DOCUMENT RESUME

ED 356 451 CS 011 261

AUTHOR Snyder, Jon; And Others

TITLE Assessment of Children's Reading: A Comparison of Sources of Evidence.

INSTITUTION Columbia Univ., New York, NY. Teachers Coll. National Center for Restructuring Education, Schools and Teaching.

PUB DATE 93

NOTE 18p.

AVAILABLE FROM NCREST, Box 110, Teachers College, Columbia University, New York, NY 10027 ($5, prepaid).

PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS Comparative Analysis; Elementary Schools; Elementary School Students; Grade 3; Pilot Projects; Primary Education; Public Schools; Readability Formulas; *Reading Ability; *Reading Achievement; Reading Research; *Reading Tests; *Student Evaluation; Test Interpretation

IDENTIFIERS Degrees of Reading Power Test; New York City Board of Education; Text Factors

ABSTRACT

A pilot study explored the nature of the relationship between test and classroom evidence and considered how these different sources of evidence were used in assessing and promoting student growth and development in reading. Test estimates of third-grade children in two public elementary programs in New York City were compared with estimates of the difficulty of the material the children were reading at the time of the testing. This was done by comparing one index of student reading abilities: Degrees of Reading Power (DRP) test scores, with a second index, i.e., the DRP readability values of books the students had actually read. Results of the comparative analysis indicated relatively close agreement between the test and classroom records for children who performed well on the test. For low-scoring students, however, there was a clear discrepancy between the two sources of evidence. According to classroom records, these students were reading and comprehending books beyond, sometimes several years beyond, the test's estimate of the reading abilities. Low-scoring students were actually reading materials that test evidence predicted would be beyond their capabilities. Several questions for future research raised by these findings are briefly discussed. (Two tables of data are included.) (RS)

C.S

Assessment of Children's Reading: A Comparison of Sources of Evidence

Jon Snyder
Edward Chittenden
Priscilla Eilington

NCREST

The National Center for Restructuring Education, Schools, and Teaching (NCREST) was created to document, support, connect, and make lasting the many restructuring efforts going on throughout the nation. NCREST's work builds concrete, detailed knowledge about the intense and difficult efforts undertaken in restructuring schools. This knowledge is used to help others in their attempts at change, to begin to build future education programs for school practitioners, and to promote the policy changes that will nurture and encourage needed structural reforms. The Center brings together the voices of practitioners and researchers, parents and students, policy makers and teacher educators.

NCREST is supported by a major grant from the DeWitt Wallace-Reader's Digest Fund. NCREST's work in New York City, through its Center for School Reform, has been supported by the Leon Lowenstein Foundation and the Aaron Diamond Foundation. Others funders have included the Center for Collaborative Education, the Danforth Foundation, the Fund for New York City Public Education, Impact II, the Metropolitan Life Foundation, the National Center for Research on Vocational Education, the New York Community Trust, the New York State Department of Education, and the Regional Laboratory for Educational Improvement of the Northeast and Islands.

Additional copies of this publication may be ordered for $5 each. All orders must be prepaid by check or money order payable to NCREST. Contact:

NCREST
Box 110, Teachers College
Columbia University
New York, NY 10027
Fax: (212) 678-4170

Assessment of Children's Reading:

A Comparison of Sources of Evidence

Jon Snyder
Edward Chittenden
Priscilla Ellington

April 1993

National Center for Restructuring
Education, Schools, and Teaching
Teachers College, Columbia University
New York, New York 10027

Preface

This article is an exploration of different sources that can be used to assess the development of the ability to read. It began with a collaborative effort between the authors and two New York City elementary schools to develop a self-study process wherein the school communities could hold themselves accountable to their principles, their students, and their families. In the course of the self-study work, the authors spent hours "hanging out" in childrens' files exploring how teachers assessed, recorded, and reported student development. These files were incredibly rich, containing books the students had read and why they had chosen to read them; writing samples based upon their reading; miscue analyses and other evidence offering insights into how the child read; notes from interviews with parents and children about reading; and marvelously descriptive anecdotes about the child's reading behaviors. Also included in these files were the child's reading test results from the DRP exam. At some point, we realized that the books that we knew the child had read could be given a DRP value and thus be compared to exam results. As described in the paper, the comparison clearly indicated that as test scores decreased, the discrepancy widened between a child's observed ability to read and his/her test results. This article analyzes this discrepancy and its educational implications.

We would like to thank the faculty and staffs at the two schools for their support of our work. In fact, without their work, this article could never have been completed. In addition, we would like to thank the Center for Collaborative Education for supporting the initial work on the self-study process which evolved into this paper. Linda Darling-Hammond helped us frame the issues and provided invaluable feedback on our successive attempts to express ourselves. Finally, we would like to thank NCREST editors Diane Harrington, Alice Weaver, and Elizabeth Lesnick for helping us share our findings with concerned educators and citizens who care.

Overview

In most public school systems, assessment of children's reading is based on two quite different sources of evidence. On the one hand, there are standardized tests, administered on a single occasion and yielding quantitative estimates of reading achievement. On the other, there is classroom-generated evidence, which is ongoing and based on teachers' everyday interactions with students and their work. The evidence from testing programs are the scores that are extrapolated from children's responses to reading-like tasks (e.g., test items). Classroom evidence is more direct and includes observations of actual reading behaviors and interests. Quite often, districts base decisions (e.g., curriculum, allocation of resources, and pupil placement) primarily on test evidence, while teachers base their decisions primarily on classroom evidence.

The purpose of this pilot study was to explore the nature of the relationship between test and classroom evidence and to consider how these different sources of evidence are used in assessing and promoting student growth and development in reading. Children's test results were compared with classroom records of reading behaviors. Specifically, test estimates of children's reading achievement were compared with estimates of the difficulty of the material the children were reading at the time of the testing. This was done by comparing two indices of student reading abilities: Degrees of Reading Power (DRP) test scores with the DRP readability values of books the students had actually read.

The results of this comparative analysis of the two sources of evidence indicated relatively close agreement between test and classroom records for children who performed well on the test. For low-scoring students, however, there was a clear discrepancy between the two sources of evidence. According to classroom records, these students were reading and comprehending books beyond, sometimes several years beyond, the test's estimate of their abilities. Low-scoring students were actually reading material that test evidence predicted would be beyond their capabilities.

Methodology

Since this was a pilot study to clarify our research design and obtain preliminary feedback, we limited the sample to third grade children enrolled in two public elementary programs in New York City. The procedures were straightforward. From each child's folder, we obtained the spring 1990 DRP test score as well as records of books the child had read during the months before and after the testing date. These were books that the teachers judged to be within the child's reading ability. If a teacher noted that a child was having difficulty with a particular text, we did not include it in that student's list of books read.

The next step was to establish a common denominator with which to compare the two sources. The DRP Readability Analysis Service provides readability values for books that may be used as estimates of text difficulty. The values are derived by applying a formula designed to generate a readability value for any piece of text. The reading formula is based primarily on language complexity (i.e., length and grammatic structure of sentence, length and familiarity of words). Thus, a passage of text with multisyllabic words, long sentences, odd tenses, and more complicated syntax would obtain a higher DRP value than would a short sentence of simple words. The higher the DRP value, the more difficult the text.[1] DRP scores, whether test- or text-generated, are obtained in

[1] DRP scores are not content based. For instance, Hemingway's ***The Old Man and the Sea***, with its characteristic short sentences and simple words, has a DRP rating of 50. Salinger's ***Catcher in the Rye*** has a similarly low DRP rating of 49. ***Blueberries for Sal*** has a DRP rating of 54. ***Chitty Chitty Bang Bang*** has a DRP rating of 59.

similar fashion by applying a consistent formula; the numbers thus generated can be used as a common denominator. The scores provide a basis of comparison between test and classroom evidence, in that both are expressed in DRP units and both reflect comparable assumptions about difficulty of reading matter. A serious drawback of using DRP units as a common denominator is the essence of those assumptions that the difficulty level of a passage can be derived from textual rather than contextual features. If one does not share this assumption, but assumes instead that reading difficulty is contextual rather than textual, then comparing test and text-generated DRP units might be considered a case of comparing rotten apples to rotten oranges.

DRP values for books were located within the ***Readability of Literature and Popular Titles*** (TASA, 1988) produced by the DRP Readability Analysis Service. It reports the DRP reading values of over 1,600 popular children's books culled from a wide variety of sources. We were able to assign a DRP value to many, but not all, of the books the children were reading. We excluded the books being read that were not included in the manual as well as books with high internal variability. With these exclusions, the records of twenty-seven students had comparable data. When the child had read two or more books for which we could obtain a DRP number, we used the median.

Findings

DRP values for books the children were reading were compared with the DRP test score that indicated the child's Independent Reading Level. Table 1 summarizes the results. The left-hand column records the DRP independent reading level generated by the child's performance on the test. The central column records the DRP level of the book(s) the child was reading concurrent to taking the exam. The column on the right shows the difference between the two sources of evidence in DRP units. For instance, one student, as indicated in Table 1, received a DRP independent reading level of 18 on the test. During the same time period, she was reading ***Amelia Bedelia*** with a mean DRP value of 42. The difference between her test score and the book she was actually reading was 24 DRP units. In this case, test evidence about this child placed her at the seventeenth percentile while classroom evidence placed her at the sixty-fifth percentile. This is the difference between qualifying for a remedial program and being significantly above average in a regular classroom. This pattern of difference was not unusual for students with low test scores. As indicated in Table 1, for those students who scored well on the exam there was a clear alignment of classroom and test assessment data. As test scores decreased, however, the gap widened between a child's observed work and test score.

In short, the test results were accurate indicators for high test achievers but became increasingly less accurate as test achievement decreased. At the lowest levels of test achievement (the bottom quartile), the test evidence indicated that certain students could barely read. In fact, they not only could read, but were able to read material almost as challenging as that read by their high-scoring classmates.

Discussion

It is important to note that the classroom practices of the participating teachers made it possible to gather and rely upon classroom data because of systematic school practices designed to utilize and share classroom-generated evidence. The fact that the teachers collected and recorded their observations concerning the books that children were reading, when, and at what level of difficulty and understanding, made this research possible. It would have been impossible to compare two sources of evidence if only one had been collected. In addition, these systematic school practices provided us with convincing documentation of the students' comprehension of the books read (e.g., miscue analysis, student journal entries, book reports, records of individual and group

Table 1

TEST RESULTS: INDEPENDENT LEVEL (IN DRP UNITS)	READABILITY OF BOOKS BEING READ (IN DRP UNITS)	DIFFERENCE BETWEEN TEST AND BOOKS READ (IN DRP UNITS)
15	43 (Curious George)	28
15	38	23
15	45	30
18	42 (Amelia Bedelia)	24
20	55 (Clue in the Old Stagecoach)	35
23	60 (Aesop's Fables)	37
23	41	18
23	40	17
27	40	13
29	51	22
29	50	21
29	47	18
29	43	14
29	42	13
50TH PERCENTILE	50TH PERCENTILE	50TH PERCENTILE
32	56 (Tuck Everlasting)	24
36	51 (A Wrinkle in Time)	15
36	46	10
37	53	16
39	40	1
40	48	8
46	54	8
48	52	4
50	54	4
50	51	1
54	52	-2
54	51	-3
56	52	-4

Note: A DRP score of 39 equates to a normal curve equivalent (NCE) of 51. Ninety percent of third-grade textbooks fall within DRP levels 44 to 53.

discussions, art and project work arising from these books) (MacDonald and Snyder, 1992). This documentation convinced us, for this sample, that the classroom-generated evidence was trustworthy and an accurate reflection of student reading ability.

While the work reported here is from a pilot study whose purpose was to clarify research design and generate preliminary feedback, several significant questions are raised. A first set of questions centers on why the test- and classroom-generated evidence were aligned for high-scoring students but not for low-scoring students. With the higher scores, it is possible that test (e.g., the test ceiling) and/or classroom characteristics (e.g., the lack of higher-level books available to students) can account for the seeming congruence between sources of evidence rather than any inherent compatibility.

It is difficult, however, to locate methodological explanations for the striking differences between evidence sources for students with low test scores. It seems clear that for many students, test results are not an accurate reflection of reading ability. When shown the results from this study, participating teachers offered two explanations. First, they argued that the test uses expository rather than narrative text samples, and, hence, the format is not a good indicator of student classroom reading work. Second, in the schools sampled, classroom work is based on student strengths and interests, while the test uses decontextualized and less inherently interesting passages, creating differences related to motivation. Similarly, the test situation is artificial, requiring a different type of performance, and this, for some students, creates unfamiliarity and/or anxiety. The essence of the teachers' explanations is that tests measure surface textual features rather than what students bring to the reading task. Since reading is an interpretive and constructive task, students read "better" in areas or topics with which they have experiential familiarity. Thus, reading "ability" is about conceptual maps and schemata at least as much as language structure (Anderson, 1984; Bussis, 1982). For instance, give one of the authors of this paper a biochemistry text with short easy sentences, and he may still fail to "read" it.

Another set of factors may revolve around the nature of exam preparation and the evolution of the testing industry. Item validation techniques play an important role in the construction of standardized achievement tests. According to Buros, "These techniques have been harmful to the development of the best possible measuring instruments" (1977, p. 12). The "dictates of objective measurements" have led to instruments that include questions teachers regard as poor and exclude questions teachers regard as good. He concludes, "Methods of statistically validating achievement tests insidiously tend to strengthen the status quo, to impede curricular progress, to perpetuate our present grade classification, to differentiate rather than to measure, to conceal unlearning, and to give an illusory sense of continuous learning" (p. 12). As validity checks increasingly became the domain of professional psychometricians, test sample selection was removed from children and teachers in classrooms and standardized tests became less accurate assessors of the ability to read.

Implications

There are implications of this study beyond analysis of the numbers generated. One such broader set of issues centers on the challenge to systematically include multiple sources of evidence to assess and support student growth. The matrix in Table 2, on the following page, provides a structure for considering this challenge.

The vertical column lists three categories of sources of evidence to assess student growth and development in reading (as well as in other areas), which, ideally, inform each other (Chittenden, 1991). One cluster is test and testlike events. Examples of such evidence include:

- standardized tests;
- skills worksheets/tests provided by many publishers;
- teacher-made work sheets and informal tests.

Table 2

SOURCE OF EVIDENCE	Classroom	School	District	State
Test/Test-Like Events	Sometimes Used	Nearly Always Used	Nearly Always Used	Nearly Always Used
Observations	Nearly Always Used	Sometimes Used	Rarely Used	Rarely Used
Performance Samples	Nearly Always Used	Sometimes Used	Rarely Used	Rarely Used

A second cluster of sources is observation and documentation of students at work. Examples of such evidence include observations of:

- what children read (e.g., what books children choose and why, content/authors/genres they enjoy, etc.);
- how children choose books (e.g., by author, category, theme, illustrator);
- where children read (e.g., at the library, at home, or at school -- if at school where);
- how children read (e.g., do they reread the same piece or is once enough, do they start with illustrations and then read the book or vice versa; do they read from cover to cover or do they skip around).

A third cluster of sources is analysis of student work itself (e.g., performance samples). Examples of such evidence include:

- informal reading inventories;
- miscue analysis;
- documenting oral reading (e.g., running records);
- discussions with children about their reading;
- children's writing.

Both teacher observation and analysis of student work were employed with the classroom evidence used in this study -- books children read. Teachers observed children selecting and reading these books, but they also listened to children read from them, discussed them with their students, and analyzed student work resulting from the books.

The horizontal column in Table 2 lists, in increasing size, institutional levels with the responsibility for the education of children. To uphold their moral and legal obligations, each of these institutions requires accurate assessment of student growth. The grid indicates which levels of the educational enterprise utilize which sources of evidence.

This research suggests what many teachers have argued for years: The careful analysis of day-to-day student work and observation of students over time is not only good teaching, but also provides a more accurate assessment of student growth and development than does once-a-year high-stakes testing. If this is indeed the case, then it would behoove districts and states to find mechanisms to tap into this more accurate vein of data to increase the potential for policy gold. One key task, then, is to move the more accurate classroom sources of evidence into the larger arenas of the school, the district, the state, and beyond (from left to right in Table 2).

The federal Chapter 1 program provides an example of the possible policy pitfalls of failing to use classroom evidence for decision making. Chapter 1 is a program specifically for students in lower socioeconomic communities who score poorly on standardized tests. The students (and thus the school programs -- as well as the national program writ large) are enrolled and evaluated primarily on standardized test scores. Yet our data indicate that standardized test scores provide questionable bases for the very clientele Chapter 1 has been designed to serve. If, as Messick (1989) and Linn (1991) argue, social consequences are a key validity criterion, then classroom evidence needs to be legitimated. Furthermore, policy makers and educators need to create mechanisms for tapping into these legitimate sources of evidence.

Another logical step toward moving more accurate data into realms further removed from the classroom might be to reduce some of the $500 million spent annually on standardized multiple choice tests (Darling-Hammond and Snyder, 1992) and increase opportunities for teachers to develop their abilities to use, record, and report observation and performance data. Each increase in the capacity to generate accurate assessment data increases the probability of better decision making at all levels of education. In an era of limited resources, it seems wise to allocate those resources to obtain data that are as accurate, and, therefore, as useful as possible.

In addition, teachers in this study were adamant that when performance and observation data were shared through opportunities for teacher collaboration, the school, as a whole, began to recognize and utilize data more effectively. As a result, the teachers claimed, the school became more responsible for, and responsive to, students. Put simply, children began to read better when teachers shared accurate assessment data. The teachers were not promising that test results would improve -- only that children would become better readers.

Future Research

The findings of this pilot study pave the way for further research in this area. First, utilizing the basic methodological approach of this work, the sample must be enlarged to include more children in more contexts. We are undertaking the same process in several more schools, and with students in grades two through six. Second, further research needs to be carried out comparing different forms of reading samples. For instance, what specific analyses of student reading do teachers utilize and what kinds of evidence do they obtain? The third set of research questions suggested by this study revolves around the relationships between levels of the educational enterprise and sources of evidence. How can larger institutions use classroom-level data? Why are classroom data more accurate? How can, or should, performance analysis and observation data be aggregated? How does teacher collaboration move classroom evidence into the school program? We think these questions can best be approached with in-depth case studies of school sites, districts, and states that are promoting, through education and policy initiatives, the increased usage of observation and performance data.

References

Anderson, R. (1984). "Role of Reader's Schema in Comprehension, Learning, and Memory," in R. Anderson, J. Osborn, and R. Tierney (Eds.), ***Learning to Read in American Schools: Basal Readers and Content Texts***, 243-257. Hillsdale, NJ: Erlbaum.

Buros, O. (1977). "Fifty Years of Testing: Some Reminiscences, Criticisms, and Suggestions." ***Educational Researcher*** *6*(7), 9-15.

Bussis, A. (1982). "Burn it as the casket: Research, reading instruction, and children's learning of the first R." ***Phi Delta Kappan 64*** (December, 1982).

Chittenden, E. (1991). "Authentic Assessment, Evaluation, and documentation of student performance," in V. Perrone (Ed.), ***Expanding Student Assessment***, 22-31. Alexandria, VA: ASCD.

Darling-Hammond, L., and Snyder, J. (1992). "Reframing Accountability: Creating Learner-Centered Schools," in A. Lieberman (Ed.), ***The Changing Context of Teaching***, 11-36. Chicago: University of Chicago Press.

Linn, R. (1991). "Alternate Forms of Assessment: Implications for Measurement." Paper presented at the American Educational Research Association, Chicago, IL: April 6, 1991.

MacDonald, M., and Snyder, J. (1992). "Learner Centered Accountability in Two Elementary Schools." Paper presented at the 1992 Annual Meeting of the American Educational Research Association, San Francisco, CA, April 20-24, 1992.

Messick, S. (1989). "Meaning and Values in Test Validation: The Science and Ethics of Assessment." ***Educational Researcher*** *18*(2), 5-11.

TASA, Inc. (1988). ***Readability of Literature and Popular Titles***. Brewster, NY: TASA, Inc.