DOCUMENT RESUME

ED 415 238 TM 027 840

AUTHOR	Shepard, Lorrie A.
TITLE	Measuring Achievement: What Does It Mean To Test for Robust Understanding? William H. Angoff Memorial Lecture Series.
INSTITUTION	Educational Testing Service, Princeton, NJ. Policy Information Center.
PUB DATE	1997-00-00
NOTE	32p.; Paper presented at the annual William H. Angoff Memorial Lecture Educational Testing Service, (3rd, Princeton, NJ, September 19, 1996).
PUB TYPE	Opinion Papers (120) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Academic Achievement; *Achievement Tests; Comprehension; Educational Assessment; Elementary Secondary Education; Knowledge Level; *Measurement Techniques; *Performance Based Assessment; Standardized Tests; *Student Evaluation; *Test Bias; Test Format
IDENTIFIERS	*Teaching to the Test

ABSTRACT

In testing, educators have two competing intentions. They want to be fair in allowing students to demonstrate what they know without creating testing conventions and formats that let students pretend to know. Teaching-the-test literature shows that test scores can be inflated, meaning that they can go up without a generalized increase in knowledge. Students can appear to know what they really do not know, as can be illustrated by comparing results from the National Assessment of Educational Progress with those from more familiar standardized tests. Performance assessments are intended to overcome this problem. They are also intended to overcome the format effects of multiple-choice tests that have distorted instruction and allowed students to pretend to know more than they actually do. Even with performance assessments, students may rely on familiar, rote routines, and so pretend to know more than they really do. As a videotape presented as part of the discussion illustrates, asking in different ways is the way to assure that students really know what they are doing, and that their understandings generalize across contexts. Those who are concerned with test bias explore the opposite side of the coin--that students really know, but are not able to show their knowledge and abilities because of some aspect of the test. These two perspectives can be reconciled by careful and thoughtful assessment that approaches student knowledge in different contexts. (Contains 16 figures and 12 references.) (SLD)

MEMORIAL LECTURE Series

MEASURING ACHIEVEMENT:

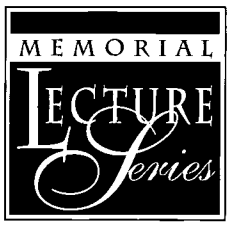
William H. Angoff
1919 - 1993

William H. Angoff was a distinguished research scientist at ETS for more than forty years. During that time, he made many major contributions to educational measurement and authored some of the classic publications on psychometrics, including the definitive text "Scales, Norms, and Equivalent Scores," which appeared in Robert L. Thorndike's Educational Measurement. Dr. Angoff was noted not only for his commitment to the highest technical standards but also for his rare ability to make complex issues widely accessible.

The Memorial Lecture Series established in his name in 1994 honors Dr. Angoff's legacy by encouraging and supporting the discussion of public interest issues related to educational measurement. The annual lectures are jointly sponsored by ETS and an endowment fund that was established in Dr. Angoff's memory.

The William H. Angoff Lecture Series reports are published by the Policy Information Center, which was established by the ETS Board of Trustees in 1987 and charged with serving as an influential and balanced voice in American education.

Measuring Achievement:
What Does it Mean to Test for Robust Understandings?



The third annual William H. Angoff Memorial Lecture was presented at Educational Testing Service, Princeton, New Jersey, on September 19, 1996.

Lorrie A. Shepard
University of Colorado at Boulder

Policy Information Center
Princeton, NJ 08541-0001

Preface

In the third William H. Angoff Memorial Lecture, Dr. Lorrie Shepard asks: "What does it mean to test for what students really know?" When they give the correct answer on a test, "Do they really know it?"

A lot of attention has been given to questions of test "fairness." Do tests permit students to fully demonstrate what they know and can do? Do students know *more* than what is indicated by tests results? Here, Dr. Shepard addresses the opposite circumstance — where test results indicate students know and understand things that they really do not.

Students may be able to demonstrate their knowledge — or get the right answer — in one context, but then be unable to do it in another context. They often have "fragile understandings." Dr. Shepard argues for the importance of "robust understandings," and guides us toward improvements in teaching and assessment.

I would like to thank the following individuals for their contribution to this publication: Ric Bruce designed the report; Carla Cooper provided desktop publishing services; Jim Chewning coordinated production; and Shilpi Niyogi was the editor.

Paul Barton
Director, Policy Information Center

Preamble

I first want to thank you very much for this invitation and for the chance to be here to honor Bill Angoff. I welcome the opportunity to come to ETS and to talk with all of my friends and fellow researchers. Although a distant colleague whom I saw only occasionally, Bill was someone who was very special to me. I think back to my first meeting with him at Johns Hopkins, where a symposium was held that eventually led to Ron Berk's book on test bias.[1] Bill presented thoughtful but pointed criticisms of statistical bias indices that he himself had been using; and, in that same talk, I believe he was the first to suggest that "item-bias" methods should instead be called only "item-discrepancy" methods — a recommendation that researchers in the field now follow as a matter of course. Subsequently, I remember that Bill was so remarkably enthusiastic about the early work that I was doing on measurement and the identification of learning disabilities. He had a way of sort of congratulating you for thinking about hard problems. I always felt that he doted on all of us, his colleagues in the measurement community, even when he was quite young to be doting. So I appreciate very much being here.

[1] Berk, R. A. (1982). *Handbook of Methods for Detecting Test Bias.* Baltimore, MD: Johns Hopkins University Press.

Introduction

The meaning of my title, "What Does it Mean to Test for Robust Understandings?" might not be apparent until I have had a chance to give you some examples of "robust" understandings. Alternative ways of entitling today's presentation, "What Does it Mean to Test for What Students Really Know?" or "Do They Really Know It?" anticipate, or at least forewarn you of two competing intentions. We want to be fair in allowing students to demonstrate what they know without creating testing conventions and formats that let them pretend to know.

When I have done research on test bias, I've said, "Ah, there's something about the measurement that is misleading and is preventing us from understanding what students really do know." When I'm in this role of examining the potential bias in tests, I become the advocate of the student and test taker. The student knows the subject matter but the test is unfair. It occurred to me recently that I was taking exactly the opposite stance when I conduct research on the effects of teaching the test. In this context, we have documented how students could appear to know, when they capitalized on the measurement format that was just like the test preparation they had had, but could not demonstrate that same knowledge if asked in even a slightly different way (Koretz, Linn, Dunbar, & Shepard, 1991). I'm going to talk first about the teaching-the-test literature. I will go pretty far down that path, giving you examples not only of what students can't do when they've been prepared for a specific format but also some examples from performance assessments illustrating how we've tried to redress the problems with teaching for the test. Then I will turn to the other side of the coin and use some examples, in fact, of ETS research on test bias. Finally, in the last part of my talk I'd like to consider what these contradictions mean for classroom assessment and what they mean for large scale assessment and possible attempts to model what's going on. "Do they know it, or don't they?" How can we know? That's the measurement question.

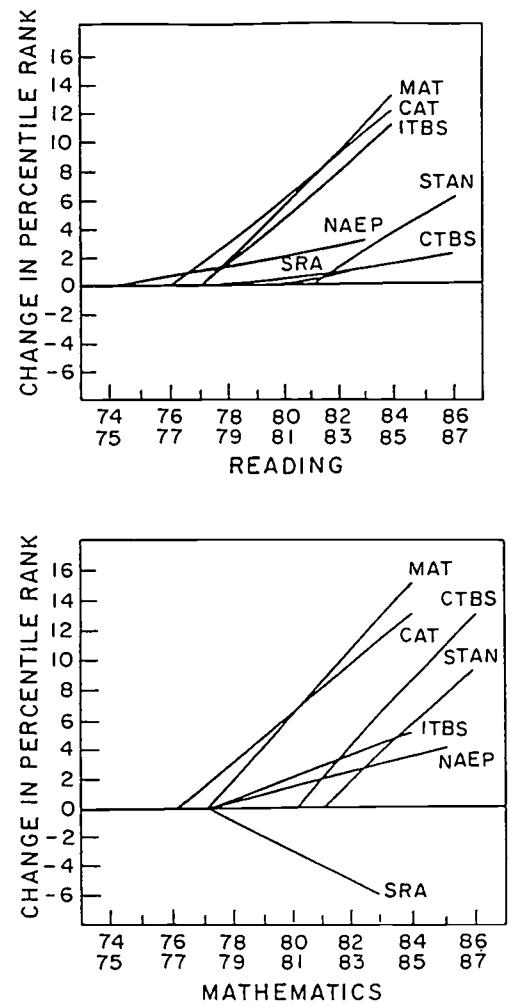
Research on Teaching-the-Test

From the teaching-the-test literature, we know that test scores can be inflated, meaning that test scores can go up without there being a generalized increase in knowledge. Students can appear to know when they don't really know. I'd like to take a minute and show you at least one example of the kind of data that leads to this conclusion because occasionally, at meetings of the American Educational Research Association or in reviews of journal manuscripts, such statements are written off as the unwarranted beliefs of standardized-test bashers. So, what kind of evidence do we have that test scores can be inflated without there being a generalized increase in knowledge?" Bob Linn (1995) presented some of this data in the first William H. Angoff Memorial Lecture.

In 1987, a West Virginia physician, J. J. Cannell, reported the scandalous finding that all 50 states claimed that their students were achieving above the national average. Linn and colleagues at the University of Colorado (Linn, Graue, & Sanders, 1990) conducted a systematic national study intended to replicate Cannell's findings using a representative sample of districts as well as the 50 states. They confirmed that, yes, indeed, nearly all states and a disproportionate number of districts across the nation were reporting achievement averages above the national norm. This pattern was especially pronounced in mathematics and in the elementary grades. Linn and his colleagues also conducted analyses to examine whether these glowing reports were real or spurious. For example, such findings could represent true gains in student achievement (although still false claims about the relative standing of states and districts) if the problem was "old norms." This would be the case if achievement in the entire nation was rising but test results were still being reported in relation to an outdated national average.

Linn, et al. (1990) used data from the National Assessment of Educational Progress (NAEP) to evaluate the interpretation that achievement gains were real versus other possible explanations such as test familiarity. Figure 1 shows estimated changes in percentile rank for third graders from late seventies norms to late eighties norms on six popular standardized tests.

Figure 1 - Estimated Change at the Median in National Percentile Ranks of Achievement Test Scores at Grade 3 (NAEP, Age 9)

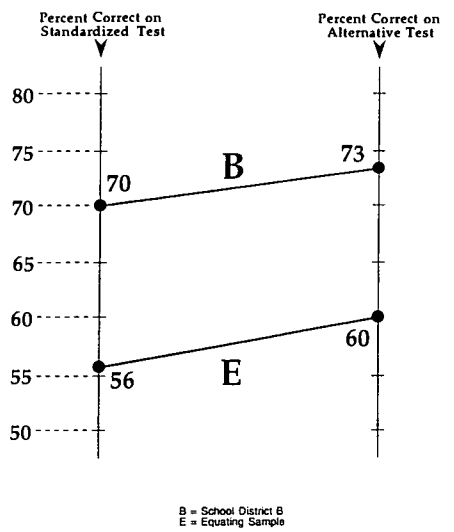


Source: Linn, Graue, & Sanders, 1990.

We call these the "pick-up-sticks" graphs. Also shown is change in performance for 9 year olds on NAEP over the same time period. The important comparison is the NAEP line, because it is a secure test and unlikely to have been taught to. Down at the bottom, in mathematics, SRA looks like it's declining over the period; in reading, two tests show achievement rising less steeply than NAEP. But the dominant trend in both graphs is steeply rising pick-up sticks suggesting achievement gains on locally administered standardized tests that were much greater than occurred on NAEP. With NAEP as the benchmark, we think there is some evidence of inflated achievement gains on most of the other tests. The students don't really know, when measured on the NAEP, what they appear to know on more familiar, locally administered, standardized tests. Linn, et al. (1990) and Linn (1995) presented other data that support this interpretation as well. "Zig-zag" graphs also show the effect of test familiarity on test score gains. This is the frequently observed pattern whereby test scores rise steadily over a period of years until a new form of the test is introduced where upon

Figure 2 - Mean Percent Correct on a Standardized-Test Item and Alternative-Test Item in a High-Stakes District (B) and Equating Sample (E)

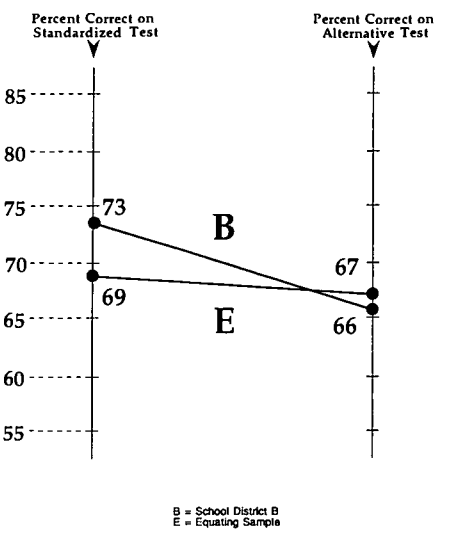
Standardized Test		Alternative Test
21	A. 21	23
x4	B. 25	x3
	C. 48	
	D. 84	



Source: Flexer, 1991.

Figure 3 - Mean Percent Correct on a Standardized-Test Item and Alternative-Test Item in a High-Stakes District (B) and Equating Sample (E)

Standardized Test		Alternative Test
	A. 721	
764	B. 731	Add 764 and 67.
+67	C. 831	
	D. 830	



Source: Flexer, 1991.

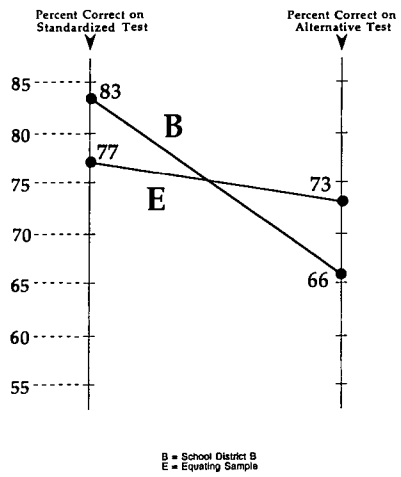
the scores drop precipitously and then rise again slowly to the previous level.

Other evidence that levels of achievement reported on high-stakes accountability tests were not "real" was gathered by Koretz, et al. (1991) in an experimental study. In large school districts selected because of accountability pressure focused on raising test scores, random subsamples of students were administered unfamiliar standardized tests and alternative tests constructed item-by-item to match the district-administered test but using a slightly more open-ended format. Student performance dropped as much as a half standard deviation on the unfamiliar tests suggesting that students did not really know all that they appeared to know on the publicly reported measures. The following examples, as reported by Flexer (1991), are item level analyses of the Koretz, et al. test comparisons.

Figure 2 shows you a multiplication item from District B's regularly administered standardized test along with the corresponding alternative test item. As you can see the items are highly similar except for the difference in response mode. In this case, there was

Figure 4 - Mean Percent Correct on a Standardized-Test Item and Alternative-Test Item in a High-Stakes District (B) and Equating Sample (E)

Standardized Test		Alternative Test
	A. 63	
87	B. 53	Subtract 24 from 87.
-24	C. 64	
	D. 62	



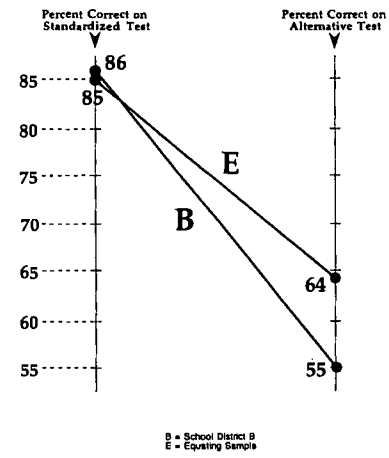
Source: Flexer, 1991.

Figure 5 - Mean Percent Correct on a Standardized-Test Item and Alternative-Test Item in a High-Stakes District (B) and Equating Sample (E)

Standardized Test		Alternative Test
4	A. 9	Which number sentence goes with
x3	B. 12	
	C. 15	
	D. 18	

x	x	x	x
x	x	x	x
x	x	x	x

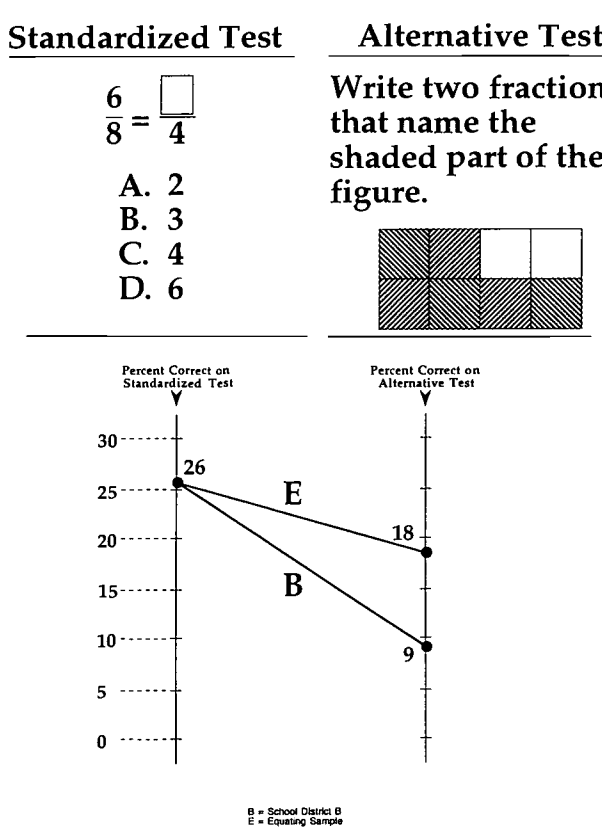
A. 3 x 4 =
B. 3 + 4 =
C. 3 x 12 =



Source: Flexer, 1991.

also no difference in student performance. The graph at the bottom of Figure 2 shows the percent correct earned on each item in District B, the high-stakes district where teaching-to-the-test was likely to occur, and in the Equating sample. Because we knew from past research that open-ended items were likely to be more difficult than multiple-choice questions, the test score comparisons reported in Koretz, et al. (1991) required that parallel sets of tests be equated statistically as well as by matching the content of items. Randomly equivalent equating samples were drawn from districts where both the standardized test and the alternative test were unfamiliar.

Figure 6 - Mean Percent Correct on a Standardized-Test Item and Alternative-Test Item in a High-Stakes District (B) and Equating Sample (E)



Source: Flexer, 1991.

Figure 3 shows another pair of items with a slightly greater difference in format. By changing from a vertical addition problem (with multiple-choice answers) to a horizontal problem, performance decreased slightly in the equating sample but declined by a greater amount in the high-stakes district. If these lines could be assumed to be perfectly reliable, the difference in the decline could be taken as the teaching-to-the-test effect. In contrast, lack of format effect would reflect generalized knowledge or "robust" mastery of the skill. These effects, showing differential losses for students in the high-stakes district, get slightly bigger over the next three examples. Figure 4 is just the subtraction version of the vertical and horizontal format change. Figures 5 and 6 illustrate a slightly greater change in format, using in each case pictorial representations of the multiplication and fraction problems. Notice that these unfamiliar but conceptually straightforward questions show a remarkable drop-off in performance for both groups of students, suggesting that these items are not interchangeable to the learner who is developing understanding, even if they appear to the adult test maker to tap the same conceptual knowledge. The teaching-to-the-test interaction effect is also apparent suggesting that practice on only one type of format can worsen the failure to generalize.

At the University of Colorado at Boulder, I teach the assessment component of subject-matter methods courses required as part of the teacher licensure program. I use examples like these when I work with students preparing to be teachers to try to convince them that they need to ask their students things

in different ways. This is a difficult concept to get across because would-be teachers and experienced teachers alike are inclined to believe that students have mastered a concept if they can perform a task as instructed. The first response of nearly every classroom observer is, "If they know it, they know it." I say, "Well, do they really know it, if we get this much difference in performance with such a subtle change in item format?"

In addition to the measurement problem created by spurious test score gains, the "teaching-the-test" literature has also documented the effect of high-stakes pressure on curriculum and instructional practices. In schools and districts under fire to raise test scores, elementary teachers reported devoting less time to untested subjects such as science and social studies (Shepard, 1991). Moreover, even basic skills instruction in reading and mathematics became distorted as textbook problems and teacher-made worksheets conformed narrowly to the format of multiple-choice tests. Although textbooks in the 1990s have already made a change for the better in response to the National Council of Teachers of Mathematics (NCTM) standards and other standards efforts, a look back at pages from textbooks from the 1980s make the case that both first-time presentation of content as well as end-of-chapter tests closely resembled skills and formats from standardized tests. For example, because it appeared on standardized tests, second and third graders spent significant amounts of time learning to write out, from numerals to words, the narrative translation of 3,467 rather than doing problems to check on their understanding of what the numeral means (not what it "says"). Not surprisingly, some children who can write out three thousand, four hundred, sixty seven do not have the place understanding to be able to subtract 26 from 3,007.

Performance Assessments

Without going into all the rhetorical claims about the benefits of performance assessments, I'd like to show you some examples from the classroom-level performance assessment project that we carried out in third-grade classrooms in three different Colorado schools for one school year. (The study also included three control schools.) These performance tasks with student responses are actually from the Maryland mathematics assessment, which we used as an independent outcome measure for the project. In showing you these examples, I have several purposes in mind. The first is simply to illustrate what I mean by open-ended tasks. A formal part of our research project, in fact, involved showing samples of performance assessments as well as standardized questions to parents for their review and comment (Shepard & Bliem, 1995). We learned that letting parents see actual performance items immediately dissipated their concerns. Whereas hearing about performance assessments had led them to believe that they would be less rigorous, as soon as they saw these kinds of problems, parents were satisfied that they reflected the kinds of skills and knowledge they wanted students to have.

My second purpose, and the one centrally important to the main point of this talk, is to illustrate how performance tasks are intended to prevent the format effects of multiple-choice tests that have distorted instruction and in turn enabled students to pretend to know. Open-ended tasks are more challenging and directly reflect desired understandings and applications. Therefore, it is less likely that students could do these items correctly and still not understand the underlying concepts. (Note in the next section, however, I discuss how even performance assessments may allow students to rely on familiar routines and give the appearance of mastery.) I would also like to note, at least by example, the greater diagnostic value of this type of assessment — how much greater insight we can gain about children's thinking than from simple, right-wrong problems. And I will at least give reference to project data suggesting that experience and "practice" with these kinds of instructional and assessment tasks improved student learning in a way that generalized to performance on the independent outcome measure.

In the lemonade problem in Figure 7, the kids have to use the number of cups in a pitcher to figure out how many pitchers are needed to get 46 cups of lemonade. The first student here filled out the table correctly and then said, "I looked at the pattern and saw that there was not a 46, so I took 48, so there was also some for my friend and I." We, of course, scored all of these booklets by hand, and I can assure you that a lot of students had plans for those extra two glasses of lemonade. At the bottom is another student's explanation, "From pitchers #11 to 12 it went 44, 48 cups so I just put 11 1/2." Many of the third graders in our study could answer this kind of problem. It shows you what open-ended tasks can do. It looks like pretty standard curriculum except that students are having to explain their answers, and it's more of an application problem, not straight multiplication, to be sure. In fact, some kids wouldn't know this as multiplication, at least it's not their typical way of learning multiplication.

More complete examples and data from this project are reported in Shepard, Flexer, Hiebert, Marion, Mayfield, and Weston (1996). Overall there was evidence of a small but interpretable positive gain showing that, when students had experience with these kinds

Figure 7 - Sample Student Responses on Maryland Mathematics Assessment Problem Set Two (Lemonade Step 4) Illustrating Correct Answers and Explanations

STEP 4

Now you want to know how many pitchers you will need for 46 cups of lemonade. You can see from the table below that a one-quart pitcher will hold 4 cups, and 2 one-quart pitchers will hold 8 cups. Continue the pattern in both rows of the table until you find the number of pitchers needed to hold 46 cups of lemonade.

Pitchers	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Cups	4	8	12	16	20	24	28	32	36	40	44	48	52	56	60

How many one-quart pitchers will you need for 46 cups of lemonade? Write your answer on the line below.

12

Explain how you got your answer. Write on the lines below.

I looked at the patteren and saw that there was not a 46. so I took 48 so there is also some for my friend and I.

Explain how you got your answer. Write on the lines below.

From pichers #11 to 12 it when t 44 48 cups so I just put 11½

Source: Shepard, Flexer, Hiebert, Marion, Mayfield, & Weston, 1996.

of problems, they learned to do things that third graders the year before had been unable to do. Keep in mind that we did not give students exactly this table or a closely parallel version of it. But the kinds of instructional tasks that were introduced did give students a lot more practice over the course of a year in thinking about patterns and in thinking about functions; in addition, the kind of multiplication work that they did, using Marilyn Burns's circles and stars and thinking about multiplication as successive addition, probably helped kids do better on problems like this without our teaching to the test in the narrow sense.

More examples of student work are shown in Figure 8. These are examples of wrong answers, but show you nonetheless how students could be gaining in understanding as a result of the project because they are able to explain their answers and give explanations that are mathematical,

"I counted by four's," instead of the prevalent answer the year before, "I thought in my head and got the answer." In our qualitative analysis, this type of wrong answer occurred frequently where students could extend the table and give a mathematically based explanation but could not use the table to answer the question correctly. "I counted by four's, which is 60, I went into the ones which is 15." The other responses are from different students, "On the cups as you go along you count four more each time." "First I saw that the (y) were counting by four, so I counted by fours until there was no more room and got the answer 57."

The data in Figure 9 are from a matched pair of classrooms from low socioeconomic participating and control schools. In the low socioeconomic participating classroom, there was no gain at the top end of the scale. But what you

Figure 8 - Sample Student Responses on Maryland Mathematics Assessment Problem Set Two (Lemonade Step 4) Illustrating Wrong Answers, But Table is Completed Correctly and Explanation Describes Pattern

Pitchers	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Cups	4	8	12	16	20	24	28	32	36	40	44	48	52	56	60

How many one-quart pitchers will you need for 46 cups of lemonade? Write your answer on the line below.

15

Explain how you got your answer. Write on the lines below.

I counted by fours which is 60 the I went in the ones which is 15.

Explain how you got your answer. Write on the lines below.

On the cups as you go along you count four more each time

Explain how you got your answer. Write on the lines below.

first I saw that the where counting by 4ors So I counted by fours until there was no rome and got the answer 57.

Source: Shepard, Flexer, Hiebert, Marion, Mayfield, & Weston, 1996.

Figure 9 - Comparison of 1992 and 1993 Student Responses on Maryland Mathematics Assessment Problem Set Two (Lemonade Step 4) from the Classrooms with the Greatest Gains in the Low Socioeconomic Participating and Control Schools

	Participating		Control	
	1992	1993	1992	1993
I. Extends table, Answers correctly, Explains (explains either pattern or point in chart).	13%	13%	31%	19%
II. Extends table, Answers correctly, Inadequate explanation.	4%	0	0	12%
III. No answer but stops table at right place, Explanation describes pattern.	0	0	0	0
IV. Extends table, Wrong answer (60, 15, 11, other), Explanation describes pattern.	0	42%	8%	35%
V. Extends table, Wrong answer (60, 15, 11, other), Inadequate explanation.	17%	29%	8%	35%
VI. Cannot extend table.	63%	8%	46%	31%
VII. Blank	4%	8%	0	0

Source: Shepard, Flexer, Hiebert, Marion, Mayfield, & Weston, 1996.

Figure 10 - Comparison of 1992 and 1993 Student Responses on Maryland Mathematics Assessment Problem Set Two (Lemonade Step 4) from the Classrooms with the Greatest Gains in the Low Socioeconomic Participating and Control Schools

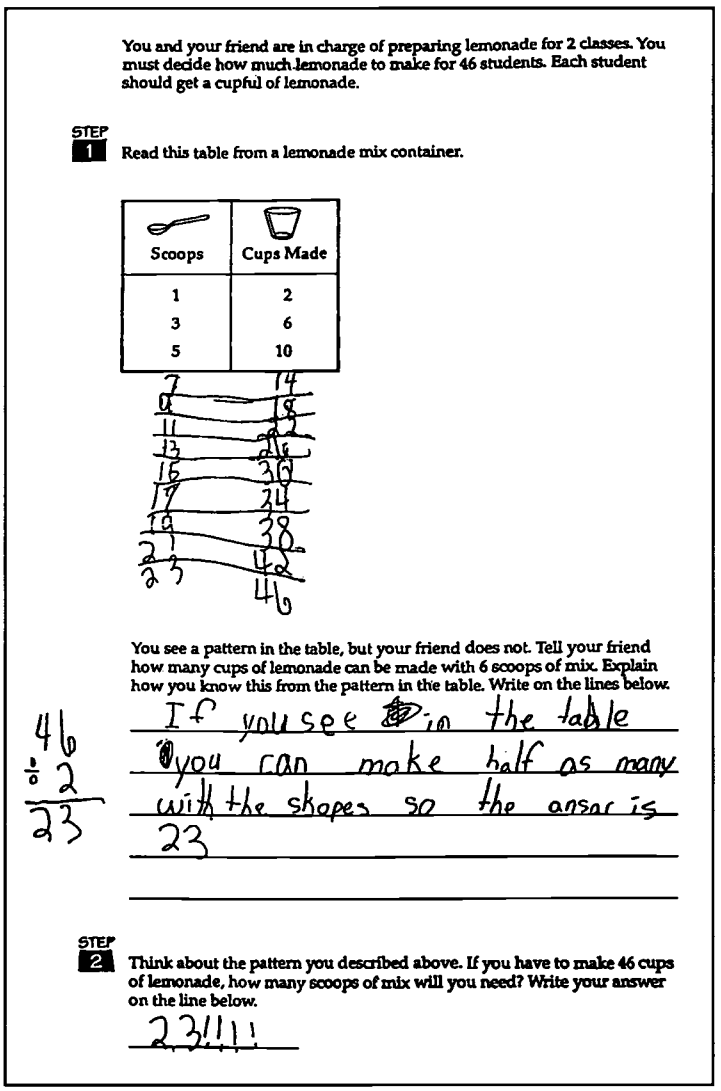
	Participating		Control	
	1992	1993	1992	1993
I. Extends table, Answers correctly, Explains (explains either pattern or point in chart).	19%	43%	56%	43%
II. Extends table, Answers correctly, Inadequate explanation.	8%	0	0	4%
III. No answer but stops table at right place, Explanation describes pattern.	0	5%	0	0
IV. Extends table, Wrong answer (60, 15, 11, other), Explanation describes pattern.	12%	29%	39%	9%
V. Extends table, Wrong answer (60, 15, 11, other), Inadequate explanation.	31%	9%	0	30%
VI. Cannot extend table.	31%	9%	6%	13%
VII. Blank	0	5%	0	0%

Source: Shepard, Flexer, Hiebert, Marion, Mayfield, & Weston, 1996.

see from the baseline data in 1992 to the end of the project year is a large boost in the number of children who could extend the table and give explanations like those shown in Figure 8. In that same teacher's classroom the year before, 67% of the students left the whole problem blank or gave meaningless answers; only 16% were similarly unable to respond after the project year. Figure 10 provides similar data for a matched pair of classrooms in high socioeconomic schools. Here the boost occurred at the top end. That is, in the baseline year, kids were already further along in how many of them could do all or part of the problem; what you see after the project year is an increase from 19% getting all parts of the problem right to 43% getting it all the way right, with additional gains in the number who could extend the table and explain their answers. The whole class distribution was shifted upwards. Note that these comparisons with control schools (rather than baseline year data) are a little bit hard to follow because there is a general pattern of declining performance for all the schools in the district on both their standardized tests and the alternative assessments that we administered. The small positive gains made by teachers in participating classrooms have slightly greater import against a backdrop of declining scores.

These last examples show you a different segment of the lemonade series of tasks. In Figure 11 the problem reads, "You and your friend are in charge of preparing lemonade for two classes. You must decide how much lemonade to make for 46 students. Each student should get a cupful of lemonade." Then, after the table, "You see a pattern in the table, but your friend does not. Tell your friend how many cups of lemonade

Figure 11 - Sample Student Responses on Maryland Mathematics Assessment Problem Set One (Lemonade Step 1-2) Illustrating A Correct Answer and Explanation



You and your friend are in charge of preparing lemonade for 2 classes. You must decide how much lemonade to make for 46 students. Each student should get a cupful of lemonade.

STEP 1 Read this table from a lemonade mix container.

Scoops	Cups Made
1	2
3	6
5	10

7 14
9 18
11 22
13 26
15 30
17 34
19 38
21 42
23 46

You see a pattern in the table, but your friend does not. Tell your friend how many cups of lemonade can be made with 6 scoops of mix. Explain how you know this from the pattern in the table. Write on the lines below.

46 ÷ 2 = 23

If you see in the table you can make half as many with the skopes so the ansar is 23

STEP 2 Think about the pattern you described above. If you have to make 46 cups of lemonade, how many scoops of mix will you need? Write your answer on the line below.

23!!!!

Source: Shepard, Flexer, Hiebert, Marion, Mayfield, & Weston, 1996.

can be made with 6 scoops of mix. Explain how you know this from the pattern in the table. Write on the lines below." The student response shown in Figure 11 is a little bit unusual because few third-grade students solved the problem by division. Most used either addition or multiplication. This student answered, "If you see in the table you can make half as many with the scoops, so the answer is 23." What did occur frequently, however, was that students — especially after the project year — marked up the booklet and extended the table downward even though they were not told to do so. This is another crude indication of how experience with this kind of problem, instead of picking-right-answers math items, can affect how children approach and conceptualize problems. Because they are asked to show their thinking, perhaps it also provides a more trustworthy indication of what they really know. Figure 12 includes other right

Figure 12 - Sample Student Responses on Maryland Mathematics Assessment Problem Set One (Lemonade Step 1-2) Illustrating Other Correct Answers and Explanations

You and your friend are in charge of preparing lemonade for 2 classes. You must decide how much lemonade to make for 46 students. Each student should get a cupful of lemonade.

STEP 1 Read this table from a lemonade mix container.

Scoops	Cups Made
1	2
3	6
5	10

You see a pattern in the table, but your friend does not. Tell your friend how many cups of lemonade can be made with 6 scoops of mix. Explain how you know this from the pattern in the table. Write on the lines below.

you add the same neber so exapel (6+6=12) (6x2=12)

If you put 1 scoop it will make 2. Then if you have 3 scoops it will make 6. So every scoop you do you will have to dubble that number.

With 6 scoops of mix you should be able to make 12 cups of lemonade I figed this out because 1+1=2, 3+3=6, 5+5=10 so 6+6=12 so that means you have 12 full cups of lemonade.

STEP 2 Think about the pattern you described above. If you have to make 46 cups of lemonade, how many scoops of mix will you need? Write your answer on the line below.

23

$$\begin{array}{r} 23 \\ +\ 23 \\ \hline 46 \end{array}$$

Source: Shepard, Flexer, Hiebert, Marion, Mayfield, & Weston, 1996.

answers that let you see the more typical pattern of either saying "6+6=12"or "6x2=12." "If you put one scoop it will make two, then if you make three scoops it will make six, so every scoop you do you'll have to double the number."

Now, let's look at the next page of student work, Figure 13. This is to make the last of the points about the benefits of performance assessments, which is the diagnostic value of these kinds of open-ended assessments. Their response represents a large category of kids. This student also wrote on the booklet extending the table downward. "Because on scoops it goes, 1, 3, 5, I saw that they're doing all odd, so I put odd, why cups was all even, and 4 in the middle. What I mean is, 2+4=6, and 6+4=10, and so on." First of all, in a traditional assessment, many of the students in this category would just be wrong, because many were not accurate enough in

Figure 13 - Sample Student Responses on Maryland Mathematics Assessment Problem Set One (Lemonade Step 1-2) Illustrating Different Way of Explaining the Pattern in the Table

STEP 1 Read this table from a lemonade mix container.

Scoops	Cups Made
1	2
3	6
5	10
7	14
9	18
11	22
13	26
15	30
17	34
19	38
21	42
23	46

You see a pattern in the table, but your friend does not. Tell your friend how many cups of lemonade can be made with 6 scoops of mix. Explain how you know this from the pattern in the table. Write on the lines below.

because on scoops it's go 1,3,5 I
saw that their doing all odd so I
put odd why cups was all
even and 4 in the mitel.
What I mean is 2 + 4 = 6 and 6 + 4 = 10
and so on.

STEP 2 Think about the pattern you described above. If you have to make 46 cups of lemonade, how many scoops of mix will you need? Write your answer on the line below.

23

Source: Shepard, Flexer, Hiebert, Marion, Mayfield, & Weston, 1996.

extending the table to get all the way to the correct answer of 23. Nor could they state the function rule of doubling to arrive at the answer computationally. But there's a tremendous amount of mathematical thinking going on here. We got more of this at the end of the year's project than before, and suddenly the teacher can understand how students are approaching the problem. Lots of students never made the left/right correspondence, but saw a pattern that they could explain in the downward extension of the table. These kinds of examples are exciting because they help us appreciate what's going on in each child's thinking. The next thing to do or ask instructionally is very different given what the child does know than if we presumed from a wrong answer that the child didn't understand patterns.

Marilyn Burns: "Children's Understandings are Fragile"

While I'm praising performance assessments, remember that I'm still in the part of my talk where I'm concerned about students being able to pretend to know. They can pretend to know on standardized tests, if we keep asking them to demonstrate skills in exactly the same format. In theory, we invented (or returned to) performance assessments to get away from that. But I also want to acknowledge that even with performance assessments, students may rely on familiar, rote routines and pretend to know.

I'm going to show you a six-minute segment from a staff development videotape by Marilyn Burns describing classroom assessment in mathematics (Mathematics: Assessing Understanding, 1993). The class we will see is a combination second and third grade. (*The audio transcript of this segment follows, with apologies to the reader because it cannot do justice to the video interaction the audience was able to see. Key summarizing statements by Marilyn Burns as narrator are underlined as well as critical points in assessing an individual student's knowledge.*)

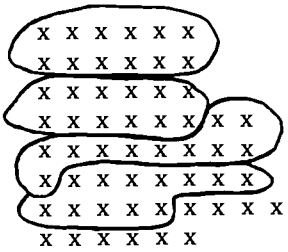
First, I should say that I stand in awe of Marilyn Burns and am grateful to her for providing these rich and powerful examples of assessment aimed at children's understandings. When I use this tape with students preparing to be teachers, it serves several purposes. The first and most important is one of attitude and philosophy. The comfortable relationship between Marilyn and this little girl is striking. It helps my students to understand that you don't have to be mean to your students or "catch them out" in what they don't understand. This is really a struggle between me and many students, especially those who are preparing to be elementary teachers. Often, they are really not so sure about this assessment business in the abstract because it implies being judgmental and mean. Marilyn's videos show her interacting with students in a way that says, "I'm trying to figure out what you know." The tone is clearly supportive and helpful, and there is no shock and dismay or labeling of "wrong" answers. Seeing examples like this helps to reassure prospective teachers. While they may not be quite this gifted in interacting with their students, finding out what their students don't know is a reasonable thing for a good teacher to do.

This example also illustrates that a good assessor must understand normal developmental progressions underlying concept mastery. When this student is unable to do a task that is too advanced, Marilyn is able to back down an implied developmental progression with her questioning. She knows how to back up, back up, back up to the place where the student is finally able to do things, and to check on understanding in that way. Although she had reason to start where she did because she thought the student understood the 49 problem in front of the class, she backs up as soon as she realizes that we're in the 20's and this is not making sense. She backs up to where she can come to an understanding of what the student does know in representing numerals and what she doesn't understand about place value.

The key point to be made in connection to today's talk, "How Should We Measure to Check for Robust Understandings," is that what Marilyn is doing here is asking in different ways. Even in that apparently manipulative-based classroom, the kids have gotten in the habit of drawing those stars. They draw out all the circles, they put a circle around ten stars, and then they put another circle around ten stars.

Marilyn introduces the tape by saying that she can use classroom discussion to learn what children are thinking but only by asking them to explain their reasoning.

As this segment begins, the teacher has drawn on the board a series of stars. Most have been circled in groups of 10.



The class counts with her the last nine stars, "1, 2, 3, 4, 5, 6, 7, 8, 9."

Teacher: "Do I draw a circle around those?"

Class: "No."

One of the girls in the class, Cena, says (pointing), "Look, you've got (stops and counts 1, 2, 3, 4), you've got 4 10s and you like put a 4 right there and you've got nine stars left over, and then you put nine right there." (The teacher writes 49.)

Class: "49."

Cena: "49."

When I have used this tape in an elementary math-science methods course, I stopped the tape at this point and allowed University students preparing to be teachers to discuss in groups what they have been able to observe about what this student, Cena, knows. Some of the preservice teachers noticed ways in which Cena is or is not confident with the correspondence between the numerals and number, but nearly all were fairly satisfied that she understands place value. As the tape continues, you're going to see Cena participating in an individual assessment. This is the same little girl, even though in one case her hair is braided and in the other case it's long.

Marilyn (as narrator): "Children's understandings are often fragile. What they know in one setting doesn't always transfer to another."

Marilyn: "Put the tiles so that you have groups of 10. And would you count out loud so I can hear what you're doing."

Cena: "1, 2, 3, 4, 5, 6, 7, 10." (Cena counts one group of 10.)

Marilyn: "Do you have enough to make another pile of 10?"

Cena: (Nods yes.)

Marilyn: "Let's see."

Cena: "1, 2, 3, 4, 5, 6, 7, 10." (Cena counts another group of 10.)

Marilyn: "So how many groups of 10 do you have?"

Cena: "2"

Marilyn: "And how many extra tiles do you have?"

Cena: "4" Marilyn: "Do you know how many tiles you have altogether?"

Cena: (Shakes her head, no.)

Marilyn: "How would you find out?"

Cena: "Count."

Marilyn: "So, how would you count them?"

Cena: "Like, 1, 2, 3, 4." Marilyn: "And let's see you do that."

Marilyn (as narrator): "After grouping the tiles into 10s, Cena wasn't able to use this information to determine the number of tiles. She needed to count and chose to do so one by one." [I might also note that Marilyn's matter of fact tone suggests that she is not shocked, as many viewers of the tape are, to realize that a child may not understand automatically that 2 10s and 4 is 24.]

Marilyn: "Do you know how to write the number 24?"

Cena: "Yes." Marilyn: "Would you do that for me."

Cena: (Writes the number 24.)

Marilyn: "Now, suppose I said that I didn't want 24 tiles anymore, I wanted only 16 tiles. Could you take some tiles away so you're left with only 16? How would you solve that problem?"

Cena: "Um, by counting back(?)" (Counts 6.)

Marilyn: "Could you put those away?" "How many do you think are left now?"

Cena: "11 (?)"

Marilyn: "Do you want to count and check?"

Cena: "1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 (smiles when she realizes there are more than 11), 12, 13, 14, 15, 16, 17, 18."

Marilyn: "So you've got 18. Let's put all these 18 in a pile. And you wrote the number 24 so well, do you know how to write 18?"

Cena: (Writes 18.)

Marilyn (as narrator): "I settled for 18 tiles rather than the 16 I had asked for. Either number would allow me to further assess Cena's understanding."

Marilyn: "I agree that's how you write 18. Cena, can you tell me with the tiles what 8 means? Put them right up here next to the eight."

Marilyn (as narrator): "I'm interested in Cena's interpretation of the digits in the numeral. Children often write numerals correctly yet have no concept of place value."

Marilyn: "So I just want to see just the eight."

Cena: "1, 2, 3, 4, 5, 6, 7, 8. (Counts eight tiles on the eight.)

Marilyn: "So this is just eight tiles. When you wrote the number eighteen, you wrote a one and an eight. Can you show me what the one means?"

Cena: (Puts one tile on the tens place.)

Marilyn: "And when you counted 18 there were all of these tiles together, so if this is the eight and this is the 1, where do those fit? (pointing to the remaining tiles.)

Cena: "Over here." (Puts back with other extra tiles).

Marilyn: Oh, over there. So now we have...how many tiles do you think we have here?"

Cena: "Nine."

Marilyn: "So if you put eight here and one here, we don't have eighteen any more, we just have nine."

Marilyn: "Suppose I asked you to put four more tiles there."

Cena: (Counts four more.)

Marilyn: "You had nine and now you have four more. Do you know how many you have all together? Can you figure that out in your head?"

Cena: "If we have nine, then we put four more. Then we have...." Marilyn: "How are you trying to figure that out?"

Cena: "In my brain."

Marilyn: "What are you doing in your brain?"

Cena: "Counting."

Marilyn: "Do you want to do it out loud so I can hear?"

Cena: "1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13."

Marilyn: "And if I said to you, like I did when we first started, can you make a pile of 10? Do you have enough there to make a group of 10?"

Cena: (Shrugs.)

Marilyn: "Don't know? How would you find out? Do you think you do or do you think you don't?"

Cena: "Do."

Marilyn: "You do. Do you think you will have any extras?"

Cena: (Nods Yes.)

Marilyn: "Do you know how many extras you'd have?"

Cena: (No response.)

Marilyn: "How many do we have here all together?"

Cena: "13."

Marilyn: So if we took ten away and made a group of 10 how many extras do you think you'd have?"

Cena: "1." Marilyn: "Do you want to try it and see?"

Cena: "1, 2, 3, 4, 5, 6, 7, 8, 9, 10."

Marilyn: "And how many extras do you have?"

Cena: (Groups 10.) "3."

Marilyn: "Can you write the number 13?"

Cena: (Writes 13.)

Marilyn: "So, I look at the number and I see a clue there that tells me you will have three extras. Do you see a clue there that will tell you you have three extras?"

Cena: "Yes."

Marilyn: "What's the clue that you see?"

Cena: (Nods vigorously and stacks 3 tiles on 3.)

Marilyn: "And what do you think the one means? That's the three extras, what do you think that one means?"

Cena: (Puts one tile on the numeral 1.)

Marilyn (as narrator): "Partial understanding is natural to the learning process. Cena needs a great deal more experience to connect what she does understand to the symbols."

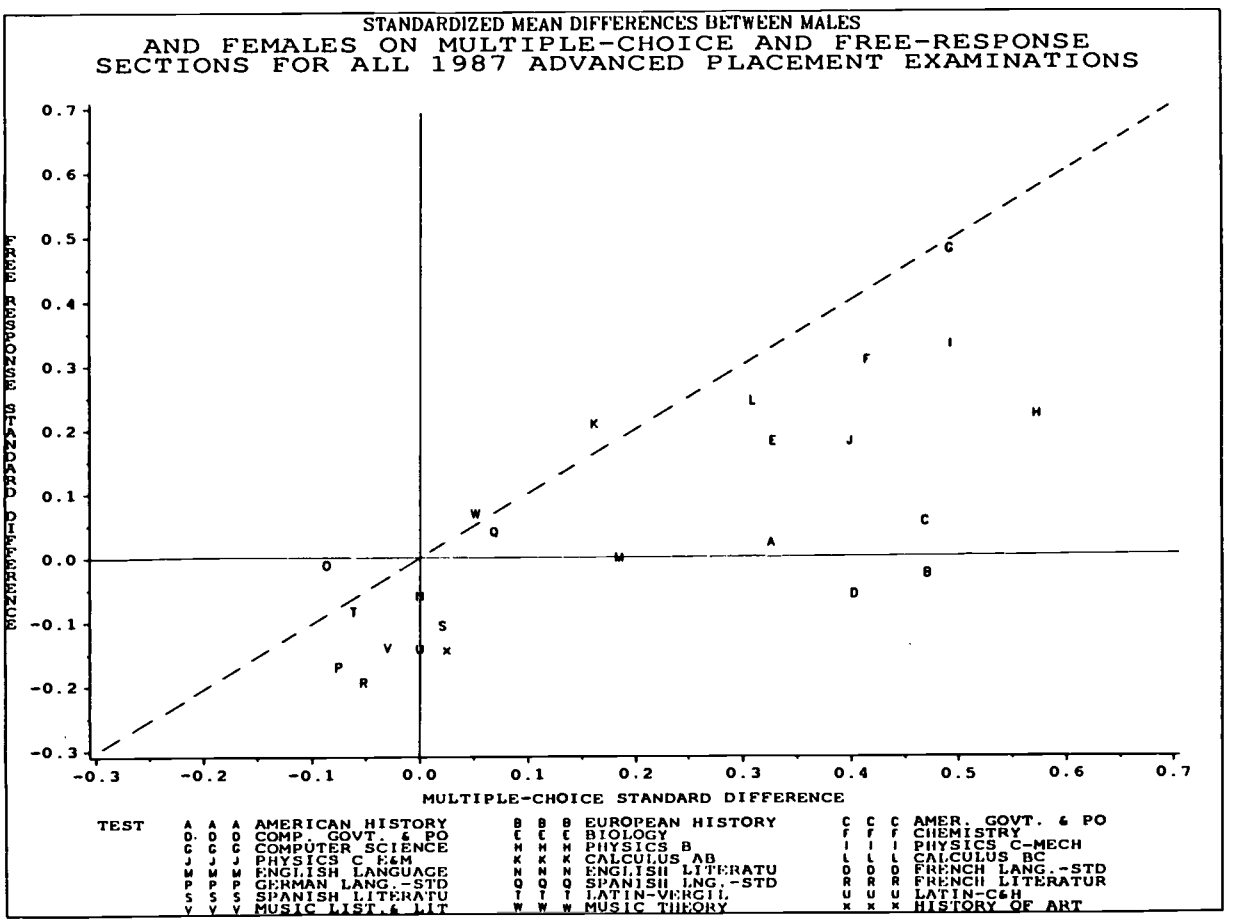
They do exactly what they've been trained to do. And, somehow, because it's been routinized, they can look like they understand these things, even though we have clear evidence that one little girl did not understand what she was doing. Just as in the teaching-the-test literature, this is an example of kids pretending to know. What good assessment has to do is ask in different ways to uncover misunderstandings. It should discover partial understanding and understandings that are "fragile," which means that the child's apparent knowledge does not generalize across contexts.

Research on Test Bias

If "pretending to know" is the worry of the teaching-the-test research, the other side of the coin when looking at test bias is the concern that students "really do know," but are prevented from showing what they know by some unnecessary, construct-irrelevant difficulty in the test. I don't have as complete a set of examples for this side of the argument, but researchers here at ETS, such as Janice Scheuneman have been among those who have documented specific instances of construct-irrelevant features that make some test items differentially difficult for some groups of test takers. For example, Scheuneman (1979) found that negatively worded School Language items were unusually difficult for African American children taking the *Metropolitan Readiness Test*. Although such items might indeed contribute to prediction of children's likely school success, they are misleading as indicators of children's general level of language development. In the context of test bias, the question is whether if asked a different way, the children would be able to demonstrate their understanding.

Figure 14 is from a study by ETS researchers Alicia Schmitt, John Mazzeo, and Carol Bleistein (1991). The graph shows the mean standardized differences between males and females on various Advanced Placement Examinations with the multiple-choice test sections plotted on the X axis and the constructed response portion of each exam on the vertical axis. Results do not fall along the 45 line as would be expected if the two parts of the test were equally difficult for males and females. For example, on the American and European History essay exams (tests denoted A and B), males and females performed roughly the same with a mean difference near zero, but on the multiple-choice portions of the exams, males out performed females by .3 and .5 standard deviation units, respectively. This type of pattern, where males do relatively better on multiple-choice tests and females do relatively better on essay tests has been replicated in many other large-scale assessment programs, not just the AP exams shown here. Of course, data such as these still leave us with the question, "Is this bias, or isn't it?" Are there two slightly different constructs represented by the two parts of the test, one ensuring broad curricular coverage and tapping knowledge of historical periods in the case of multiple-choice questions, and the other involving historical argument and ability to use primary data to support an argument in the case of the essay portion of AP history? Follow-up studies such as the one by Bridgeman and Lewis (1994) have advanced our understanding but still have not fully resolved the question as to whether multiple-choice questions can really be called biased against women. It can be said, however, that women are unfairly disadvantaged by the use of multiple-choice questions when the goal is to predict performance in college history courses, because essay exams are used prevalently in college history courses and women earn grades equal to men's on the criterion measure. The point here is that women on average have a better opportunity to demonstrate their competence in history when essay exams are used.

Figure 14 - Standardized Differences Between Males and Females on Multiple-Choice and Free Response Sections for All 1987 Advancement Examinations



Source: Schmitt, Mazzeo, & Bleinstein, 1991.

Conclusions: Implications for Classroom and Large-Scale Assessment

These two perspectives, or these two different stances — on the one hand, arguing for giving the students the benefit of the doubt, and the other hand trying to push a little harder to see whether they really know — these two sides can be reconciled by trying to sort through incongruities or inconsistencies in performance in terms of either construct relevant or irrelevant sources of difficulty. If a student can perform similar tasks in one context and not in another, do I conclude that his knowledge is incomplete or that one of the settings is affected by some unfair artifact? Only careful reasoning about what is the same and what is different across task performances can help to resolve the dilemma.

For my students preparing to be teachers, I try to emphasize two equally important principles. First, assessments should let students show what they know. My students like this principle. They can identify with it. It resonates with their suspicions about examinations and assessments. They are eager to learn about multiple ways that students can demonstrate proficiency because that's going to help them be fairer in their own classroom assessments. Even on this point, however, we have to do some work to think about how it should actually be implemented in classroom practice. When many preservice teachers first think about multiple modes of assessment, they think it means choice, "Let one student do it this way, and another student do it a different way." My perspective is to use a variety of assessment methods, so that each student will have the chance to perform using the mode that they do best, but to have all students respond to all methods so that I can see and try to understand how mode of assessment affects performance. I'm also committed to helping each student work at getting better at the thing they do least well as well as the thing they do best. Isn't that part of helping students learn? For special needs students, we may need to make the same accommodation across all tasks (which means choosing the most advantageous assessment mode), but for most students, the practical way to honor both principle one and principle two below is to use multiple methods of assessment and to have all students participate in all of the methods.

The second principle is that assessments should not let students appear to know when they don't really know. As I've said previously, prospective teachers are not as eager to embrace this principle because they fear it will take them out of their supportive and nurturing role. That's why Marilyn Burns's skill as an assessor is such a powerful example: illustrates how asking focused questions that get at real understanding can be an act of kindness and good teaching. Good assessment should be so entwined with good teaching that it becomes impossible to see where one leaves off and the other begins. In the videotape segment, there was only one point where Marilyn intervened and directly "taught," "I look at the number and I see a clue there that tells me you will have three extras. Do you see a clue there?" And, because her assessment is giving her close insights about where she can best extend Cena's understandings, Cena is indeed able to respond. But in other interactions Marilyn is not just gathering information, she is also teaching as she asks Cena to perform tasks that highlight the connections between objects, number, and numerals. If a student doesn't understand a concept, even recognizing some dissonance — as when they exchanged a smile over the student's misestimate about the number of objects — can be a first step in developing further understanding.

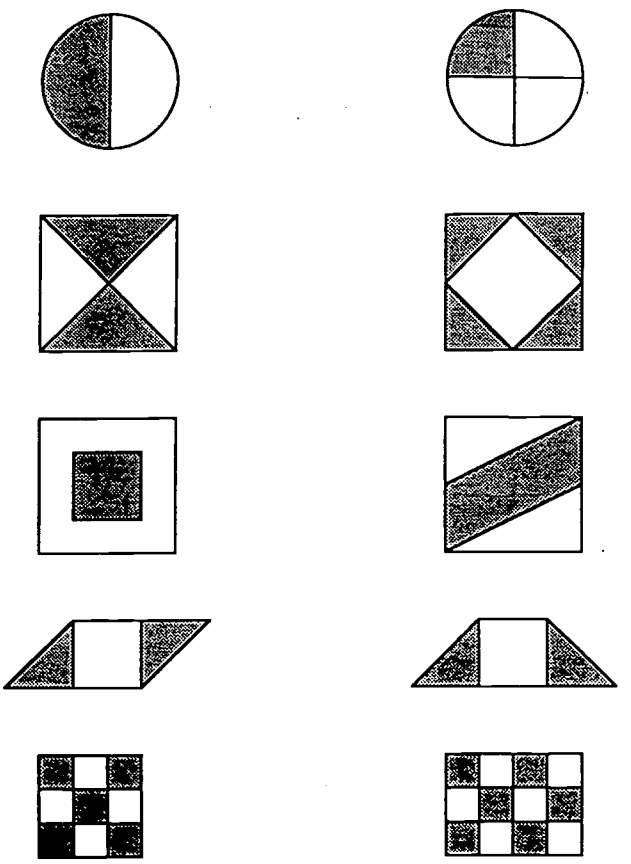
My term, "robust" understandings comes from Marilyn talking about children's understandings being fragile. Kids know it one way, but they don't know it the other way. That's what sent me back to the measurement question, "Is it bias?" Or is it not really knowing? What is it?" They could appear to know, but performance may be highly dependent on format and context. Ultimately it is important to realize that this is not just a measurement problem. The problem of fragile understandings is at the heart of teaching and learning. How should we help students learn in ways that ensure transfer and generalized knowledge? Good teaching constantly asks about old understandings in new ways, calls for new applications, and draws new connections to help develop robust understandings. What this means to me for both assessing and teaching in the classroom

Figure 15 - Examples of Different Ways to Ask About the Concept One-Half

Assessing Mathematical Understanding 9/89

Name _____ Grade _____

1. Ring each shape that has one half shaded.



2. Suppose there were a sale and everything was 1/2 off--you could buy something for 1/2 the original cost.

   How much would an item cost that originally sold for $10.00? _____

   How much would an item cost that originally sold for $1.98? _____

   How much would an item cost that originally sold for 75¢? _____

Source: Assessing Mathematical Understanding, 1989.

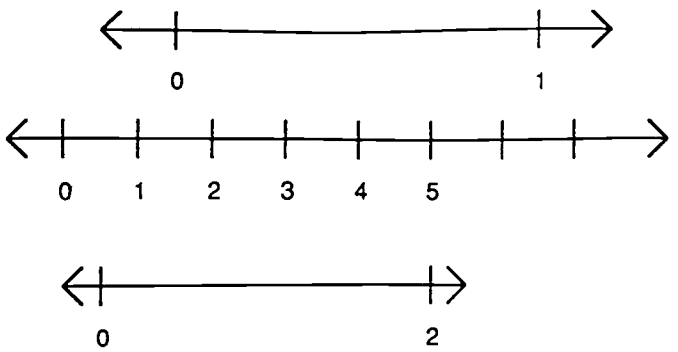
is that after introducing a concept or problem type in one particular form, I must constantly work to extend that knowledge and ask about the concept in new ways. So I'm not going to hit students with every possible application at first, but I won't just let them rest on the one comfortable way that we did a problem before.

Figures 15 and 16 are two pages of simple problems from Assessing Mathematical Understanding all aimed at checking for understanding of one-half. They illustrate how the same concept can be approached in different ways. So, in the first set, kids are supposed to decide which of these shapes actually represents a half. If you also asked children to explain their reasoning you would hear talk about "two equal parts," or the "same number of squares" in the checkerboard example, which is fundamental to the idea of one-half but not something that children

Figure 16 - More Examples of Ways to Ask About the Concept One-Half

Assessing Mathematical Understanding 9/89

3. Mark approximately whre the number 1/2 would be on each number line.



6. Circle all below that are equivalent to 1/2.

1 ÷ 2	2 ÷ 1	32 ÷ 64	48 ÷ 24
1/2	2/1	6/12	18/9
0.5	0.05	0.50	1.2
100%	50%	120%	5%
1/4 + 1/4	3/8 + 1/8	1 X 1/2	1 ÷ 1/2

5. What is 1/ 2 of:

100 _____	48 _____	12 _____	5 _____
2 _____	1 _____	0 _____	1/2 _____
3/4 _____	2/3 _____	6/7 _____	3/5 _____

Source: Assessing Mathematical Understanding, 1989.

are always helped to "see" as a part of instruction about fractions. Then in the next set, "Suppose there was a sale and everything was one-half off, you could buy something for half the original cost. How much would an item cost that originally sold for $10.00, $1.98? How much would an item cost that originally sold for 75 cents?" On the next page are more 1/2 problems involving number lines, division, fractions, percents, and so forth. An obvious goal is to help students understand the interchangeability of decimals and fractions as well as the areas, distances, and objects they represent. Many teachers don't teach in a way that lets kids ever see the connection between decimals and fractions. If you haven't been in an elementary classroom recently, you may not realize that some children even have compartmentalized knowledge of money problems and decimals (with greater accuracy on both money problem computation and estimation). So these examples help us think about the teaching implications that go hand in hand with the assessment implications. When I talk about teaching to develop robust understandings, I'm returning to the very old idea that we've always had about teaching for transfer.

Large-scale assessments face similar problems in trying to represent accurately what students know. These issues aren't quite the same as the classroom issues, but they are related. When performance does not "generalize" from one type of assessment task to the next, we want to know why. When is it measurement artifact? When is it non-generalizable knowledge suggestive of "fragile" or incomplete learning? When is it non-generalizable measurement, reflective of specialization or depth of knowledge not captured by assuming that all items in a domain are interchangeable? When I say that we need to find the explanations for lack of generalization, I'm referring to the need to sort out the construct-relevant versus the construct-irrelevant explanations for inconsistencies in performance. So sometimes it might be a measurement artifact. For example, a child confronted for the first time by comprehension questions following a story might be baffled and appear to be a poor reader yet when asked to retell the story could give a reasonable account. This would be an example of unfairness or bias caused by lack of familiarity with test format. But sometimes non-generalizable performance is suggestive of fragile understandings as in the videotape. Asking in a different way from the familiar format gave a truer picture of the student's understanding. In this case we would believe the lower score not the familiar-format performance. A point worth noting is that it would be very hard, just from the statistical summary of the data, to know which of those two has occurred. You would need to do more close-hand investigations, think-aloud studies, and comparisons to classroom work, to figure it out.

Lastly, we have to be aware that non-generalizable measurement (i.e., non-equivalent performance across tasks in the same assessment) could be reflective of specialized curriculum and depth of knowledge in some areas but not others. Statistically this would look the same as fragile understandings because it is also "incomplete" knowledge but would have different import for documenting achievement depending upon the structure of the knowledge domain. Many of our existing measurement models and assumptions made

sense when achievement was conceived of in terms of discrete skills measured using formats that were equally familiar to all test takers. As we attempt to develop measures of more advanced content, the assessments cannot simply be harder versions of basic skills tests. In my talk this morning, "Testing for Advanced Achievement without a Syllabus," I used the example of two graduate students in measurement each required to take each other's comprehensive exams or dissertation orals. Although there would certainly be some common content, most of the questions that tap advanced knowledge would be tailored to the specific type of problem the student had been working on; taking each other's exams would give a misleading picture of achievement. Thus far, there has been very little attention to how curricular differences among examinees taking the same large-scale assessment might affect how standards are set or how the generalizability of the assessment itself should be evaluated. What if subgroups of students are following two or more different instructional pathways, as opposed to being at different stages on one pathway? This would have implications for the statistical models that we choose. The type of validity studies I have proposed will help us understand what's going on with the measurement by helping us to connect measurement results more closely to the learning that has occurred.

References

Bridgeman, B., & Lewis, C. (1994). The relationship of essay and multiple-choice scores with grades in college courses. *Journal of Educational Measurement*, 31, 37-50.

Cannell, J. J. (1987). *Nationally normed elementary achievement testing in America's public schools: How all 50 states are above the national average* (2nd ed.). Daniels, WV: Friends for Education.

Flexer, R. J. (1991). Comparisons of student mathematics performance on standardized and alternative measures in high-stakes contexts. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL, April.

Koretz, D., Linn, R. L., Dunbar, S. B., & Shepard, L. A. (1991, April). The effects of high-stakes testing on achievement: Preliminary findings about generalization across tests. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Linn, R. L. (1995). *Assessment-based reform: Challenges to educational measurement.* Princeton, NJ: Educational Testing Service.

Linn, R. L., Graue, M. E., & Sanders, N. M. (1990). *Comparing state and district test results to national norms: Interpretations of scoring "Above the national average,"* CSE Technical Report 308. Los Angeles: Center for Research on Evaluation, Standards, and Student Testing.

Mathematics: Assessing Understanding. Individual Assessments: Part I. (1993). White Plains, NY: Cuisenaire Company of America.

Scheuneman, J. (1979). A method of assessing bias in test items. *Journal of Educational Measurement*, 16, 143-152.

Schmitt, A. P., Mazzeo, J., & Bleistein, C. (1991, April). Are gender differences between Advanced Placement multiple-choice and constructed response sections a function of multiple-choice DIF? Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Shepard, L. A. (1991). Will national tests improve student learning? *Phi Delta Kappan*, 232-238.

Shepard, L. A. (1995). Parents' thinking about standardized tests and performance assessments. *Educational Researcher*, 24, 25-32.

Shepard, L. A., Flexer, R. J., Hiebert, E. H., Marion, S. F., Mayfield, V., & Weston, T. J. (1996). Effects of introducing classroom performance assessments on student learning. *Educational Measurement: Issues and Practice*, 15, 7-18.

MEMORIAL LECTURE Series

ETS

POLICY INFORMATION CENTER
Educational Testing Service
Princeton, New Jersey 08541-0001

04202-13515 • S107M3 • 204924 • Printed in U.S.A.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

NOTICE

REPRODUCTION BASIS

☑ This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☐ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").