Silent			X	X
Text				
Single sentence			X	X
Short passage				X
Medium passage	X			
Long passage		X		
Assessment				
Picture selection			X	
Cloze				X
Multiple-choice	X			
Short answer		X		
Retell		X		

Note. GORT = Gray Oral Reading Test–3; QRI = Qualitative Reading Inventory–3; PIAT = Peabody Individual Achievement Test; WJPC = Woodcock–Johnson Passage Comprehension subtest.

is identical in format to the PIAT–R and PIAT–R/NU (Markwardt, 1989, 1997).[1] Table 1 shows how across our tests we covered the range of test format options. We have two tests that are silent reading and two that are oral reading. The passage length varies from a single sentence to long passages up to 785 words. The types of tasks used to assess comprehension include (a) picture selection in the PIAT, where the child must select from among four pictures the one that best represents the meaning of the sentence just read; (b) the cloze technique in the WJPC, wherein the child is presented with a text in which one word is omitted and the child demonstrates understanding by providing the missing word; (c) multiple-choice comprehension questions in the GORT; (d) open-ended, short-answer questions in the QRI, some of which are literal and some inferential; and finally, because the QRI involves two assessments of comprehension, (e) retelling the passage in the QRI.

METHOD

Participants

The sample consisted of 510 children. Because they were taking these tests as part of a behavioral genetic study of comprehension skills, all were twins (180 identi-

[1] We use the PIAT rather than the PIAT–R to maintain continuity with earlier data collection on the project, but its format is identical to the PIAT–R.

cal, 290 fraternal) and their siblings ($n = 40$). The children ranged in age from 8 to 18 years, with the median at 10.5. They were recruited from 27 different school districts in Colorado by first identifying twins from school birth-date records and then sending letters to the families requesting participation. For inclusion in our analyses, the participants had to have English as their first language, no uncorrected sensory deficits, and Full-Scale IQ greater than 85 as measured by the Wechsler Intelligence Scale for Children–Revised (Wechsler, 1974) or the Wechsler Adult Intelligence Scale–Revised (Wechsler, 1981).

One potential problem of using twins as participants is possible nonindependence of the data when related individuals constitute the sample. Although whatever potential bias created by using twins and siblings would be constant across the analyses of each test because the same participants took all the tests, we also took steps to ensure that the results were not specific to a twin sample. The analyses were redone twice, once using one of the randomly selected twins of each pair, and a second time using the other member; siblings were excluded from these analyses to maintain independence of observations. The results were identical to the original analyses that included all individuals, both in terms of significance and in terms of the obtained values (to the second decimal place in almost all cases). Reported here are the values from the full sample.

Measures

Reading comprehension. The reading comprehension tests were the WJPC (Woodcock et al., 2001), in which children silently read short passages and provide a missing word to demonstrate their comprehension; the QRI (Leslie & Caldwell, 2001), in which grade-level expository and narrative stories (250–785 words) are read aloud, and comprehension is assessed by (a) number of ideas recalled from an idea checklist provided for each passage in a passage retelling and (b) short-answer comprehension questions; the GORT (Wiederholt & Bryant, 1992), in which expository and narrative passages (80–150 words) are read aloud, and multiple-choice comprehension questions for each passage are read to the child by the examiner; and the PIAT of Reading Comprehension (Dunn & Markwardt, 1970), which has participants silently read a sentence and choose which of four pictures expresses the meaning of the sentence.

Listening comprehension. A composite measure of listening comprehension was based on the combined age-adjusted z scores for the following tests, the first two of which are comparable to their reading comprehension versions: Woodcock–Johnson Oral Comprehension subtest (Woodcock et al., 2001), in which children listen to short passages and provide a missing word to demonstrate their comprehension; QRI (Leslie & Caldwell, 2001), in which participants listen to passages, both narrative and expository, and then retell the passage and answer comprehension questions; and the KNOW-IT Test (Barnes & Dennis, 1996; Barnes, Dennis, & Haefele-Kalvaitis, 1996), in which children are first taught a knowledge base relevant to the story that they will hear, then listen to the long story, and then answer short-answer literal and inferential comprehension questions.

Word decoding. Word recognition of isolated words was assessed with a composite score computed from two tests. One was the PIAT Word Recognition subtest (Dunn & Markwardt, 1970). Participants read across rows of increasingly difficult unrelated words until they reach an error criterion. There is no time constraint. The other test was the Timed Oral Reading of Single Words (Olson, Forsberg, Wise, & Rack, 1994), which assessed word recognition accuracy for a series of increasingly difficult single words presented on the computer screen. For responses to be scored as correct, they had to be initiated within 2 sec. The test's age-adjusted correlation with PIAT word recognition is .88. The composite measure of word recognition used in our analyses was created by combining the age-adjusted z scores for the two measures.

Nonword decoding. The Nonword Reading task consisted of reading 45 one-syllable (e.g., *ter, strale*) and 40 two-syllable (e.g., *vogger, strempick*) nonwords aloud (Olson et al., 1994). Percent correct scores were calculated for each task. Test–retest reliability is .86.

RESULTS

Descriptive Statistics

The first column of Table 2 presents descriptive statistics for all of the standardized tests for our full sample. For each of the different tests, it is clear that the means and standard deviations for our sample are comparable to the population norms, or slightly above. In addition, distribution plots of each measure on our sample showed normal distributions. Table 2 also presents the means and standard deviations on each of the standardized tests broken into two subgroups: one defined by chronological age and one defined by reading ability, where reading ability was based on raw scores from the PIAT Word Recognition test. These subgroups were not used in the regression analyses reported in the results—all analyses were performed using age and reading ability as continuous variables. However, for purposes of illustrating our developmental findings later in Figure 2, we used median splits on age and reading age, and we present the means of these groups here for comparison. The QRI measures are not in this table because the QRI is not a stan-

TABLE 2
Standard Score Means and Standard Deviations for Each Standardized Measure

Standardized Measures	*Overall*	*Age Groups*		*Reading Age*	
		Older	*Younger*	*Low*	*High*
GORT-3 Comprehension	10.99 (3.08)	11.68 (3.24)	10.25 (2.71)	9.87 (2.54)	12.10 (3.17)
WJ Passage Comprehension	102.18 (10.41)	103.48 (9.96)	100.77 (10.7)	97.54 (9.30)	106.87 (9.36)
PIAT Comprehension	107.44 (12.6)	105.57 (12.10)	109.46 (12.90)	104.28 (13.15)	110.68 (11.22)
PIAT Word Recognition	105.50 (12.21)	104.67 (11.94)	106.39 (12.46)	99.56 (10.80)	111.47 (10.53)

Note. Standard deviations are in parentheses. GORT = Gray Oral Reading Test–3; WJ = Woodcock–Johnson; PIAT = Peabody Individual Achievement Test.

dardized test. Note, however, that we standardized each of the two QRI measures before conducting any of our analyses by taking the raw scores and standardizing them across the full sample, regressed on age and age squared; these standardized residuals were then used in all analyses.

Intercorrelations Among Reading Comprehension Tests

Table 3 presents the correlations among the reading comprehension tests. They range from a low of .31 for the correlation of the GORT with the QRI Retellings, to a high of .70 for the correlation between the PIAT and the WJPC. In general, except for the correlation between the PIAT and the WJPC, the correlations among the tests are rather modest given that they are all purporting to be measures of the same construct. Even the correlation between the two assessments of the QRI, which are both assessing comprehension of the same passage, is only .41. These

TABLE 3
Intercorrelations Among the Reading Comprehension Tests

	GORT	*QRI–Retell*	*QRI–Qs*	*PIAT*	*WJPC*
GORT	1.0				
QRI–Retell	.31	1.0			
QRI–Qs	.38	.41	1.0		
PIAT	.51	.45	.44	1.0	
WJPC	.54	.48	.45	.70	1.0

Note. GORT = Gray Oral Reading Test–3; QRI = Qualitative Reading Inventory–3; PIAT = Peabody Individual Achievement Test; WJPC = Woodcock–Johnson Passage Comprehension subtest.

modest correlations suggest that these reading comprehension assessments may be measuring different component skills of comprehension.

Factor Analysis

An exploratory principal components factor analysis using oblique rotation, to allow for correlations between the factors, was performed. The analysis included not only the five reading comprehension assessments but also the listening comprehension composite and the word and nonword decoding measures. Two factors emerged with eigenvalues greater than 1. We refer to these factors as Comprehension and Decoding, and the correlation between them was *r* = .52. Table 4 shows the pattern matrix of factor loadings. All the reading comprehension tests load on the comprehension factor, but the factor loadings for the PIAT and the WJPC are lower (.37, .43) than the other tests (.62 –.79). Furthermore, the PIAT and the WJPC, but not the other reading comprehension measures, also load highly on the decoding factor; in fact, they load considerably higher on decoding than on the comprehension factor.

Regression Analyses

A pair of hierarchical regressions was run for each of the five reading comprehension tests to determine how much of the variance in performance on each test was accounted for uniquely by word decoding and by listening comprehension and how much was shared variance. For each pair of regressions, word decoding was entered as the first step followed by the listening comprehension composite in one analysis (Model 1), whereas in the other analysis, the order of entering was re-

TABLE 4
Pattern Matrix Showing the Factor Loadings of the Reading Comprehension Tests (in Bold), the Word and Nonword Decoding Composites, and the Listening Comprehension Composite

	Comprehension Factor	*Decoding Factor*
Listening Comprehension Composite	.88	–.06
GORT–3 Comprehension	.69	.04
QRI Reading–Questions	.79	–.06
QRI Reading–Retell	.62	.11
PIAT	.37	.58
WJPC	.43	.54
Word Decoding Composite	.07	.90
Nonword Decoding	–.13	.96

Note. GORT–3 = Gray Oral Reading Test–3; QRI = Qualitative Reading Inventory–3; PIAT = Peabody Individual Achievement Test; WJPC = Woodcock–Johnson Passage Comprehension subtest.

versed, listening comprehension composite entered before the word decoding composite (Model 2). The results are shown in Table 5. As can be seen in this table, both listening comprehension and word decoding account for significant independent variance in each of the tests. However, what is also evident is that the tests vary dramatically in the amount of variance accounted for by decoding, as seen in the ΔR^2 column for Model 2, Step 2.

This discrepancy in the influence of word decoding skills on performance can most easily be seen in Figure 1. This figure shows the amount of variance in each of the five reading comprehension measures that is unique to word decoding, unique to listening comprehension, and shared between the two. Two findings are most salient from Figure 1. One is that more total variance is accounted for in the PIAT and the WJPC than the other three measures. The other finding is that this is because most of the variance in these two tests is accounted for by word decoding and its shared variance with listening comprehension. Only 5% of the variance on the PIAT and only 7% of the variance on the WJPC is accounted for independently by listening comprehension skills. Thus, the answer to our question of whether reading comprehension tests differ in the degree to which they are assessing component skills appears to be yes because the PIAT and WJPC are more sensitive to individual differences in decoding skill than are the GORT or the QRI measures.

This split in sensitivity to word decoding is similar to the difference found by Nation and Snowling (1997) between the Suffolk and the Neale reading tests. What is interesting about our finding, however, is that it is not just the test using the cloze format (the WJPC) that is so heavily influenced by decoding skill; the PIAT, which uses multiple-choice selection of pictures representing the meaning of the

TABLE 5
Regression Analyses from the Full Sample Predicting Comprehension from Word Decoding and Listening Comprehension on Each of the Five Reading Comprehension Assessments

	GORT		*QRI–Retell*		*QRI–Qs*		*PIAT*		*WJPC*	
	R^2	ΔR^2	R^2	ΔR^2	R^2	ΔR^2	R^2	ΔR^2	R^2	ΔR^2
Model 1										
1. Decoding	.197*		.165*		.150*		.540*		.542*	
2. Listening Comp	.293*	.096*	.305*	.141*	.321*	.171*	.587*	.047*	.611*	.069*
Model 2										
1. Listening Comp	.218*		.259*		.287*		.246*		.291*	
2. Decoding	.293*	.075*	.305*	.047*	.321*	.033*	.587*	.341*	.611*	.319*

Note. Comp = comprehension. GORT = Gray Oral Reading Test–3; QRI = Qualitative Reading Inventory–3; PIAT = Peabody Individual Achievement Test; WJPC = Woodcock–Johnson Passage Comprehension subtest. *$p < .01$.

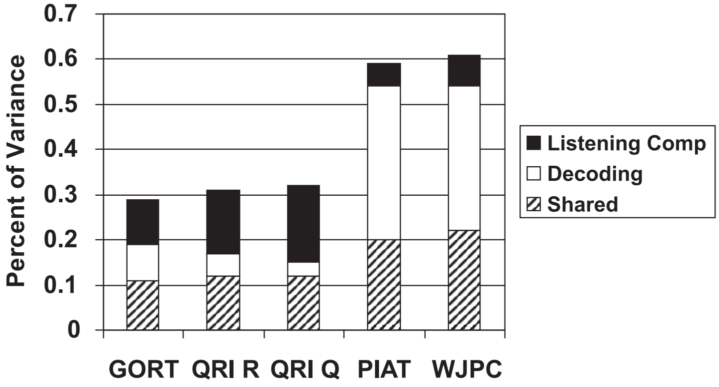

FIGURE 1 The proportion of total variance in each of the reading comprehension tests that was accounted for independently by word decoding skill, by listening comprehension skill, or shared. *Note.* GORT = Gray Oral Reading Test–3; QRI R = Qualitative Reading Inventory–3 Retellings; QRI Q = Qualitative Reading Inventory–3 Comprehension Questions; PIAT = Peabody Individual Achievement Test; WJPC = Woodcock–Johnson Passage Comprehension subtest.

sentence, shows the same pattern as the WJPC's cloze-test format shows. This indicates that other factors besides the format of the test item are responsible for this pattern of results. We offer our analysis of what we think these are in the Discussion section.

The GORT is the only test that we examined that was also examined by Cutting and Scarborough (2006). It is amazing how well our findings on the GORT converge with findings by Cutting and Scarborough. We both report the same amounts of variance independently accounted for both by decoding skill, .075 in both studies, and in listening comprehension/oral language, .096 in ours and .093 in theirs. There were only differences between the studies in the amount of shared variance, with theirs being larger because their oral language measure focused more on vocabulary tests than on the oral discourse measures we used, and vocabulary is an important component of word decoding skill.

Developmental Differences

Because our sample included children across a broad age range, we could determine not only whether there were differences between the tests in what predicted reading comprehension, but also whether this differed with developmental level. Developmental differences were assessed both as a function of chronological age and of reading ability, defined by raw scores on the PIAT Word Recognition test.

To determine if word decoding skill differentially predicted reading comprehension as a function of chronological age, we again ran the Model 2 regression analysis in which listening comprehension is entered as the first step. The amount of variance accounted for after this first step represents that explained by listening comprehension and its shared variance with word decoding, shown in the first row of Table 6. Then to determine whether the amount of variance accounted for by word decoding interacted with age, we entered an interaction term for the interaction of decoding and age. We created the interaction term by multiplying each child's chronological age by their composite word decoding *z* score. The second row of Table 6 shows the additional variance accounted for in each test by the interaction of age and decoding. As can be readily seen, the amount of variance accounted for by the interaction is much larger for the PIAT (.31) and WJPC (.27) than for the other tests (all ≤ .06), although all were significant because of our large sample. We tested the significance of the differences in variance explained by the interaction of age and word decoding across the tests by using Fisher *Z* tests. There is not a significant difference between the PIAT and the WJPC in the amount of variance accounted for by the interaction of word decoding and age; *Z* is less than 1 ($p = .19$). However, comparing either the PIAT or the WJPC against the other three tests, the *Z* statistics were always greater than 5 with $p < .001$, whereas those three tests were not significantly different from each other. Thus, the interaction term analyses show that there are developmental differences across tests in what is being assessed.

Perhaps the easiest way to see these developmental differences across tests is to examine the top half of Figure 2. Although the regression analyses examining whether word decoding interacts with age were conducted across the full sample, for purposes of illustrating developmental trends across the tests, the top half of Figure 2 displays the amount of variance accounted for in each test by decoding and listening separately for the two halves of our sample, using a median split on

TABLE 6
The Percentage of Variance Accounted for on Each of the Five Reading Comprehension Tests by the Interactions of Word Decoding Either With Chronological Age or With Reading Age (Raw Score on PIAT Word) After First Accounting for Listening Comprehension and Its Shared Variance With Decoding

	GORT ΔR^2	*QRI–Retell* ΔR^2	*QRI–Qs* ΔR^2	*PIAT* ΔR^2	*WJPC* ΔR^2
Step 1. Listening Comprehension	.218*	.259*	.287*	.246*	.291*
Step 2. Chronological Age × Decoding	$.060*_{a}$	$.033*_{a}$	$.031*_{a}$	$.312*_{b}$	$.270*_{b}$
Step 2. Reading Age × Decoding	$.062*_{a}$	$.034*_{a,b}$	$.017*_{b}$	$.261*_{c}$	$.255*_{c}$

Note. Values with different subscripts are significantly different from each other at $p < .01$. GORT = Gray Oral Reading Test–3; QRI = Qualitative Reading Inventory–3; PIAT = Peabody Individual Achievement Test; WJPC = Woodcock–Johnson Passage Comprehension subtest. $*p < .01$.

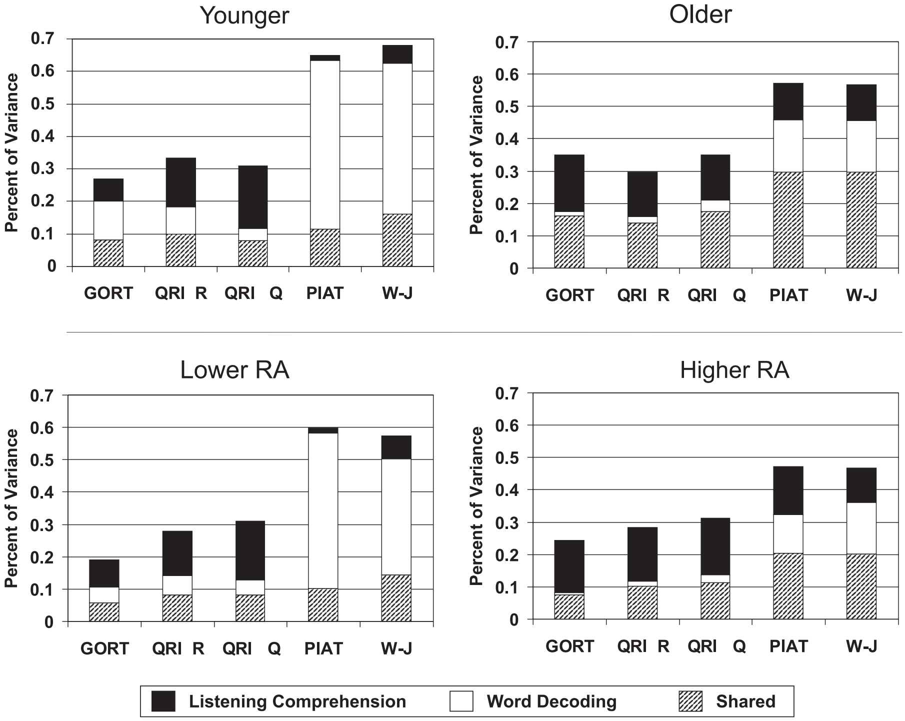

FIGURE 2 The proportion of total variance in each of the reading comprehension tests that was accounted for independently by word decoding skill, by listening comprehension skill, or shared for groups defined by chronological age (top figures), and reading age on Peabody Individual Achievement Test (PIAT) Word Reading (bottom figures). *Note.* GORT = Gray Oral Reading Test–3; QRI R = Qualitative Reading Inventory–3 Retellings; QRI Q = Qualitative Reading Inventory–3 Comprehension Questions; W-J = Woodcock–Johnson Passage Comprehension subtest.

age where the *younger* group's mean age was 9.1 years and the *older* group's was 13.1 years. It is readily apparent from this figure that decoding skill accounts for more variance when children are younger than when they are older, a finding that is well established in the literature (Hoover & Tunmer, 1993). What is new in our results is that there are such large discrepancies across tests in these developmental differences. As this figure shows, and as the values for the interaction terms of age with decoding (second row of Table 6) showed, there are dramatic differences across tests as a function of age in the amount of variance accounted for by word decoding. These developmental differences are large on the PIAT and WJPC, whereas on the GORT and QRI measures, they are small.

Because reading problems are defined relative to expectations for what is appropriate for age and grade level, it seemed equally important to look at possible differences in what the tests were measuring as a function of word reading ability, independent of whether that was above or below expectations. The same analyses performed for chronological age were therefore repeated using raw scores on the PIAT word recognition test instead of chronological age. The bottom row of Table 6 shows the differences between the tests in how the amount of variance accounted for by decoding interacts with a child's word decoding ability. The pattern for reading age analyses is similar to the results we previously reported for chronological age. Again, the amount of variance accounted for by the interaction is much larger for the PIAT (.261) and WJPC (.255) than for the other tests (all ≤ .06), although all were again significant because of our large sample. Fisher *Z* tests assessing the significance of the difference in the amount of variance accounted for by the interaction terms again showed that the PIAT was not significantly different than the WJPC ($p = .45$), but both were significantly different than the other three tests, the *Z* statistics all had $p < .001$.

Again, the easiest way to see these developmental differences across tests is to examine Figure 2, where the bottom half displays the amount of variance accounted for in each test by decoding and listening separately for the two halves of our sample defined by a median split on PIAT word reading raw scores. The difference between the two ability groups is most evident in the larger amount of variance independently accounted for by decoding skill in the children with lower word reading ability. This is most apparent on the PIAT and WJPC where the amount of variance accounted for independently by decoding declines from .48 to .12 for the PIAT and from .36 to .16 for the WJPC. Thus, again we are seeing evidence that the PIAT and WJPC are different than the other tests not only in terms of how much of their variance is accounted for by decoding skill, but also because what they measure depends on developmental level. If children are young or have low reading ability, these tests are more assessments of decoding skill, whereas for more advanced readers, they also assess listening comprehension skills and shared variance with decoding.

DISCUSSION

Comprehension is a complex cognitive construct, consisting of multiple component skills. Even though this complexity is recognized theoretically, when it comes to assessment, there is a tendency to ignore it and treat tests as if they are all measuring the same "thing." This is reflected in the fact that researchers who measure comprehension rarely give information on why they chose the particular test that they used. Implicit in this behavior is the suggestion that it does not really matter which test was used because they are all measuring the same construct.

Many researchers who use comprehension tests often have considerable experience with using word reading assessments. We think that experiences with the interchangeability of word reading tests may underlie the tendency to assume similar comparability of comprehension tests. Word reading instruments tend to correlate very highly; as noted in the Method section, our two measures of word reading correlate $r = .88$. However, the intercorrelations we observed among our five reading comprehension measures were much lower. Even our most highly correlated tests (the PIAT and the WJPC) correlate less highly than most word reading measures. The modest correlations we observed among reading comprehension tests suggest that the assumption that reading comprehension tests are all fairly comparable is not correct.[2] They suggest that these tests are not all measuring the same thing—a point that was substantiated by all our subsequent analyses.

The results of our factor analysis showed that two of the reading comprehension tests, the PIAT and the WJPC, load highly on decoding, whereas the GORT and the QRI measures do not. The regression analyses showed that most of the variance on these two tests is accounted for by individual differences in decoding skill, whereas decoding plays a rather small role in accounting for performance on either the GORT or the QRI.

The analyses examining developmental differences as a function of chronological age and of reading age further supported the conclusion that the tests are not all measures of the same skill. Although our findings replicate previous research showing more influence from decoding skills for younger and less skilled readers (Catts, Hogan, & Adolf, 2005; Hoover & Tunmer, 1993; Kirby, 2006), they also extend this finding in two important ways. One way is that they show that the very same test, particularly the PIAT and the WJPC, can measure different skills depending on developmental level. The other is that they show that there are greater differences between what reading comprehension tests are measuring when children are younger, less skilled or reading disabled. For less skilled and younger readers, the PIAT and WJPC differ dramatically from the GORT and the QRI, whereas for more skilled and older children, the differences between the tests are much smaller.

Illustrations of Comprehension Test Items

We think that inspection of test items from these reading comprehension tests makes it fairly apparent why the tests differ in their sensitivity to word decoding skills. Because we cannot publish actual items from standardized tests, we simply

[2] Modest correlations can also be interpreted as reflecting lowered reliability of the measures. It should be noted, however, that because comprehension involves so many different component processes, the lower correlations are likely to reflect differential assessment of the components. For example, even within the same test, modest correlations can occur between performance on different passages because the reader may have knowledge about one topic and not another.

describe our variations that parallel their sentence forms but that use different vocabulary.

On the PIAT (and the PIAT–R), children read a single sentence; then the sentence is removed and four pictures are presented for the child to show comprehension of the sentence by selecting the picture that best represents the sentence's meaning. The four choices depict alternatives that would correspond to incorrect decodings of one of the words in the sentence. So, if the sentence were *The patients were amazed by the giraffe in the lobby*, then the wrong answer pictures would depict events using a word similar to *patients*, like *parents*, and a word similar to *giraffe*, like *graffiti*. Thus, the child would be required to select among pictures of *patients amazed by the giraffe, patients amazed by graffiti, parents amazed by the giraffe,* and *parents amazed by graffiti.* In short, correct word decoding is the essence of choosing between the alternatives.

On the WJPC, the correct response also often hinges on correct decoding of a single word. To illustrate the importance of one word, we use Xs to substitute for that word. For example, consider this passage: *I thought that the painting was too XXXX. I did not, however, feel like arguing about the* __________. The typical response in examples like this is for the child to fill in the blank by saying "painting"; so that the completion would be *I did not feel like arguing about the painting*. However, that is incorrect. The correct response is *size* because the word Xed out was *enormous*. A child who has the comprehension skills to know that the blank must refer back to something in the first sentence demonstrates that knowledge by saying *painting*. However, the test scores the child as not having those skills because the assessment of comprehension for this item rests on decoding that one word, *enormous*.

How Test Format Differences Affect the Components of Comprehension Measured

Two previous comparisons of reading comprehension tests (Francis et al., 2005; Nation & Snowling, 1997) involving tests with a cloze procedure found that the cloze test differed from other tests in that most of its variance was accounted for by decoding skill. Our finding that the PIAT and the WJPC are so similar is interesting because it suggests that it is not just tests using a cloze format that are so heavily influenced by decoding skill. We now see that the PIAT, which uses multiple-choice selection of pictures representing the meaning of the sentence, shows the same pattern as the WJPC's cloze-test format, suggesting that some other factor besides the format of the test item is responsible for this pattern of results.

We would like to suggest that one factor that the PIAT and the WJPC share that leads them to be so heavily influenced by decoding skill is that the passages are all short. The PIAT uses single sentences; most of the WJPC items involve two-sentence passages, although some are also only a single sentence (only one item uses three sentences). In our view, there are two reasons why using one- or two-sentence passages results in a reading comprehension test that assesses decoding skill more than comprehension. One is that the assessment of comprehension in a short text is likely to be based on the successful decoding of a single word. This was illustrated previously in our example of a cloze item where failure to decode just one word, *enormous*, leads to an incorrect response, and in the PIAT where decoding confusions appear to be the sole basis for constructing the alternatives on each test item.

Another reason that short passages tend to be more influenced by decoding is that decoding problems are likely to be more catastrophic in short passages than in longer passages. In a single sentence, there frequently are no other words for the child to use to help determine the correct decoding of difficult words, such as *magician*. In a longer passage, however, the text is likely to describe events, such as pulling a rabbit out of a hat, which would allow the child to use this context to determine the correct decoding. Our speculations about this are reinforced by our findings that decoding skill accounted for much less variance on the QRI measures, where the passages are quite long and decoding problems can often be rectified by context.

What Does the GORT Assess?

The length of the passages in the QRI provides the contextual support needed to minimize the impact of individual differences in decoding in predicting comprehension. But why is decoding playing such a small role in explaining the variance on the GORT? We think the answer lies in the fact that the examiner, not the child, reads the test questions on the GORT and many are passage-independent. Keenan and Betjemann (2006) recently showed that most of the GORT items can be answered with above-chance accuracy without even reading the passages. Thus, not only do children not need to read the test questions on the GORT, they really do not even need to read the passage itself to answer the questions correctly. This suggests that a child with little decoding skill can do as well on the comprehension questions as a child with excellent decoding skills. In fact, when Keenan and Betjemann examined what was the best predictor of performance on the comprehension questions by people who actually read the GORT passages, it was not how accurately they read the passages, but rather how easily the question could be answered without reading. Keenan and Betjemann concluded that the GORT is not assessing decoding skills as much as reasoning from prior knowledge. The findings in this paper on the GORT converge nicely with Keenan and Betjemann's findings.

CONCLUSION

We believe that our findings have important implications both for research and clinical assessment. For research, it means that the answers to research questions could vary as a function of the specific test used to assess comprehension. To illustrate, suppose one was interested in assessing the extent to which decoding skill and comprehension skill are associated with similar genes (e.g., Keenan et al., 2006). If one used the PIAT or the WJPC to assess comprehension, the answer would more likely be that the same genes appear to be involved in word reading and comprehension, especially if the data were from young or less skilled readers, because these comprehension tests assess mainly decoding skill in these readers. In fact, such a conclusion has been reported by Byrne et al. (2007), who assessed comprehension with the WJPC in twins tested at the end of first grade. If they had used the QRI, we contend that their findings would at least have more potential to show different genes associated with decoding and comprehension because what is being assessed by the QRI is not just decoding.

Similarly, if a clinician used the WJPC or PIAT to test a child's comprehension skill, a child with only poor decoding skills would appear as if he or she also had poor comprehension skills. A child who really did have both poor decoding and poor comprehension could appear as if he or she had good comprehension skills if that child was tested on the GORT, where many of the questions can be answered without comprehending the passage by just using prior knowledge.

We hope that our findings that different reading comprehension tests measure different skills, and that sometimes even the same test measures different things depending on age and ability, will motivate researchers to further examine what the various comprehension tests that are available on the market are measuring. In pursuing this research, it will be important to recognize that the answer to what a comprehension test is measuring might also be affected by the specific variables used to carve up the variance in reading comprehension—such as using a global measure of listening comprehension of discourse versus using only a single component of oral language like vocabulary—as well as the specific operational definitions of these variables (cf. Bowyer-Crane & Snowling, 2005; Keenan, Betjemann, & Olson, 2007). Progress in science and validity of diagnoses depend on measurement instruments. As we have shown, when the construct being measured is as complex as comprehension, those instruments can tap different aspects.

ACKNOWLEDGMENTS

This research was supported by a grant from NIH HD27802 to the Colorado Learning Disabilities Research Center, for which R. Olson is PI and J. Keenan is a co-PI. Rebecca Betjemann was supported by NIMH training grant T32 MH016880-25. Portions of these data were presented at the Midwestern Psychological Association Meeting, 2005 and the Society for the Scientific Study of Reading, 2006.

We thank Laura Roth, Amanda Miller, and Sarah Priebe for discussions of the data; Laura Roth for comments on the manuscript; all the participants and their families; and all the testers and scorers.

REFERENCES

Barnes, M. A., & Dennis, M. (1996). Reading comprehension deficits arise from diverse sources: Evidence from readers with and without developmental brain pathology. In C. Cornoldi & J. Oakhill (Eds.), *Reading comprehension difficulties: Processes and intervention* (pp. 251–278). Mahwah, NJ: Erlbaum.

Barnes, M. A., Dennis, M., & Haefele-Kalvaitis, J. (1996). The effects of knowledge availability and knowledge accessibility on coherence and elaborative inferencing in children from six to fifteen years of age. *Journal of Experimental Child Psychology, 61*, 216–241.

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.

Byrne, B., Samuelsson, S., Wadsworth, S., Hulslander, J., Corley, R., DeFries, J. C., et al. (2007). Longitudinal twin study of early literacy development: Preschool through Grade 1. *Reading and Writing: An Interdisciplinary Journal, 20,* 77–102.

Catts, H. W., Hogan, T. P., & Adolf, S. M. (2005). Developmental changes in reading and reading disabilities. In H. W. Catts & A. G. Kamhi (Eds.), *The connections between language and reading disabilities*. Mahwah, NJ: Erlbaum.

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.

Davis, F. B. (1944). Fundamental factors of comprehension of reading. *Psychometrika*, *9*, 185–197.

Defries, J. C., Filipek, P. A., Fulker, D. W., Olson, R. K., Pennington, B. F., & Smith, S. D. (1997). Colorado Learning Disabilities Research Center. *Learning Disabilities, 8*, 7–19.

Dunn, L. M., & Markwardt, F. C. (1970). *Examiner's manual: Peabody individual achievement test*. Circle Pines, MN: American Guidance Service.

Francis, D. J., Fletcher, J. M., Catts, H. W., & Tomblin, J. B. (2005). Dimensions affecting the assessment of reading comprehension. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 369–394). Mahwah, NJ: Erlbaum.

Hoover, W. A., & Tunmer, W. E. (1993). The components of reading. In G. B. Thompson, W. E. Tunmer, & T. Nicholson (Eds.), *Reading acquisition processes* (pp. 1–19). Adelaide, Australia: Multilingual Matters.

Keenan, J. M., & Betjemann, R. S. (2006). Comprehending the Gray Oral Reading Test without reading it: Why comprehension tests should not include passage-independent items. *Scientific Studies of Reading*, *10*, 363–380.

Keenan, J. M., Betjemann, R. S., & Olson, R. K. (2007). *How do the specific measures used for decoding and comprehension influence the assessment of what a reading comprehension test measures?* Manuscript in preparation.

Keenan, J. M., Betjemann, R. S., Wadsworth, S. J., DeFries, J. C., & Olson, R. K. (2006). Genetic and environmental influences on reading and listening comprehension. *Journal of Research in Reading, 29,* 79–91.

Kirby, J. (2006, July). *Naming speed and fluency in learning to read: evaluation in terms of the simple view of reading*. Paper presented at the annual meeting of the Society for the Scientific Studies of Reading, Vancouver, British Columbia.

Leslie, L., & Caldwell, J. (2001). *Qualitative Reading Inventory–3*. New York: Addison Wesley Longman.

Markwardt, F. C. (1989). *Peabody Individual Achievement Test–Revised.* Bloomington, MN: Pearson Assessments.

Markwardt, F. C. (1997). *Peabody Individual Achievement Test–Revised* (normative update). Bloomington, MN: Pearson Assessments.

Nation, K., & Snowling, M. (1997). Assessing reading difficulties: the validity and utility of current measures of reading skill. *British Journal of Educational Psychology*, *67*, 359–370.

Olson, R. K. (2006). Genes, environment, and dyslexia: The 2005 Norman Geschwind memorial lecture. *Annals of Dyslexia, 56*(2), 205–238.

Olson, R., Forsberg, H., Wise, B., & Rack, J. (1994). Measurement of word recognition, orthographic, and phonological skills. In G. R. Lyon (Ed.), *Frames of reference for the assessment of learning disabilities: New views on measurement issues* (pp. 243–277). Baltimore: Brookes.

Pearson, P. D., & Hamm, D. N. (2005). The assessment of reading comprehension: A review of practices—past, present, and future. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 13–69). Mahwah, NJ: Erlbaum.

RAND Reading Study Group. (2002). *Reading for understanding: Toward an R & D program in reading comprehension.* Santa Monica, CA: RAND.

Rimrodt, S., Lightman, A., Roberts, L., Denckla, M. B., & Cutting, L. E. (2005, February). *Are all tests of reading comprehension the same?* Poster presentation at the annual meeting of the International Neuropsychological Society, St. Louis, MO.

Wechsler, D. (1974). *Examiners' manual: Wechsler Intelligence Scale for Children-Fourth Edition*. San Antonio, TX: Psychological Corporation.

Wechsler, D. (1981). *Examiner's manual: Wechsler Adult Intelligence Scale–Revised*. New York: Psychological Corporation.

Wiederholt,, L., & Bryant, B. (1992). *Examiner's manual: Gray Oral Reading Test–3*. Austin, TX: Pro-Ed.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock–Johnson III tests of achievement*. Itasca, IL: Riverside.