DOCUMENT RESUME

ED 471 756 SE 067 015

AUTHOR	Confrey, Jere; Carrejo, David
TITLE	A Content Analysis of Exit Level Mathematics on the Texas Assessment of Academic Skills: Addressing the Issue of Instructional Decision-Making in Texas.
PUB DATE	2002-00-00
NOTE	13p.; In: Proceedings of the Annual Meeting [of the] North American Chapter of the International Group for the Psychology of Mathematics Education (24th, Athens, GA, October 26-29, 2002). Volumes 1-4; see SE 066 887.
AVAILABLE FROM	ERIC/CSMEE Publications, 1929 Kenny Road, Columbus, OH 43210-1080. Tel: 800-276-0462 (Toll Free).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	Academic Achievement; Educational Change; *High Stakes Tests; *Mathematics Education; *Problem Solving; Secondary Education
IDENTIFIERS	*Texas Assessment of Academic Skills

ABSTRACT

High stakes tests are increasingly used to monitor systemic improvements in mathematics and teachers are expected to rely on the results of such tests to adapt their instructional practices. We examine the Texas Assessment of Academic Skills (TAAS) over a period of three years to examine to what extent its results can be used to guide instructional decision-making. We present the results of an expert content analysis of the 10th grade TAAS mathematics test for 1999, 2000, and 2001 which reveal that problem solving objectives mask significant content emphases. We further examine the variation in raw scores by objective across grades and years to show this information is not reliable enough to guide changes in instruction. We examine the sampling of topics within objective to gauge the distribution across topics. Finally, we attempt unsuccessfully to account for changes in difficulty using a combination of changes in sampling, item characteristics and composition of distracters. This leads us to question the utility of providing teachers raw data by objective and points to the urgency of developing better methods to link content analyses and psychometric methods of scoring. (Author)

A CONTENT ANALYSIS OF EXIT LEVEL MATHEMATICS ON THE TEXAS ASSESSMENT OF ACADEMIC SKILLS: ADDRESSING THE ISSUE OF INSTRUCTIONAL DECISION-MAKING IN TEXAS

Jere Confrey
University of Texas at Austin
jere@mail.utexas.edu

David Carrejo
University of Texas at Austin
carrejo@mail.utexas.edu

High stakes tests are increasingly used to monitor systemic improvements in mathematics and teachers are expected to rely on the results of such tests to adapt their instructional practices. We examine the Texas Assessment of Academic Skills (TAAS) over a period of three years to examine to what extent its results can be used to guide instructional decision-making. We present the results of an expert content analysis of the 10th grade TAAS mathematics test for 1999, 2000, and 2001 which reveal that problem solving objectives mask significant content emphases. We further examine the variation in raw scores by objective across grades and years to show this information is not reliable enough to guide changes in instruction. We examine the sampling of topics within objective to gauge the distribution across topics. Finally, we attempt unsuccessfully to account for changes in difficulty using a combination of changes in sampling, item characteristics and composition of distractors. This leads us to question the utility of providing teachers raw data by objective and points to the urgency of developing better methods to link content analyses and psychometric methods of scoring.

Since 1995, the state of Texas relied on the Texas Assessment of Academic Skills (TAAS) to drive the current accountability system for Texas schools. Along with attendance and course credits, passage of he TAAS was required for a student to graduate. A new assessment, the Texas Assessment of Knowledge and Skills (TAKS) will be implemented in 2003, and although better aligned with secondary topics, it will produce similar data artifacts for teacher use. TAAS results, in the form of raw scores by test objective, scaled scores known as Texas Learning Index (TLI) scores, and item analysis percentages have been provided to teachers and administrators to guide them in improving school performance and providing quality instruction in mathematics. Our interest was in examining these data to determine if they could validly be used for these purposes. Our interest in the question was piqued by data showing discrepant results as student scores on TLI increased while raw scores have declined, typically explained by the state as differences in test difficulty (Confrey & Carrejo, 2002). We sought to understand at the level of classroom practice, if the data provided by raw score for each of thirteen objectives could be used to make instructional changes, as it is widely believed.

In the case of the TAAS test, we propose that even within a single test and its analyses, drawing data-driven conclusions reveals how psychometric traditions for creating scaled scores and equated tests seem to produce contrasting results from content analyses. The key issue of content validity (Heubert & Hauser, 1999) "whether a test measures what it purports to measure and what conclusions can be drawn from the results – and whether the conclusions or inferences drawn from the test results are appropriate" (p. 71), requires one to link these two arenas. Thus our goal is to link psychometric and content analytical traditions by building protocols to help teachers refine their analysis and increase their statistical capacity to interpret and critique data and create plans of action (Confrey and Makar, 2002). In addition, we wish to argue for the role of outside content experts when conducting analyses of the TAAS. We argue for more discipline-based protocols for conducting and interpreting item analyses, and point out that these analyses are likely to be neglected or unpublished under the current accountability system.

In this paper, we provide an account of our content analysis of the mathematics portion of the TAAS test for the 10th grade level over a period of three years (1999, 2000, and 2001). We address, based on available information, the question, "Can data provided to teachers in the form of raw scores by objective provide an accurate description of student performance and support instructional decision making?" In addressing this question, we relied on answers to three sub-questions:

1. Using an independent content protocol, do the constructed TAAS objectives provide an adequate description of the content tested?
2. How much variation is there in students' mean performance by these objectives across the grades over time?
3. Can we identify the possible factors that result in changes in performance by objective through a content sampling and item analysis?

As development of the new assessment, TAKS, is underway, examination of the TAAS, its construction, format, and content, provides important experimental groundwork for future analysis of the new test, designed to be closely aligned with TAAS.

TAAS Construction and Format

According to the test makers, the Texas Education Agency (TEA, 2001), construction of the TAAS first involves a review of state standards for mathematics, the Texas Essential Knowledge and Skills (TEKS), to determine appropriate learning objectives by grade level. Following this review, educator committees develop drafts of test objectives, linked to the TEKS, to be reviewed by teachers and other specialists. The objectives are then refined based on feedback, and sample test items are written. Educator committees develop guidelines for assessing each objective which include eligible test content, test item formats, and sample items. A test blueprint is then developed by item writers, some of whom are identified as former teachers. TEA curriculum specialists then review the items. During this process, item review committees judge the content, difficulty, and bias of each item. Further item revision occurs and then they are field tested. Field-test data is analyzed for reliability, validity, and possible bias. Data review committees then use statistical analyses to determine if items are worthy of being used. The final blueprint is then developed. Field-test items are placed in an item bank and the final tests are built from the bank and are designed to be equivalent in difficulty from one administration to the next.

TEA outlines thirteen objectives grouped in three categories: concepts, operations, and problem solving. Within concepts there are five objectives: (1) Number Concepts, (2) Relations and Functions, (3) Geometric Properties, (4) Measurement, and (5) Probability and Statistics. Within operations are (6) Addition, (7) Subtraction, (8) Multiplication, and (9) Division. Within problem solving are the four remaining objectives: (10) Estimation, (11) Solution Strategies, (12) Representation, and (13) Reasonableness. TAAS questions are clustered in groups of four under each objective with the exception of Solution Strategies and Representations which have eight items each at the exit level. Therefore, sixty questions comprise the final multiple choice exit test with each question having a possible 4-5 answers. Furthermore, TEA outlines which state mathematics standard(s), i.e. TEKS standard(s), is tested by each aforementioned TAAS objective.

Examining TAAS Content with an Expert Protocol

Our approach to the analysis began by:

1. Creating a protocol based on our selection of topics relevant to K-12 mathematics education. The topics comprising the protocol are indicative of the breadth of subject matter teachers and researchers find most relevant to many, if not most, implemented curricula. Topics were also chosen based on current research in mathematics education.
2. Obtaining copies of the tests (available from the TEA website) and categorizing each item according to the constructed protocol without referring to their respective TAAS designation. If a general topic was missing for an item, it was identified and the categories were adapted until we could account for all items on the test.
3. Comparing our categorization with that of TEA's for all three years.

Our protocol includes the following topics (followed by subtopics): a) numeration (scientific notation, sequences), b) geometry (angle, congruency, coordinate plots, formula, spatial reasoning, similarity, symmetry, and vertices, edges, and faces), c) measurement (linear/perimeter, area, volume, weight, and the Pythagorean Theorem), d) operations (addition, subtraction, multiplication, division), e) rate (ratio and proportion), f) probability (combination, experiment outcomes, and mean, median, mode), g) data and statistics, and h) equations (literal, equation with variable, inequality). For d) operations, we constructed the following table to indicate specific number types used in the items we placed under this topic (see Figure 1).

Likewise for g) data & statistics, we constructed the following table to indicate the representation format involved in the question and the representation involved in the answer choice (see Figure 2).

Our categorization separated the topics of ratio and data analysis as separate categories (see Figure 3).

Our categorization separated the topics of ratio and data analysis as separate categories (see Figure 3). We used it to examine and display content distribution on the test over the three years. The following chart shows the number of items per topic. It is in marked contrast with the test specifications which specify four items per objective for most objectives and eight for solution strategies and representation.

One can see the emphasis on certain topics such as addition and subtraction remain relatively unchanged from year to year. However, other topics such as mul-

	Whole	Integers	Fractions	Decimal	Percent	Exponents
Addition						
Subtraction						
Multiplication						
Division						
Combined ops.						
Comparison						
Representation						
Ordering/sequence						

Figure 1.

		Answer form					
		Read value	Table	Line	Bar	Verbal	Circle
Question form	Table						
	Line						
	Bar						
	Verbal						
	Circle						

Figure 2.

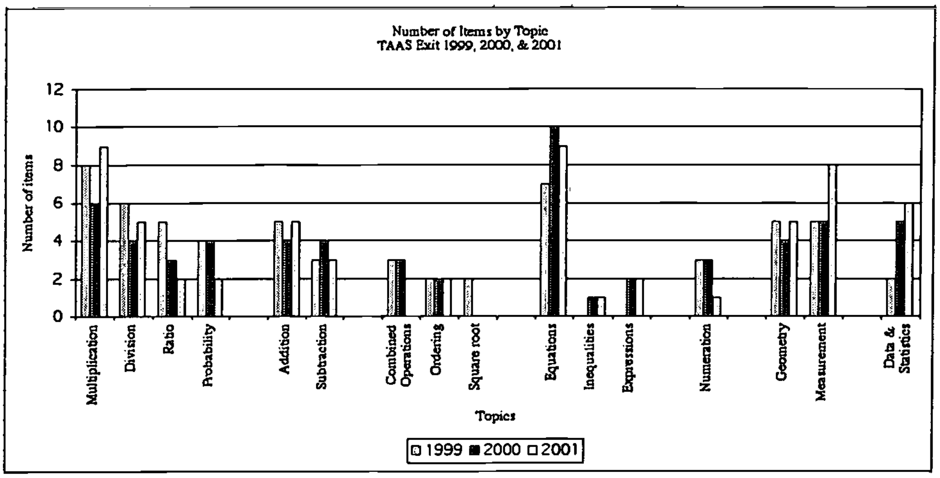

Figure 3. Distribution of items by topic, 1999, 2000, 2001

tiplication, division, ratio, equations, and data receive different emphasis from year to year. (Notably, measurement received a heavier emphasis in 2001 compared to the previous two years.) The major difference lies in our elimination of the Problem Solving objectives (Estimation, Solution Strategies, Representations, and Reasonableness) and categorization of their items into content categories. The following table lists the item numbers in each of these Objectives and the heading under which we placed them on our protocol (see Figure 4). Also note the variation in topics year to year.

The total number of items in the 1999 test related to multiplicative structures (multiplication, division, ratio, rate) is eighteen out of the sixty items. The total number of items in 2000 is eleven out of the sixty and the total number for 2001 is fifteen out of sixty. We found multiplicative structures far more heavily represented than teachers recognized, and hence its importance in passing TAAS could have been easily neglected. Furthermore, TAAS Objective 5, Probability and Statistics, contains four items, yet on the overall tests, the total number of items related to data, statistics, and probability under our protocol is six out of sixty for 1999, nine out of sixty for 2000, and eight out of sixty for 2001. This area is also underrepresented in the test specifications relative to the actual test. We conclude that constructs in multiplicative structures and data and statistics (including probability) comprise 24 out of 60 or 40% of the items for 1999, 20 out of 60 or 33% of the items for 2000, and 23 out of 60 or 38% of the items for 2001. This analysis indicates how the Problem Solving category with its four Objectives mask content that teachers should emphasize in their instruction. We understand why the Problem Solving objectives were included but point out that they

Objective 10 – Estimation					
1999		2000		2001	
Item	Protocol Topic	Item	Protocol Topic	Item	Protocol Topic
23	Mult/whole	21	Comb.Op./whole	21	Mult/whole
24	Mult/whole	31	Equation/literal	23	Mult/dec.
36	Division/whole	42	Comb.Op./dec.	26	Data/line, read
37	Division/%	50	Mult/%	27	Mult/%
Objective 11 – Solution Strategies					
1999		2000		2001	
Item	Protocol Topic	Item	Protocol Topic	Item	Protocol Topic
22	Numer/sequence	22	Numer/sequence	25	Equation/literal
25	Rate/ratio	24	Meas/perimeter	28	Meas/area
26	Equation/literal	26	Rate/ratio	29	Equation/literal
30	Rate/ratio	28	Meas/perimeter	30	Equation/literal
31	Rate/ratio	30	Equation/literal	32	Meas/volume
34	Numer/sequence	34	Equation/literal	35	Meas/perimeter
39	Numer/sequence	35	Equation/literal	36	Meas/volume
43	Comb.Op./whole	43	Comb.Op./whole	38	Mult/%
Objective 12 – Representations					
1999		2000		2001	
Item	Protocol Topic	Item	Protocol Topic	Item	Protocol Topic
29	Equation/literal	25	Equation/inequal	22	Equation/vari.
32	Equation/literal	29	Equation/literal	24	Geom/coord
33	Equation/literal	32	Data/line, table	31	Division/%
35	Comb.Op./dec.	33	Data/line, table	33	Equation/inequal
40	Equation/literal	36	Equation/literal	34	Data/line, table
41	Comb.Op./%	37	Equation/vari.	39	Equation/vari.
42	Equation/literal	38	Data/table, line	42	Equation/vari.
44	Data/Venn	39	Equation/vari.	43	Equation/vari.
Objective 13 – Reasonableness					
1999		2000		2001	
Item	Protocol Topic	Item	Protocol Topic	Item	Protocol Topic
21	Data/bar, verbal	23	Data/table, verbal	37	Data/table, verbal
27	Mult/dec.	27	Data/bar, verbal	40	Data/bar. verbal
28	Mult/dec.	40	Equation/literal	41	Mult/whole
38	Mult/whole	41	Mult/dec.	44	Mult/dec.

Figure 4. Content protocol for objectives 10-13.

should be crossed dimensionally with the Content objectives rather than designed as their own categories.

Variation in Students' Mean Performance by Objective

Our second investigation concerned the variation in student performance for each objective. Teachers were using declines or gains in performance by objective as evidence of instructional need or success. We decided to examine the variation in students' mean scores in two ways. First, we examined student performance for TAAS assessments 1999, 2000, and 2001 disaggregated by objective. Then we examined performance over three years by grade level to see if any trend lines suggested patterns of improvement or decline (see Figure 5).

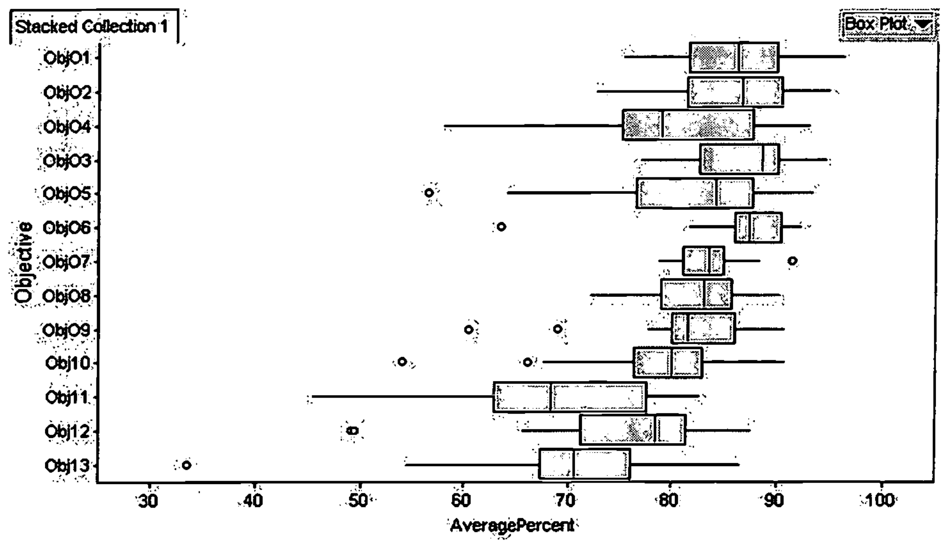

Figure 5. Variation in mean scores for all grades.

Note that two of the four Objectives related to operations; namely, Objectives 6 & 7 show little variation for all grades. However, Objectives, 4, 5, and 11, Measurement, Probability & Statistics, and Solution Strategies respectively, show considerably more variation. These data raise questions whether raw scores reported by objective are relatively stable enough over time to guide instructional decisions.

One possible explanation for the variation is that students are consistently gaining ground on particular objectives as teachers implement revised strategies for instruction. To examine this question, we examined student performance on the objectives individually by grade over three years. Our analyses revealed little consistency in results. From 1999-2000, over student performance measures on six Objectives increased, while performance decreased on the remaining seven Objectives. From 2000-2001, student performance increased on two Objectives and decreased on the remaining eleven. These data suggest that variation in student performance by Objective was due to improvements over time. Below, we provide the trend charts for Objectives 4, 5, and 11 for all grades as illustrative of the variation found.

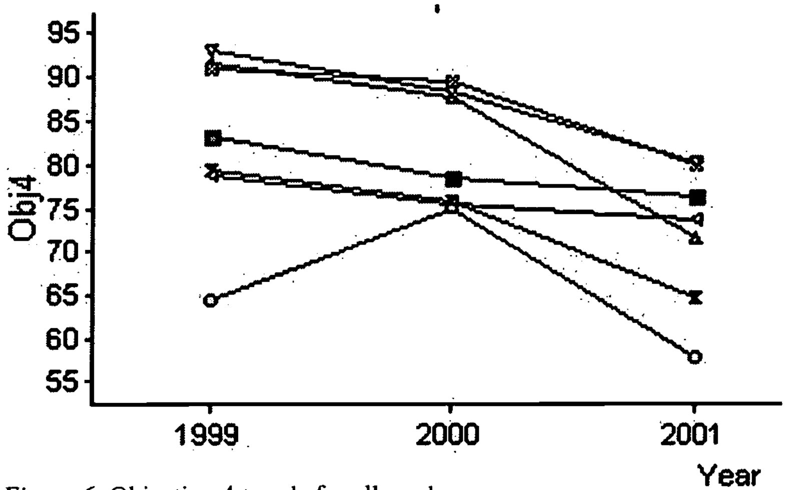

Figure 6. Objective 4 trends for all grades.

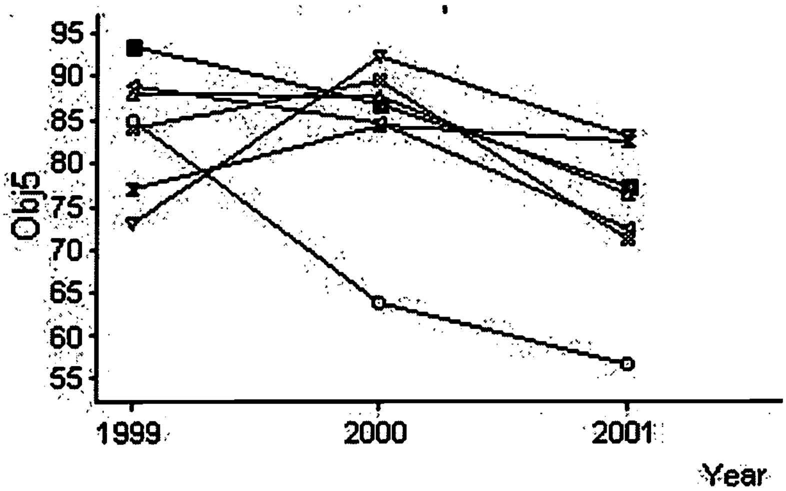

Figure 7. Objective 5 trends for all grades.

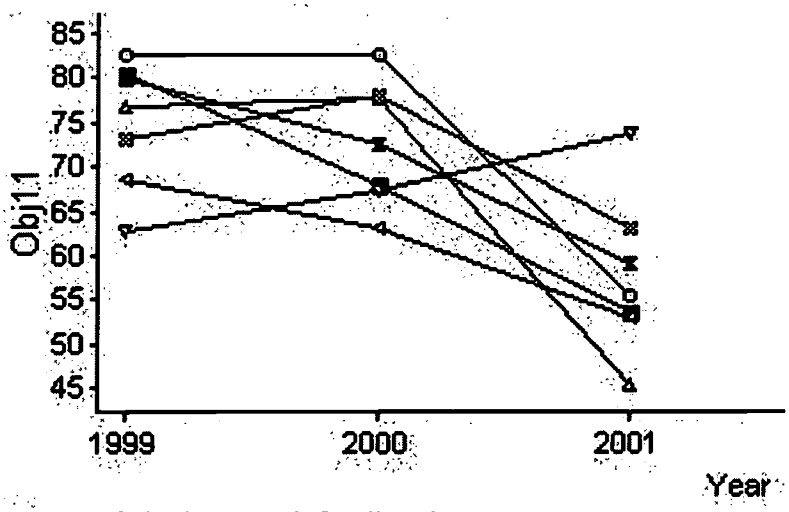

Figure 8. Objective 7 trends for all grades.

This analysis demonstrates that teachers cannot simply use changes in their mean student performance by objective to guide their planning. This seems obvious as one recognizes that the raw scores by objective are the product of not only the content of the objective but also the sampling of topics and the difficulty of the items and distractors. This led us to our next protocol for analysis.

Possible Content Factors Affecting Student Performance by Sampling and Difficulty

Returning to content analytical methods, we tapped into another possible source of information for teachers that could be useful in examining the variation in sampling of subtopics in each objective. TEA outlines which TEKS standards are being aligned with each TAAS objective. We created a table outlining the items clustered under each TAAS Objective. Within each TAAS objective, we identified which associated TEKS standard(s) each Objective tested (TEA, 2000). Each clustered item within a TAAS Objective was aligned with an associated TEKS standard (for example, see Figure 9). Our conjecture was that too little variability would permit the test to lose validity as teachers could virtually drill students on likely topics for inclusion. We expected that a valid test over time would sample proportionately across subtopics. We also expected that if teachers were assuming consistency in certain items, changes in the sampling of those objectives would lead to drops in performance.

Continuing this process for all the TAAS Objectives, we quantified the amount of variability in item sampling between 1999, 2000 and 2001. For example, since two items in 2000 are sampled from a new topic while two items remained from the old topic, we coded this as a 50% change. Eight item objectives can produce changes such

TEKS standards aligned with TAAS Objective 1: Number Concepts	1999	2000	2001
(1)(A) compare and order rational numbers	2, 15	2, 5	8, 20
(1)(C) approximate the value of irrational numbers	8, 17		
(1)(D) express numbers in scientific notation, including negative exponents		10, 13	3
(d)(3)(A) use patterns to generate the laws of exponents and applies them			16

Figure 9. TEKS to TAAS alignment for objective 1.

as 37.5%. Likewise, one item in 2001 differs in alignment with the 2000 test. This constitutes a one out of four or 25% shift in content emphasis. Overall, we calculated the percent change in content alignment from 1999 to 2000 and then from 2000 to 2001 (see Figure 10).

Obj	1		2		3		4		5		6	
Years	99/00	00/01	99/00	00/01	99/00	00/01	99/00	00/01	99/00	00/01	99/00	00/01
% change	50	25	75	25	100	50	100	50	0	50	0	0

Obj	7		8		9		10		11		12		13	
Years	99/00	00/01	99/00	00/01	99/00	00/01	99/00	00/01	99/00	00/01	99/00	00/01	99/00	00/01
% change	25	0	0	0	0	0	50	25	37.5	12.5	62.5	37.5	0	0

Figure 10. Percent change in content alignment.

Our initial analysis showed that from 1999 to 2000 and 2000 to 2001, five of thirteen objectives were unchanged. For 1999-2000, an additional four objectives and for 2000-2001, an additional seven objectives showed under fifty percent changes, showing only four objectives in the first year and no objectives in the second year were altered more than fifty percent. We examined the variation in student performance to see if there was a simple relationship between student performance and variation in sampling (see Figure 11). There was not. This is an area that requires a more sophisticated form of analysis.

Our final analysis included an attempt to quantify difficulty in three ways: task consistency, task characteristics, and distractors. Within task consistency, we considered whether the content being tested was comparable between years. When the content topic was the same, we noted changes in task characteristics including number type, language use, single-step versus multi-step procedures, and the use of the money context. Distractor analysis was performed at the level of identifying answer choices that could easily be disregarded an/or common misconceptions. We identified the Objectives that a) appeared to have task consistency, were amenable to test preparation, and should produce relatively stable performance; b) objectives with varied difficulty in the items (harder or easier) across years and c) objectives with varied items

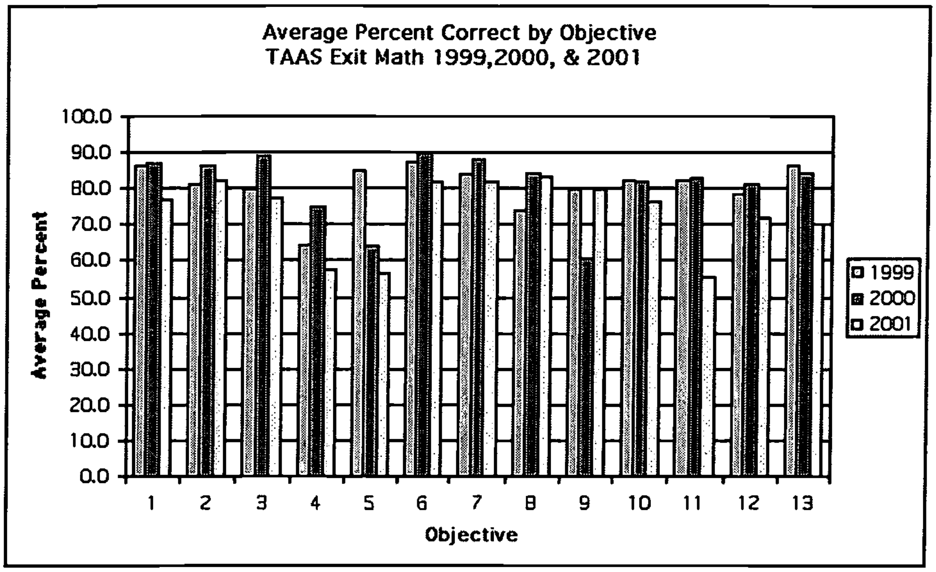

Figure 11. 10th grade student performance by objective.

and incommensurability in assessing difficulty. We compared our prediction and the item analysis provided by TEA of actual results of the percentage of students passing.

The results of our analyses showed the unreliability of making predictions based on our measures of item difficulty. We predicted that the 2000 test would be more difficult on all objectives based on our measures of difficulty except for Objectives 6 and 7 that we predicted would remain the same. According to the chart, we were correct for Objectives 5 and 9. Other results showed we were incorrect or only minimally correct for Objectives 10 and 13. Between 2000 and 2001 we had similar results. This is not surprising as test equating on TAAS is actually done using a Item Response Theory approach based in Rasch analysis. The difficulty levels are not publicly released.

Based on our analysis of TAAS for tenth graders, we found it difficult to draw any conclusions about student performance and improvement based on objective level analysis. These results suggest serious doubts about how teachers, after quantifying such results, are supposed to use this information to make a judgment about performance and thereby influence their instructional decision-making.

Conclusions

Our research was focused on examining whether the data provided to teachers for instructional decision-making were valid for this purpose. We worked to extend the analysis of the data in relation to the published test structure, and found that there were a number of problems in reconciling the results. The problem solving objectives obscured the underlying content dimensions making it difficult to judge the relative needs of students as regards topics. The variability in the student raw scores by objectives makes it unlikely that variations in student performance year to year represent real changes in student knowledge. A lack of trend data by objective suggests this variability is unlikely to represent systematic improvements in instruction. Finally, it did not appear that one could easily construct a content valid analysis of changes in difficulty in terms of topic selection, item characteristics or distractors.

In future work, we plan to continue to work to develop protocols that can validly guide teachers in undertaking content analyses of test results that can inform instructional decision-making. We hope to be able to link such analyses with the methodologies of test equating to determine whether there is a way to resolve the competing influences of psychometric test analysis and score preparation and content analyses. We encourage our colleagues in mathematics education to become similarly involved in close analysis of content dimensions of testing, to ensure that reform efforts at the curricular and instructional level are consistent with the messages given teachers from high stakes tests.

References

Confrey, J., & Carrejo, D. (2002). Can high-stakes testing in Texas improve instruction? To appear in the proceedings of the 25th annual Philosophy of Mathematics Education Conference. Athens, GA.

Confrey, J., & Makar, K. (2002). *Developing teachers' statistical inquiry through immersion in high-stakes accountability data.* This volume.

Heubert, J.P. & Hauser, R.M. (Eds.). (1999). *High stakes: Testing for tracking, promotion, and graduation.* Washington, DC: National Research Council.

Texas Education Agency. (2000). *Educator's guide to TEKS-based assessment: Exit level TAAS 1999-2000.* Student Assessment Division. Available: http:// www.tea.state.tx.us [2002, May].

Texas Education Agency. (2001). *Test development process.* Division of Student Assessment. Available: http:// www.tea.state.tx.us/student_assessment/develop/ devproc.html [2002, May].

NOTICE

Reproduction Basis

X	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").