Research Series No. 53

A CLOSER LOOK AT STANDARDIZED TESTS

Donald Freeman, Therese Kuhs,

Lucy Knappen, and Andrew Porter

Published By

The Institute for Research on Teaching
252 Erickson Hall
Michigan State University
East Lansing, Michigan 48824

Printed and Distributed
by the
College of Education
Michigan State University

June 1979

This work is sponsored in part by the Institute for Research on Teaching, College of Education, Michigan State University. The Institute for Research on Teaching is funded primarily by the Teaching Division of the National Institute of Education, United States Department of Health, Education, and Welfare. The opinions expressed in this publication do not necessarily reflect the position, policy, or endorsement of the National Institute of Education. (Contract No. 400-76-0073)

Institute for Research on Teaching

The **Institute for Research on Teaching** was founded at Michigan State University in 1976 by the National Institute of Education. Following a nationwide competition in 1981, the NIE awarded a second contract to the IRT, extending work through 1984. Funding is also received from other agencies and foundations for individual research projects.

The IRT conducts major research projects aimed at improving classroom teaching, including studies of classroom management strategies, student socialization, the diagnosis and remediation of reading difficulties, and teacher education. IRT researchers are also examining the teaching of specific school subjects such as reading, writing, general mathematics, and science, and are seeking to understand how factors outside the classroom affect teacher decision making.

Researchers from such diverse disciplines as educational psychology, anthropology, sociology, and philosophy cooperate in conducting IRT research. They join forces with public school teachers, who work at the IRT as half-time collaborators in research, helping to design and plan studies, collect data, analyze and interpret results, and disseminate findings.

The IRT publishes research reports, occasional papers, conference proceedings, a newsletter for practitioners, and lists and catalogs of IRT publications. For more information, to receive a list or catalog, and/or to be placed on the IRT mailing list to receive the newsletter, please write to the IRT Editor, Institute for Research on Teaching, 252 Erickson Hall, Michigan State University, East Lansing, Michigan 48824-1034.

Abstract

The content covered by four commonly used standardized tests of elementary school mathematics was examined. Striking differences between tests were discovered. For this and other reasons, it is likely that there will be significant discrepancies between the content a teacher presents to students and the content tested by a standardized test. Such mismatches between content taught and content tested have negative effects on the utility of standardized tests for facilitating instruction and may also result in underestimates of student achievement.

A Closer Look at Standardized Tests

Donald Freeman, Therese Kuhs, Lucy Knappen, and Andrew Porter[1]

It is commonly argued that teachers should use scores from standardized tests to facilitate instruction. Specifically, teachers are encouraged to use standardized test results to evaluate student achievement on both a group and individual level, to identify students with learning problems, and to assess the effectiveness of instructional strategies which have been used. However, the use of standardized tests for any of these functions must be tempered by the teacher's knowledge of the extent to which the content of the test parallels the content of instruction.

Several factors contribute to differences among teachers in what is taught. These include use of different textbook series, the variety of mandated objectives across school districts, and the fact that teachers are sensitive to differences in ability among students. As will be shown later, there are also striking differences in content tested by different standardized tests. Because of the profound variety in content taught in schools and content tested, significant mismatches between the content of classroom instruction and the content of a standardized test are likely.

When a teacher elects to teach a topic that is not included on the standardized test used, there are two probable consequences. First,

---

[1]Donald Freeman is a senior researcher with IRT's Content Determinants group. Therese Kuhs is a research intern with that group. Lucy Knappen is the group's teacher collaborator, and Andrew Porter is the group coordinator.

student performance on the test will not reflect the level of learning of that topic. Second, the amount of time spent on instruction for that untested topic may reduce the amount of instructional time spent on other topics which are tested. Of course, achievement cannot be expected on topics which are tested but not taught. The net result of inconsistencies between what is taught and what is tested is an underestimate of student achievement for the curriculum as offered.

Our primary purpose is to assist teachers in determining the relationship between the content of mathematics instruction they provide and the content tested by the standardized measure they may be asked to use. First, we describe a taxonomy of mathematics which teachers could use in analyzing the content of their mathematics curriculum. Next, we give the results of a content analysis of the four most widely used standardized tests of elementary school mathematics: the Stanford Achievement Test (SAT), the Iowa Tests of Basic Skills (Iowa), the Metropolitan Achievement Tests (MAT), and the Comprehensive Tests of Basic Skills (CTBS).[2] Finally, significant discrepancies between content taught and content tested are outlined, and implications for classroom practice are discussed.

A Taxonomy of Mathematics Content

Test publishers generally recognize the need to consider test content in interpreting test results. They have, therefore, attempted to provide analyses of test content in the manuals which accompany each series. However, the descriptors provided tend to be very general (e.g., basic concepts) or very specific (e.g., matching figures with common fractions).

[2]Copyright dates of the tests which were analyzed are as follows: SAT -- 1973; Iowa -- 1971; MAT -- 1970; CTBS -- 1976.

Even when publishers provide descriptions of test content at a useful level of detail, teachers are provided little assistance in using the same classification scheme to analyze the content of their instruction.

To facilitate the description of content taught and content tested within a common framework, we have developed a taxonomy of elementary school mathematics. The taxonomy consists of a classification matrix with three dimensions: (1) mode of presentation (how questions are asked); (2) nature of material (type of numbers or mathematical terms used); and (3) operation (process which is required). The intersection of these three dimensions results in a classification matrix of 468 cells, where each cell represents a topic that a teacher may elect to cover or not to cover. We have developed a manual to assist teachers and others in using the taxonomy to classify the content of elementary school mathematics (Kuhs, Schmidt, Porter, Floden, Freeman, & Schwille, 1979).[3] Figure 1 provides the outline of the classification matrix.

Similarities and Differences in Content Covered

One of the most significant features of the taxonomy is its flexibility in describing content at different levels of detail. Each cell in the matrix represents a specific topic (e.g., solving story problems that require the multiplication of multiple digit whole numbers). At the same time, descriptions which correspond to the marginals of Figure 1 (rows or either of the two types of columns) represent general topics that might be included in the mathematics curriculum (e.g., fractions, division with remainder, story problems).

---

[3]Copies of the manual are available from the Institute for Research on Teaching. Ask for Research Series No. 4. (A modest fee is charged to cover the costs of reproduction.)

Figure 1: A Classification Matrix for Elementary School Mathematics

Classification of _____

By _____ Date _____

MODE OF PRESENTATION

Nature of the Material / Operation	Graphics, Figures, Tables or Physical Objects												Operation(s) Specified												Operation(s) Not Specified (Story Problems)											
	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12
Whole Numbers — single digits																																				
Whole Numbers — single digit and multiple digit																																				
Whole Numbers — multiple digits																																				
Fractions — single																																				
Fractions — multiple																																				
Decimals																																				
Percents																																				
[a] Alternate Number Systems																																				
Place Value																																				
Number Sentences																																				
[b] Algebraic																																				
[a] Essential Units of Measurement																																				
[a] Geometric Figures																																				
Other																																				

Operations

1. Add
2. Subtract w/o borrowing
3. Subtract with borrowing
4. Add or subtract fractions
5. Multiply
6. Divide w/o remainder
7. Divide with remainder
8. Combination
9. Grouping
10. Identify Equivalents
11. Identify Rule (Order)
12. Identify Terms*

* Be sure to *identify specifics* on attached page.

Table 1 reports the nature of material (rows in Figure 1) contained in the various grade levels of the four standardized tests. The categories included in Table 1 have been deliberately selected from the total analysis to illustrate striking differences in the content covered by the four tests. Variations among tests can be identified by an examination of the different tests which might be used at a given grade level.

Consider, for example, the tests which might be administered in the spring to a fourth-grade class. Over 63% of the items on the SAT involve whole numbers and only about 5% of this test deals with fractions. On the MAT, 53% of the items are whole number problems and 7% involve fractions; on the Iowa only 45% of the test deals with whole numbers, while 8% involve fractions. The CTBS presents an interesting option at the fourth-grade level. Use of the lower level test (grades 2.5 - 4.9) offers a test wherein 66% of the problems involve whole numbers and only 2% include fractions. However, if the higher level test is selected (grades 4.5 - 6.9), the emphasis on whole number problems is drastically reduced (37%), while problems dealing with fractions are increased to constitute 18% of the test. Examination of the varying emphasis on place value problems reveals further differences in the content of the tests for fourth-grade (MAT - 4.3%; SAT - 5.4%; Iowa - 9.2%; CTBS first level - 18.4%; CTBS second level - 3.1%).

Table 1 also shows changes in the nature of material across the sequence of levels of a given test. This analysis provides an index of the content emphasized on standardized tests at a particular grade level and highlights the increases and decreases of topic emphases which occur in successive grades.

An examination of the table reveals some interesting differences among tests. The decline in the percentage of single digit whole number

Table 1: Nature of Material: Item Distributions Across Levels of Standardized Tests*

MAT	(2.5-3.4)	(3.5-4.9)	(5.0-6.9)
Whole Numbers (total)	57.4	53.0	28.7
(single digit only)	26.9	14.8	4.3
Place Value	7.4	4.3	4.3
Fractions	2.8	7.0	19.1
Decimals	0.9	6.1	8.7
Percents	0	0	3.5

SAT	(2.5-3.4)	(3.5-4.4)	(4.5-5.4)	(5.5-6.9)
Whole Numbers (total)	64	60.4	63.4	47.5
(single digit only)	27	26	16.1	12.5
Place Value	6	10.4	5.4	5.0
Fractions	2	2	5.4	13.3
Decimals	0	2	4.5	4.2
Percents	0	0	0.8	2.5

CTBS	(2.5-4.9)	(4.5-6.9)
Whole Numbers (total)	66.3	36.7
(single digit only)	18.4	3.1
Place Value	5.1	3.1
Fractions	2.0	18.4
Decimals	2.0	14.3
Percents	0	1.0

Iowa	GR3	GR4	GR5
Whole numbers (total)	54.7	44.8	40.2
(single digits only)	17.3	12.6	12.4
Place Value	9.3	9.2	6.2
Fractions	4.0	8.0	17.5
Decimals	1.3	8.0	7.2
Percents	0	0	0

* Entries represent the percent of items on the entire test which encompass this material.

problems, for example, is more pronounced on the MAT and CTBS than it is on the other two tests. Place value problems are gradually deemphasized across grade levels on all of the tests except the CTBS. These and other differences depicted in Table 1 once again illustrate that *standardized tests at a given grade level do not all measure the same content.*

Consideration of the problems that require calculations on each test lends further insight into the content assessed. Table 2 reports the operations (numbered columns in Figure 1) that are tested on each of the four standardized measures and gives further descriptions of the whole number and fraction problems included on each test.

The limited assessment of division with a remainder (2% or less) may come as a surprise to fourth-grade teachers who emphasize the development of that skill. Also, for some grade levels, the MAT and CTBS contain a significant number of items that measure skills in adding, subtracting, multiplying, and dividing fractions. This emphasis contrasts sharply with the SAT, which does not include many items involving the addition or subtraction of fractions at any grade level. Notice that multiplication of whole numbers is most heavily emphasized at a particular level of each test (MAT - 7.8% at 3.5 to 4.9 level; SAT - 12.5% at 3.5 - 4.4 level; CTBS - 13.3% at 2.5 - 4.9 level; Iowa - 5.2% at Grade 5 level). Such comparisons suggest at least some differences in grade levels where initial content mastery is expected by the authors of each text.

Implications for Classroom Practice

In view of the striking differences in content covered by the four tests, it is very likely that the match between content taught in an established mathematics curriculum and content tested will be greater for some tests than others.

Table 2: Computation Problems: Item Distributions Across Levels of Standardized Tests*

Problem Description	Grade Levels			
MAT	(2.5-3.4)	(3.5-4.9)		(5.0-6.9)
Whole Numbers				
Add	25	16.5		4.3
Subtract w/out borrowing	13.9	5.2		1.7
Subtract with borrowing	0.9	6.1		2.6
Multiply	3.7	7.8		5.2
Divide w/out remainders	2.8	6.1		3.5
Divide with remainders	0	0.9		1.7
Fractions	(2.5-3.4)	(3.5-4.9)		(5.0-6.9)
Add or Subtract (like denominators)	0.9	2.6		6.0
Add or Subtract (unlike denominators)	0	0		0.9
Multiply	0.9	2.6		3.5
Divide	0	0		2.6
SAT	(2.5-3.4)	(3.5-4.4)	(4.5-5.4)	(5.5-6.9)
Whole Numbers				
Add	20.0	11.5	8.0	3.3
Subtract w/out borrowing	18.0	13.5	7.1	2.5
Subtract with borrowing	1.0	3.1	4.5	2.5
Multiply	5.0	12.5	9.8	8.3
Divide w/out remainders	0	0	11.6	4.2
Divide with remainders	0	0	0.9	1.7
Fractions				
Add or subtract (like denominators	0	0	0	0.8
Add or subtract	0	0	0	0.8
Multiply	0	0	2.7	1.7
Divide	0	0	0.9	0

(Table 2 cont.)

CTBS	(2.5-4.9)	(4.5-6.9)
Whole Numbers		
Add	12.2	6.1
Subtract w/out borrowing	11.2	2.0
Subtract with borrowing	3.0	4.1
Multiply	13.3	7.1
Divide w/out remainders	13.3	7.1
Divide with remainders	0	2.0
Fractions		
Add or subtract (like denominators)	0	5.1
Add or subtract (unlike denominators)	0	2.0
Multiply	0	3.0
Divide	0	2.0

IOWA	GR3	GR4	GR5
Whole Numbers			
Add	14.7	9.2	4.1
Subtract w/out borrowing	10.7	5.2	4.1
Subtract with borrowing	5.3	6.9	5.2
Multiply	4.0	4.6	5.2
Divide w/out remainders	1.3	1.1	0
Divide with remainders	0	1.1	1.0
Fractions	GR3	GR4	GR5
Add or subtract (like denominators)	0	2.3	3.0
Add or subtract	1.3	1.1	1.0
Multiply	0	1.1	4.1
Divide	0	0	0

* Entries represent the percent of items on the entire test which encompass this material.

The best way to assure a reasonable match between test content and instructional content is to select tests carefully. If a school district has a well established mathematics curriculum, factors of cost, format, and general attractiveness should be of secondary importance to considerations of content in test selection. We recognize, however, that a district's flexibility in selecting or changing the test used is limited by budget considerations and by the fact that tests are typically purchased as packages rather than as individual tests of a single subject matter area. Thus some discrepancies between test content and instructional content will occur, even when content is considered in test selection.

A second approach to achieving a reasonable match between test content and instructional content is to change the curriculum. In our view, it would be irresponsible to alter the content of instruction simply to attain a one-to-one correspondence with test content. Rather, we recommend that the identification of a clear discrepancy between content tested and content taught in one or more specific areas should prompt a thoughtful reappraisal of how that area is treated in the mathematics curriculum. When teachers can provide a solid rationale for stressing content that receives little attention on the test, current practice in that area should be maintained. The same is true for those areas in the curriculum given limited attention, but stressed on the exam. However, when there is no solid basis for continuing the current level of content emphasis, teachers might modify what they teach to achieve a clearer match between content taught and content tested.

Although Tables 1 and 2 have been constructed primarily to identify discrepancies between content taught and content tested, the information presented in these tables has other implications for classroom practice.

Perhaps the most important of these is the use of standardized tests to diagnose individual strengths and weaknesses in specific content areas.

Because the standardized tests reviewed in this paper have been deliberately designed to provide "general" measures of student achievement, the number of items testing specific content areas (e.g., subtraction with borrowing) is often too limited to provide reliable measures of individual achievement in that area.[4] Attaining a correct answer on the one item which assesses division with remainders hardly demonstrates mastery of that skill. An incorrect answer could also be misleading. By examining the data in Tables 1 and 2, teachers should be able to identify the specific content areas which have been emphasized on the test administered in their district. Whereas judgments about the achievement of individual students in these specific areas is reasonable, judgments in other specific areas are not. This analysis should, therefore, suggest the upper limits of each test in diagnosing individual strengths and weaknesses in specific content areas.

Data in Tables 1 and 2 also provide for more meaningful interpretations of grade equivalent scores.[5] A common misinterpretation is to conclude, for example, that a third-grade student with a grade equivalent of 6.2 is capable of doing sixth-grade work. From examination of Tables 1 and 2, it should be obvious that a child who scores at a 6.2 grade equivalent on a test administered in the third-grade has merely demonstrated a high level of achievement on third-grade content. There is no assurance that this child can solve such problems as subtraction with regrouping or

---

[4] For reasons beyond the scope of this paper, the same limitation does not prevail in regard to measuring group achievement.

[5] Scores on an exam may be compared to real or estimated average scores of students at grade levels other than those for which the test was intended. A third-grade student with a grade equivalent score of 6.2, for example, has a score equal to the average score on the third-grade test for students in the second month of sixth-grade.

division with remainders, or perform other operations which are typically tested in the sixth-grade. Teachers who wish to determine if their more capable students can perform skills typical of higher grades should administer more advanced levels of the test where these skills are tested. At the other extreme, a third-grade youngster who has a grade equivalent score of 1.2 may, nevertheless, have mastered second-grade content. His low score may merely reflect problems in performing those third-grade skills which are emphasized on the test.

As this discussion suggests, there are limitations to using standardized tests for diagnosing the strengths and weaknesses of individual students in specific content areas. This limitation may be further confounded by misinterpretations of grade equivalent scores. Teachers should, therefore, collect additional information prior to determining what enrichment experiences or remedial activities are best suited for those students who score at either extreme on a standardized test of achievement. A series of comparatively short, self-designed tests of specific skills should always supplement standardized test scores in designing these activities for individual students.

Mismatches between what is taught and what is tested also have implications for evaluating instruction. If a teacher's students demonstrate gain from fall to spring on a standardized test, the gain may be uniform across all topics tested or it may be isolated to just a few areas of mathematics (Porter, Schmidt, Floden, & Freeman, 1978). Thus, to diagnose instruction, student gain (even on a subtest) must be broken down into gain for specific topics. For any topic on which students do not demonstrate gain, there are at least two explanations which fall within the teacher's sphere of responsibilities.[6] As we said

---

[6]Using student gains in achievement as an index of the quality of instruction must, of course, also take into account the aptitudes of the students receiving instruction.

previously, if the topic was not included within the teacher's goals for instruction, the teacher should not expect student gain. What remains to be answered is, under these conditions, whether or not the topic should have been included in the goals for instruction. One way in which the quality of instruction may be poor is that not enough content or the wrong content was taught. A second explanation for lack of student gain on a topic is that the instructional strategies were lacking. This distinction between content taught and strategies used seems important for evaluating instruction, since the implief remedies are quite different for the two.

Summary

There are striking differences in the content covered by four commonly used standardized tests of elementary school mathematics. For this and other reasons, it is likely that there will be significant discrepancies between the content a teacher presents to students and the content which is being tested on the standardized test administered. These mismatches between content taught and content tested have negative effects on the utility of standardized tests for facilitating instruction. They may also result in underestimates of student achievement. A conscious effort should be made to select tests which best match the existing mathematics curriculum and/or to reappraise the curriculum in terms of its match with test content. The match between content taught and content tested is a crucial context for using tests to diagnose student strengths and weaknesses as well as for assessing the strengths and weaknesses of instruction provided.

References

Porter, A.C., Schmidt, W.H., Floden, R.E., & Freeman, D.J. Practical significance in program evaluation. American Educational Research Journal, 1978, 15(4),

Kuhs, T., Schmidt, W., Porter, A., Floden, R., Freeman, D., & Schwille, J. A taxonomy for classifying elementary school mathematics content (Res. Ser. No. 4). East Lansing, Mich.: Institute for Research on Teaching, Michigan State University, 1979.

$1.50

Research Series No. 53

A CLOSER LOOK AT STANDARDIZED TESTS

Donald Freeman, Therese Kuhs,

Lucy Knappen, and Andrew Porter

Published By

The Institute for Research on Teaching
252 Erickson Hall
Michigan State University
East Lansing, Michigan 48824

June 1979

This work is sponsored in part by the Institute for Research on Teaching, College of Education, Michigan State University. The Institute for Research on Teaching is funded primarily by the Teaching Division of the National Institute of Education, United States Department of Health, Education, and Welfare. The opinions expressed in this publication do not necessarily reflect the position, policy, or endorsement of the National Institute of Education. (Contract No. 400-76-0073)

Institute for Research on Teaching

The **Institute for Research on Teaching** was founded at Michigan State University in 1976 by the National Institute of Education. Following a nationwide competition in 1981, the NIE awarded a second contract to the IRT, extending work through 1984. Funding is also received from other agencies and foundations for individual research projects.

The IRT conducts major research projects aimed at improving classroom teaching, including studies of classroom management strategies, student socialization, the diagnosis and remediation of reading difficulties, and teacher education. IRT researchers are also examining the teaching of specific school subjects such as reading, writing, general mathematics, and science, and are seeking to understand how factors outside the classroom affect teacher decision making.

Researchers from such diverse disciplines as educational psychology, anthropology, sociology, and philosophy cooperate in conducting IRT research. They join forces with public school teachers, who work at the IRT as half-time collaborators in research, helping to design and plan studies, collect data, analyze and interpret results, and disseminate findings.

The IRT publishes research reports, occasional papers, conference proceedings, a newsletter for practitioners, and lists and catalogs of IRT publications. For more information, to receive a list or catalog, and/or to be placed on the IRT mailing list to receive the newsletter, please write to the IRT Editor, Institute for Research on Teaching, 252 Erickson Hall, Michigan State University, East Lansing, Michigan 48824-1034.