Reading Horizons: A Journal of Literacy and Language Arts

Volume 23
Issue 4 *July 1983*

Article 3

7-1-1983

Reconciling Differences in Test Results: Comprehension

Barbara A. Hutson
VPI and State University

Jerome A. Niles
VPI and State University

RECONCILING DIFFERENCES IN TEST RESULTS: COMPREHENSION

Barbara A. Hutson & Jerome A. Niles
VPI & STATE UNIV., VA

In planning an instructional program for Brenda you have discovered that one of her tests indicates an instructional level of fourth grade for comprehension and another test shows comprehension at the high second grade level. How can both results be accurate? How do you decide about their accuracy? If both are true what does that indicate about her profile of abilities? How can you turn what appears to be a testing anomaly into useful diagnostic information?

Inaccurate, Misleading, or Irrelevant Test Results

There is always the possibility that one of your test results is inaccurate. Many of the diagnostic tests have only one or two brief passages per grade level. Some prior experience with the topic, a relevant schema, may help students in answering questions even on passages they cannot actually read well; the lack of such experience can distort comprehension even when a student accurately decodes the passage. For example, on the Diagnostic Reading Survey (Spache, 1972) there is a passage that talks about shifting gears as a metaphor for shifting speeds in reading. A bright third grader could decode the passage, but was stumped by a question about shifting from gear to gear, for which nothing in her experience had prepared her. An error on this question brought her below the criterion for comprehension at the seventh level. Was this result accurate? Perhaps not, though in this case it didn't matter all that much—it was obvious that she could read orally with comprehension passages several years above her grade level.

It's also possible that the test you are using is intended only for global differentiation. For example, the Gray Oral Reading Test (Gray & Robinson, 1967) gives a reasonably precise estimate at lower grade levels but has a standard error of estimate of more than one year at the upper levels. This means that for a student with a tested grade equivalent of tenth grade on the Gray, his/her "true" score is likely to range between eighth and twelfth grades (the score plus or minus two standard errors of measurement). Because of the imprecision of measurement on many tests, you may not have strong grounds for interpreting differences between tests or subtests unless scores are two or three years apart or other observations support these findings.

Some testing strategies produce results that are not inaccurate but are potentially open to misinterpretation. You may decide to administer a test in a nonstandard way, but if you do, you must take that into consideration in interpreting results. For instance, allowing bright students to begin two thirds of the way through a test may deprive them of the benefit from practice on easier items and they may thus receive a somewhat deflated score. More often, though, the problem is an inflated score. Going past the specified cutoff point of decoding errors for oral reading on a test such as the Standard Reading Inventory (McCracken, 1966), for example, may yield valuable information, yet if you want to use the test norms you must score responses in terms of the normal cutoff and only report the later responses as additional information. (You also need to consider whether the nonstandard administration will "spoil" that test for use with this student during the next year.)

If you are diagnosing a student who often declines to answer questions, you may decide to test limits by pushing harder or waiting longer than usual for a response, or to probe by modifying the test item to determine the conditions under which he/she can succeed. If you want to test the limits of students' thinking but are using an informal reading inventory that provides only literal questions, you may want to add some inferential questions or to have the students recall the story in order to assess their grasp of the theme and structure (unless the passage is too short or too devoid of plot or motivation to stimulate a revealing retelling). Any kind of deviation from a standard presentation may be well-justified, but you need to consider whether your presentation has so altered the test that it is unreasonable for you to use the norms or grade designations based on the assumption of a standardized presentation. If you find a discrepancy between a test result derived for a standard presentation, a conservative procedure is to accept the standard measure as a reasonable estimate of a student's usual performance but also to use the probed responses on this test or the nonstandard presentation of another test as an indication of the range of response available to the student under optimal conditions.

In addition, some tests may be irrelevant. If you are interested in assessing comprehension, a test of vocabulary in isolation such as the Wide Range Achievement Test, though it provides a score called "Reading," misses the mark by a wide margin. High scores on such a test, however, rule out decoding skills as a source of comprehension problems. Such narrow-band tests, though, should not be interpreted as a measure of comprehension.

If, however, you've checked and found that none of the troublesome test results are not inaccurate, misleading, or irrelevant, you face perhaps the most intricate problem in diagnosis, determining why two tests that supposedly assess the same thing yield different results for a given student.

Examining Differences Between Tests

Sometimes your test results are accurate, reasonably precise and obtained in a standard manner, yet two findings are incontro-

vertibly different. That's when (after deciding whether the discrepancy is important enough to investigate) a true professional brings to bear all of his/her knowledge and analytic skills in attempting to reconcile test differences, perhaps the most demanding aspect of diagnosis. What are the differences in the responses required to demonstrate competence on these two tests? Even on two tests that supposedly measure the same ability there may be important differences in (1) modes of presentation and response; (2) thinking processes required; or (3) scoring procedures and criteria for success. If you consider carefully these differences between tests, you may resolve discrepancies or, better yet, obtain a more finely differentiated profile of abilities for a student.

Differences in Modes of Presentation and Response

Reading/language tests vary in the way materials are presented and the responses by which reading performance is measured. Presentation differences such as page format can produce significant disparities in test scores, particularly at lower grade levels. For example, tests which have the questions separate from the passage can be a problem and tests which require a separate answer sheet can be a disaster for some students. Other students, especially in the earlier grades, might be disturbed by the cloze format for comprehension of the Woodcook Reading Mastery Tests (1973) or the complex task structure for the Word Meaning subtest of the Test of Reading Comprehension (Brown, Hammill, & Wiederholt, 1978), for example, unless they've had prior experience with that format.

For some students, performance varies greatly depending on whether the material is presented orally or in print. It's not unusual for a student's score on a listening comprehension test or subtest to be higher than his/her score on a reading comprehension test. A low reading comprehension score paired with a much higher listening comprehension score presents a much different diagnostic picture than a low reading comprehension score paired with an equally low listening comprehension score.

Tests also differ in the responses by which they ask the reader to demonstrate comprehension. The primary dimensions of variation for response mode are oral versus written and recall versus recognition (production versus selection). Each year Mark consistently scored better on the end of the year achievement test than he did on teacher-made tests of comprehension and in the workbook. This discrepancy frustrated his parents and puzzled his fifth grade teacher, Miss Long, who could not understand why Mark did not do better in class. Mrs. Sherman, the reading teacher, was asked to consult on the problem. After observing Mark's classwork in reading and his test performance, she found one possible explanation for the score differences.

Mark had a severe writing problem. In fact he even had difficulty copying material from the board, much less spelling words recognizably. Mark's writing problem precluded successful performance in classroom reading where success depended primarily upon written responses to comprehension questions. On the other hand, Mark's contributions in discussion reflected good comprehension.

Discussion performance, however, was not part of the criteria for grading reading performance in Mark' s class. Mrs. Sherman pointed out the probable reasons for Mark's differences in performance in comprehension and explained to Miss Long the importance of providing alternative measures of comprehension performance.

Miss Long thought Mrs. Sherman's discovery was an important one and she immediately brought another child to her attention. Miss Long observed that Cindy did not do well in her written work or the group discussions, yet her achievement test scores were as impressive as Mark's. After reviewing Cindy's classwork and test performance, Mrs. Sherman found that Cindy consistently did better on measures which gave her multiple choices and asked her to select a response than on measures which asked to create a response. The achievement test she took each year used the recognition format to measure reading ability. Miss Long and Mrs. Sherman discussed this difference and planned some trial teaching lessons to collect more information to solve the problem of Cindy's apparent difficulty in producing responses on comprehension measures.

One of the most common kinds of discrepancy is the difference between a student's performances on measures of oral and silent reading comprehension. Since both kinds of measures are frequently used in assessing and evaluating reading performance, it is crucial that the diagnostician understand and be sensitive to the differential effects that are a result of the requirements of these two tasks. Differences between a student's performances on oral and silent reading can sometimes be traced to his/her perceptions of the purpose of the task. If the student senses that the teacher is interested in correct pronunciation and fluency in oral reading, he/she may limit processing of text to the surface structure language and not attend to units of meaning. Thus, a pattern might emerge which shows one reader to have much better comprehension when reading silently than orally. The reverse may be true for another reader, who conceives of silent reading as "brushing the print with your eyes," and depends upon the auditory trace of his oral reading to aid his comprehension and memory.

Prior instruction or practice can also cause comprehension performance differences. Beginning readers typically practice much of their reading orally. Moreover, most of their pre-school experience with reading was through having accomplished readers read books orally to them. Thus beginning readers often perceive reading as a task that naturally involves production of speech, and a diagnostician might expect their oral reading to be better than their silent reading.

One type of reader who is frequently misdiagnosed because of a failure to reconcile oral and silent test performances is the highly anxious or nervous child. High levels of anxiety clearly affect the fluency with which skilled behavior can be conducted. Reading orally in a testing situation, especially if the reader has a history of failure, can be traumatic, and no amount of examiner rapport can entirely overcome this feeling. The result is a product which reflects numerous oral reading miscues and most likely a depressed comprehension score or such an intense concentration on oral accuracy that comprehension suffers. For

some of these children, the privacy of silent reading provides a comfortable haven which allows them to conduct the reading process with the required fluency.

Differences in Processing

Tests also vary in the thinking processes they require or permit. The types of processing may include location of explicitly stated answers to a literal question, transformation of explicit information in text into a slightly different form, drawing inferences about the relationship between two facts stated in the text or about the relationship of a fact in the text and information drawn from the readers' experience, and judgments about the structure or purpose of the text. One arrangement may permit a given reader to use his preferred processing strategies, while another arrangement forces him/her to use less familiar or less comfortable strategies. For example, a student who is used to being asked "What color was John's coat?" may be derailed when asked "What is the main idea of this story?" In contrast, a student who is used to reading independently to gather information relevant to solution of a broad problem may be startled if asked a question about a bit of information no bigger than his/her thumbnail. Either of these assessment procedures is legitimate and useful, but the two strategies are likely to interact with a student's experiences and expectations for comprehension questions and ultimately require different cognitive processes.

Some readers are affected more than others by the cognitive demands of the reading test. Tina, for example, integrates information from her reading well and connects it to her personal experiences. On the Silent Reading subtest of the Durrell Analysis of Reading she had little opportunity to display these skills and in fact missed some points for small factual errors. (Points are allotted on the basis of number of facts recalled, major or minor.) On the Reading Miscue Inventory (Goodman & Burke, 1971), though, she obtained a relatively high comprehension score by retelling the major points of a story in a coherent fashion. A student with a set toward surface level processing and retention of details might have had exactly the opposite pattern.

Results on comprehension tests may also vary depending on whether the questions require the student to deal with directly stated facts, simple transformations of text-explicit material or for example more inferential processing. When the test states "Before he ate dinner Jack rode his bicycle," the question might ask "What did Jack do after he rode his bicycle?" On the other hand, the test may incorporate questions which deal with more implicit relationships in the text and demand inferences and applications by the reader. Tom does well on exact recall of facts, but because he fails to combine information from the text with his experiences and common sense, he does poorly on tests such as the new Metropolitan Intermediate Survey Test (Prescott, Balow, Hogan & Farr, 1978), which taps higher level thinking skills. Performance on comprehension questions can not be lumped together indiscriminately. To obtain an accurate student profile, the diagnostician must consider the cognitive requirement of the questions and the individual differences of the reader.

Differences in Scoring and in Criteria for Success

Test scores sometimes differ because responses scored as errors on one test may not be sored as errors on another test. For example, hesitations and repetitions in oral reading are scored in oral accuracy counts that along with comprehension, determine grade levels on Silvaroli's Classroom Inventory, while on other measures, such as the Johns' Basic Reading Inventory (1981), only meaning-change errors are counted for the word recognition criteria. Thus, a reader may make 10 unexpected responses while reading, yet only four of them change the author's intended meaning. Clearly there will be significant discrepancy on how these two tests judge a reader's competence if the score is accepted on face value without thoughtful interpretation by the diagnostician.

Variation in IRI test scores can also complicate the diagnostician's effort to establish an instructional comprehension performance level. The criterion established by the authors for a number of tests is 75% while several others use 60% as their cutoff for satisfactory performance. Ignoring the fuzziness or lack of precision of comprehension criteria can obscure evidence of the reader's competence and hinder the diagnostician from assembling an accurate description of the reader's abilities.

The problem of a satisfactory comprehension criterion is especially troublesome when it interacts with the type of processing required. Some reading tests, such as the Basic Reading Inventory, (Johns, 1981) are designed to assess various features of a reader's comprehension ability. The tests examine the reader's prior knowledge through vocabulary and inference questions, reasoning ability through inference and evaluation questions and information pick-up through literal level questions. It is easy to imagine a reader who receives ten questions; he answers six of seven literal level questions correctly and misses the vocabulary, inference, and evaluation questions. Using a comprehension criterion of 75%, this student would have failed this passage. Without thoughtful reconcilation, this reader's poor comprehension performance on the Basic Reading Inventory could be quite confusing if the diagnostician was trying to compare the result to another comprehension measure which used only passage dependent literal level questions. Using tests which "average" together a number of different comprehension aspects is a common practice and the diagnostician must be aware of the effects on the data.

It's been suggested in this section that in attempting to reconcile discrepant scores on reading comprehension measures the diagnostician consider whether two tests differ in the way they present materials, the way students must respond, the kinds of processing required, the means of scoring, and the criteria set for success. Although we've discussed these separately, in practice they are generally interdependent factors. The differences we've discussed are surely not the only ones that matter, but they provide a good start toward analyzing and reconciling test differences.

Deriving a Profile of Abilities

Our intent has been to point out how test differences can

occur and how to make sense of them. We would like to take that a step further and suggest that you "bracket" your readers' comprehension ability by deliberately using tests with different characteristics. In this way you can gauge the range of their ability. A comparison of two readers, Larry and Ron, on three measures of reading helps to illustrate this point.

On one measure of comprehension Larry and Ron seemed very similar in ability, but an examination of differences from one test to another reveals different profiles of abilities. On tests given in February their scores were:

	Larry - 6th Grade	Ron - 6th Grade
Gray Oral Reading	6.0	5.0
Durrell Silent Reading Comprehension	6.0	5.0
Metropolitan Survey	4.5	7.0

For Larry there was no difference between scores on silent and oral reading on tests that emphasize literal comprehension. The Metropolitan, however, emphasizes inference, a major weakness of Larry's.

Ron's silent reading score was higher than his oral reading scores, although both scores were based on literal comprehension. His score on the Gray Oral Read ng est was brought down by a number of small, meaning-preserving errors in oral reading. Although he was not outstanding on tests composed primarily of literal questions, he performed better than his age-mates on a test which emphasized inferential questions, as the high score on the Metropolitan Survey indicated. He could use signal words and text structures, in combination with his own experiences, to infer meanings not explicitly stated.

This pattern was also observed when the examiner conducted a functional analysis of the boys' skill and efficiency in using their content area text in science and social studies. Larry could use the Table of Contents and Index if the reference was listed under the heading he expected, but if he were looking for trucks and found no such thing, it never occurred to him to look under transportation. He could use subheadings to locate major divisions of the text but could not easily skim to locate specific facts. He read carefully but became swamped with facts and had difficulty selecting key points or tying them together. Ron was a little less efficient on the mechanical aspects of content area reading, but used the structure of the material to help him locate, organize, and evaluate facts. He was a flexible reader, varying his speed and depth of processing to suit his purpose, the time available, and the difficulty of the material. The test scores, take together with purposeful observations, delineated sharply different profiles of comprehension abilities for these two boys.

Summary

While across large groups of students two tests may be highly correlated, specific characteristics of tests may interact with specific characteristics of students to yield differences in scores

for one individual on two or more tests. These differences may provide valuable information but require thoughtful interpretation. Examining and reconciling differences in test results for a student can help you not only to provide more accurate interpretations of test results but to gain a more complex and useful understanding of each student. The student's abilities, experiences and attitudes interact with specific features of each test; the thoughtful diagnostician can use the real and apparent discrepancies between tests to sketch the profile of abilities unique to a given student and to develop individual educational plans appropriate for that student.

REFERENCES

Bond, Guy, I.H.Balow, & C.J.Hoyt, New Developmental Reading Tests. Chicago: Lyons and Carnahan, 1968.

Brown, V.L., D.D.Hammill, & J.L.Wiederholt. Test of Reading Comprehension. Austin, TX: Pro-Ed, 1978.

Durrell, D.D. Durrell Analysis of Reading Difficulty. New York: Harcourt,Brace & World, 1965.

Durrell, D.D. & M.T.Hayes. Durrell Listening - Reading Series, Primary Level. New York: Harcourt, Brace Jovanovich, 1970.

Goodman,Y. & Burke,C. Reading Miscue Inventory. New York: The MacMillan Company, 1971.

Gray, W.S. & H.M.Robinson. Gray Oral Reading Test. Bobbs-Merrill, 1967.

Jastak,J.F., S.W.Bijou, & S.R.Jastak. Wide Range Achievement Test. Wilmington, Delaware, 1965.

Johns, J. Basic Reading Inventory. Toronto, Canada: Kendall/Hunt Publishing Company, 1981.

McCracken, R.A. Standard Reading Inventory. Klamath Falls, Oregon: Klamath Printing Co., 1966.

Prescott,G.A., I.H.Balow, T.P.Hogan, & R.C.Farr. Metropolitan Achievement Tests (Intermediate). New York: The Psychological Corporation, 1978.

Silvaroli, N. Classroom Reading Inventory. 3rd edition. Dubuque, Iowa: Wm. C. Brown Publishing Company, 1976.

Spache, G.D. Diagnostic Reading Scales, Revised edition. Monterey, California: CTB/McGraw-Hill, 1972.

Stroud, J.B., A.N.Hieronymus, and Paul McKee. Primary R ading Profiles, Level 2. Boston, Houghton Mifflin, 1968.

Woodcock, R.W. Woodcock Reading Mastery Tests. Circle Pines, Minn.: American Guidance Service, 1973.