Problem Choice by Test Takers: Implications for Comparability and Construct Validity

CSE Technical Report 485

Robert L. Linn, Damian W. Betebenner, and Kerry S. Wheeler
Center for Research on Evaluation, Standards,
and Student Testing (CRESST)/
University of Colorado at Boulder

September 1998

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

Project 2..2   Comprehensive Systems for Accountability and the Measurement of Progress   Robert L. Linn, Project Director, CRESST/University of Colorado at Boulder

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B60002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education, and by the Oregon Assessment Development and Evaluation Project (OADEP) PR/Award number R279A50027

The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education, or of the Oregon Departmment of Education.

PROBLEM CHOICE BY TEST TAKERS: IMPLICATIONS FOR COMPARABILITY AND CONSTRUCT VALIDITY[*]

Robert L. Linn, Damian W. Betebenner, and Kerry S. Wheeler
CRESST/University of Colorado at Boulder

Abstract

For assessments that present problems that require extended responses and substantial amounts of time, there is often a desire to allow students to choose which problem they will respond to among two or more options. Student choice of problem has the appeal that it may allow students a better opportunity to demonstrate what they know and are able to do. On the other hand, choice raises questions about the comparability of scores obtained by students who respond to different problems. Questions of comparability and validity of scores obtained when students are given a choice among alternative problems was investigated using data from the Oregon State Assessment Program Grade 10 Mathematics Assessment administered in the Spring of 1997. The assessment consisted of a multiple-choice section and a pair extended-response problems. On each of six alternate forms, two problems were presented and students were instructed to choose one to complete. Data from the six forms were analyzed to evaluate the comparability of scores obtained from responses to different tasks and the validity of the results. It was found that problems differed in popularity and that the scores students obtained differed systematically as a function of problem choice. On the other hand, confirmatory factor analysis results across forms for students choosing different problems suggest that there was similar validity for measuring the underlying constructs across problem choice. It was concluded that while choice may be justified, some form of equating adjustments would be needed before making high-stakes decisions based on performance of students on problems where choice is allowed.

Examinations that allowed test takers to choose a subset of potential problems to answer were common in the early part of this century. Wainer and Thissen (1994) have noted, for example, that most of the thirteen examinations offered by the College Entrance Examination Board in 1905 provided test takers with some degree of choice (e.g., answer 7 of 10 problems). The provision of choice continued for a

[*] Data for the research reported in this report were provided by the Oregon Department of Education.

number of years, but was gradually phased out so that only 3 of 14 examinations offered in 1941 gave test takers any choice of problems. Furthermore, the choice for those three examinations was fairly restricted (e.g., three mandatory essay problems and three pairs of parallel problems where test takers were instructed to choose one problem from each pair to answer) (Wainer & Thissen, 1994).

The widespread use of selected-response testing, of course, reduced some of the motivation for choice among problems. With fifty items, each requiring only a minute or two to answer, there is much less concern that the luck of the draw will deny a student a reasonable opportunity to show what he or she knows than when there are only a half a dozen problems to be answered.

Although the dominance of the multiple-choice format was one factor that led to a reduction in the use of choice among alternate problems on widely used tests, it was not the only factor. The emphasis on standardization, reliability, and comparability within and across test forms all pushed in the direction of asking students to respond to the same set of items. Choice among problems turns a single examination into multiple examinations, each defined by a unique combination of problems selected by test takers. Problem difficulty and the ability of test takers choosing different problems are confounded, which makes adjustments through standard equating methods problematic. Because of the lack of a satisfactory way of adjusting for differences in difficulty of the problems test takers choose to answer, Gulliksen (1950, p. 338) advised that test taker choice among "questions should *always* be avoided" (emphasis added; see Wainer & Thissen, 1994, for a more extended discussion).

Gulliksen's advice on choice became a widely accepted part of the instruction in good testing practice. Because choice undermines comparability, it was viewed as a threat to fairness.

In recent years, the dictum against test taker choice, like many other long-standing canons of good testing practice, has been challenged. The resurgence of extended essays and other time-consuming performance assessment problems has led to increased use of problem choice in a number of assessment programs. The necessary reduction in the number of problems on performance assessments increases the concern that a test taker may be penalized by the particular problems selected for the assessment. On a two- or three-problem examination, for example, there is a noticeable chance that a student who could solve 40 of the problems in a pool of 50 mathematics problems would encounter only problems that he or she could not solve.

The idea that choice will increase the likelihood that test takers will be able to demonstrate what they know and are able to do is one motivation for the re-introduction of test taker choice on examinations. A possibly more important motivation, however, is based on changes in conceptions of learning and cognition. The "less-is-more" philosophy favors in-depth study of a smaller number of topics over broader coverage in less depth. The combination of a desire for more focused, in-depth study with the press for better alignment of assessment with instruction leads to a desire for increased reliance on extended, performance-assessment tasks that probe students' understanding and ability to use their knowledge to solve complex problems.

If all students being assessed had the same focus for their in-depth study, the selection of assessment task topics for extended responses would be straightforward. Of course, such a common focus cannot be assumed for large-scale assessments. Problem choice by test takers is one approach to increasing the likelihood that students will be able to respond to problems that are more closely aligned with the topics they have had an opportunity to study. Thus, problem choice can be seen as an attempt to enhance validity. That is, if the goal is to make inferences about what students know and can do related to topics that they have had an opportunity to study, then problem choice may improve validity by increasing the likelihood that there is better alignment between the problems students respond to and the content areas they study. Thus, the question is whether such potential benefit outweighs threats to validity resulting from reduced technical characteristics such as reliability and comparability. The purpose of this study is to investigate the benefits and threats to validity of problem choice for a statewide assessment in mathematics.

Recent Research on Problem Choice

Over the past decade, research on problem choice remains rather limited with only a handful of researchers investigating the difficulties of incorporating problem choice into assessments. Building on research conducted at Educational Testing Service, Wang, Wainer, and Thissen (1995) developed an experiment in which they allowed students to indicate their preference on each of three pairs of multiple-choice items on a 20-item test. After indicating their preference on each pair, students were then required to answer both items of each pair. The question guiding the authors was whether or not it is possible to equate test forms generated by examinee choice. Such equating, the authors argue, is necessary to the production of tests that are fair regardless of chosen items.

A strong assumption underlying the equating is that a test taker's choice not to respond to a particular choice-question is statistically irrelevant after conditioning on examinee ability (i.e., ignorable non-response conditional on proficiency). In an item response theory context, this assumption implies that the item response curve for test takers who initially chose the item should be the same as the curve for those answering it after indicating a preference for the other item in the pair. Wang et al. (1995) found this to not always be the case. Choice did make a difference between the two groups. As such, their conclusions are discouraging: In order for tests be fair, "choice is either unnecessary or impossible" (p. 224).

In a non-experimental study using constructed response items allowing examinee choice, Fitzpatrick and Yen (1995) came to somewhat less pessimistic conclusions with respect to the inclusion of choice items. Using third-, fifth- and eighth-grade reading assessment data from the Maryland School Performance Assessment Program the researchers compared items across seven statistical categories including number and percentage of students selecting each choice and item-test correlations. With respect to the choice items, their results suggest that differences in difficulty and item-test correlations cannot be attributed to choice. Furthermore, the scores given on the different sets of choice items were comparable when these choice items were scaled together with the non-choice items given to all students. The authors emphasize that the non-choice items serve a crucial function in allowing the difficulty of the choice items to be statistically separated from examinee ability. They acknowledge, however, that given the nature of their study, determination of how wisely students chose is not possible.

Proponents of problem choice argue that choice should increase the fairness of a test by allowing students who have different academic backgrounds to demonstrate their knowledge better by answering questions that might be the most pertinent to what is covered in their specific district or classroom. Students can still demonstrate their knowledge about history by choosing a question that best fits the locally guided curriculum. Some research suggests, however, that choice may reduce rather than enhance fairness by undermining the validity of the test. Using data from the Advanced Placement (AP) Chemistry test, Wainer, Wang, and Thissen

(1994) found that most essay questions were not equal in terms of difficulty. Indeed, it is almost impossible for test publishers to produce essay questions of identical difficulty. They also concluded that tests requiring item choice not only measured the student's ability in the given subject area but also measured the student's conception of his or her ability to pick the easiest item to answer. That is, choice may reduce validity by introducing a construct-irrelevant (Messick, 1989) source of variance in the test scores.

In their article reviewing examinee choice, Wainer and Thissen (1994) examined whether some groups were prone to picking more difficult items than other groups. They found gender differences in choice on the tests that they analyzed. For both the AP Chemistry test and the AP United States History test, women were more likely than men to pick items of greater difficulty. They concluded that men had an advantage on these tests not because they were better at chemistry or history but because they chose easier items. In these instances, Wainer and Thissen concluded that tests allowing choice or problems to answer were biased against certain groups of students.

The Wainer and Thissen (1994) conclusion leaves test developers with a dilemma. Should tests be equated to produce equal difficulty or should test publishers simply assist students in making better choices? Statistical equating using Item Response Theory (IRT) accomplishes to some degree the goal of achieving comparability. Assisting students to make better choices may or may not enhance comparability. Clear directions are, of course, desirable in any case, but the test taker goal of achieving the highest score is not necessarily identical with the goal of maximum comparability of scores based on different problems for different students.

Method

Oregon State Assessment Program. In 1995 the Oregon Legislative Assembly passed amendments to the Oregon Educational Act for the 21st Century that changed the purpose and design of the Oregon State Assessment Program (OSAP). Prior to that legislation, the OSAP was intended for purposes of general program evaluation and public reporting for school buildings, districts and the state, and for reporting general individual-student results to students, parents and teachers. The 1995 amendments changed the OSAP to a standards-based system with a "focus on determining if students have met the standards established for the Certificate of

Initial Mastery at approximately grade 10" (Oregon Department of Education [ODE], 1996, p. 1). In addition to assessments used to determine whether students meet the Certificate of Initial Mastery (CIM) standards at Grade 10, the revised OSAP includes "benchmark" assessments and standards to be administered at Grades 3, 5, and 8 to evaluate the progress students are making toward achieving the CIM standards. Grade-12 assessments and standards are also planned. Students who meet the Grade 10 standards will receive Certificates of Initial Mastery, and Certificates of Advanced Mastery will be awarded to students who meet the Grade 12 standards. The Oregon plan also specifies that students who fail to meet the benchmark standards in the earlier grades (3, 5, and 8) will receive additional services.

The OSAP system comprises three forms of assessment: "content-based assessments, performance-based assessments, and work samples" (ODE, 1996, p. 1). The content-based and performance-based assessments are developed, administered, and reported by the state; the work samples are the responsibility of local districts. The content-based assessments are multiple-choice tests intended to measure "a student's understanding of a predetermined body of knowledge" (ODE, 1996 p. 1). Those tests are placed on a common scale across Grades 3 through 10. The performance-based assessments require students to construct responses to problems intended to measure "a student's ability to use knowledge and skills to create a complex or multi-faceted product or complete a complex task" (ODE, 1996, p. 1). In addition to a 3-point accuracy score (*precisely correct*, *essentially correct*, or *not correct*) the responses to the performance-based assessments are scored using 6-point scoring guides on each of four dimensions (1. Conceptual Understanding, 2. Processes and Strategies, 3. Communication, and 4. Verification). See Table 1 for brief descriptions of the four dimensions. In order to meet the standard on the performance-based assessment, a student must have an accuracy rating of either precisely or essentially correct *and* receive a score of at least 4 on *each* of the four dimension scores.

According to the scoring guide, answers that give a mathematically justifiable solution are scored "precisely correct" whether or not the work supporting that solution is provided by the student. An "essentially correct" score is given if the answer "would have been precisely correct had it not been for a minor error. No additional instruction appears necessary" (ODE, 1997, p. 1). If the answer is not mathematically justifiable and not simply the result of a minor error, the response is

Table 1

Descriptions of OSAP Dimension Scores for Mathematics Problem Solving

Dimension	Description
Conceptual Understanding	"Showing an understanding of the mathematical concepts related to the task (the 'what')"
Processes and Strategies	"Choosing strategies that can work, and then carry out the strategies chosen (the 'how')"
Communication	"Showing the reasoning (the 'why') behind the process, using diagrams, symbols, and/or vocabulary"
Verification	"In addition to solving the task, reviewing the work & showing that the solution process is reasonable in relation to the task (the 'proof')"

Note. Descriptions taken from 1997-98 *Official Scoring Guide* (ODE, 1997).

given an accuracy score of "not correct." The 6-point dimension scores are labeled (6) *extraordinary achievement*, (5) *thoroughly developed*, (4) *work is complete and effective*, (3) *work is partially effective or complete*, (2) *work is inappropriate or partially ineffective*, (1) *work is flawed*, and (*NE*) *no evidence*. The scoring guide elaborates each of these score points on each dimension. For example, a score of 4 on Conceptual Understanding is described as follows: "The task is translated into adequate mathematical concepts using relevant information and/or data from the task" (ODE, 1997, Official Scoring Guide).

Test takers were instructed that the people reading their responses would be looking for the following things: (1) "How well you *understood the problem* and the kind of mathematics you used. (2) How well you carried out the problem solving strategy. (3) How well you communicated your mathematical reasoning in arriving at your solution. (4) How you reviewed your work (or solved the problem in a second way) to be sure it made sense and was accurate" (ODE, 1996-97, emphasis in original). Test takers were also given a checklist to help guide their completion of the task and to encourage attention to the four dimensions. For example, one of two checklist items under Communication was "I explained what I was thinking while working the problem, including using pictures, charts or diagrams to help explain 'the why' of my steps" (ODE, 1996-97). No advice was given in the written directions on strategies for choosing which of the two alternative problems to answer. Test takers were simply told: "Choose one problem from the two below to complete" (ODE, 1996-97).

Data set. The data for the present study are drawn from the first statewide, operational administration the OSAP Grade 10 mathematics assessment in the spring of 1997. Results for the Grade 10 reading assessment are also included in some of the analyses reported below, but the focus is on the mathematics assessment, particularly the performance-based portions of that assessment.

Six forms of the mathematics assessment were administered to Grade 10 students. Each form contained a pair of performance assessment problems where students were instructed to "Choose one problem from the two below to complete." The pairs of problems on each of the six forms are briefly described in Table 2. As seen in Table 2, the first three forms present a pair of problems from the same content area (Form A: Geometry, B: Algebra, and C: Probability). On the last three forms the first problem from one of the first three forms is paired with the first

Table 2

Problem Pairs From Which Student Choose One on the Six Forms of the Grade 10 Oregon 1997 Mathematics Assessment

Form	Alternate problems	Problem content area and description
A	1. Pizza	Geometry. Round vs. Square pizza value
	2. Box	Geometry. Maximum volume of box from square piece cardboard
B	3. Marathon	Algebra. Time to Prepare for marathon given training plan
	4. Car	Algebra. Speed increase as function of increase in stopping distance
C	5. Target	Probability. Likely number of shots at target to obtain fixed score given probabilities of outcomes
	6. Probability square	Probability. Likely score based random falls on to square given scores for different areas
D	5. Target	Probability. (see description above)
	3. Marathon	Algebra. (see description above)
E	3. Marathon	Algebra. (see description above)
	1. Pizza	Geometry. (see description above)
F	5. Target	Probability. (see description above)
	1. Pizza	Geometry. (see description above)

problem from another of the first three forms, thereby presenting students with a choice of problems from different content areas (e.g., probability or algebra on Form D).

Questionnaires. Students were asked to complete five survey questions associated with the performance assessment problems in mathematics. The survey questions asked students (a) how they decided which problem to solve, (b) in what mathematics course they were enrolled, (c) their self-appraisal of how good they were at solving the kind of problems on the assessment, (d) whether they had used a scoring guide like the one with the four dimensions before, and (e) how often they had practiced the kind of problem found on the assessment. Students chose from among four or five options for each of the survey questions.

Sample. The analyses reported below are based on data for roughly 30,000 students with scores on both the content-based and performance-based assessments in mathematics. There is a minor variation in the number of students for the different analyses. For example, scores on the content-based assessment were available for a total 31,212 students who also attempted one of the performance assessment problems, and that is the sample size used for analyses found in Tables 3, 4 and 5. Most of the analyses of performance assessment problems were limited to students who had scores of 1 to 6 on the Conceptual Understanding, the Processes and Strategies, and the Communication dimension scores (i.e., students with scores of NE [no evidence] on one of these dimension scores were excluded). The elimination of students with one or more scores of NE on these dimension scores resulted in a total sample of 27,770 across the six forms. The latter sample is the one that was used for analyses reported in Table 7 and subsequent tables.

Analyses. Descriptive statistics were computed by test form for the total sample and for subgroups defined by selected demographic variables (gender, racial/ethnic group, and socio-economic status). Descriptive statistics on all parts of the assessment were also computed by problem choice on each form. Correlational analyses were conducted to examine the relationship among dimension scores, scores on different parts of the assessment (content-based and performance-assessment sections), and responses to the survey questions.

For each form, a multivariate analysis of covariance was conducted using the mathematics performance-assessment dimension scores on the problems as the dependent variables; scaled scores on the content-based section of the mathematics assessment as covariates; and problem choice and gender as the factors. Discriminant analyses were also conducted to determine the degree to which scores on the content-based section of the assessment and parents' education distinguished groups of students choosing different problems to complete on each form.

Finally, a series of structural equation analyses were conducted to compare the underlying factor structure of the scores on the mathematics content-based assessment and the dimension scores on the mathematics performance assessment problems.

Results

Descriptive statistics. The total scale score means and standard deviations on the content-based assessment in mathematics are listed separately in Table 3 for students choosing each alternative performance assessment problem on each of the six forms. Also shown in Table 3 is the number of students choosing each problem and the effect size for each form where effect size is defined as the scale score mean for students choosing the second problem in a pair minus the mean for those choosing the first problem all divided by the standard deviation of students choosing the first problem.

As can be seen in Table 3, the first problem was selected by a majority of students on each of the six forms. However, the size of the majority varied substantially from form to form (from a low of 51% choosing the first problem on Form D to a high of 72% choosing the first problem on Form C). For the repeated problems the percentage of students changed from one form to another. The Pizza problem, for example, was selected by 63% of the students administered Form A compared to 46% and 44% on Forms E and F respectively. Since both position (first or second problem) and the context of the alternate problem differed from form to form for the Pizza problem, it is unclear how much of the difference in popularity is due simply to position and how much to context.

For five of the six forms, the groups of students choosing different problems have significantly different ($p < .05$) means on the content-based assessment scale scores. The absolute value of the effect size on the scale scores for groups choosing to respond to different performance problems ranged from .09 to .34. A positive effect size indicates a higher mean for the group choosing the second (less popular) problem of a pair. For five of the six forms, the group choosing the second problem in a pair had the larger mean (significantly larger in four of those cases). For Form E,

Table 3

Number of Students Choosing Each Alternate Performance Problem and Scale Score Means, Standard Deviations and Problem Effect Size on Content-Based Assessment for Student Choosing Each Problem

		Number and (percent) selecting problem		Math scale score statistics		
Form	Problem			Mean	Standard deviation	Effect size*
A	1 Pizza	2992	(63)	233.0	10.6	
	2 Box	1791	(37)	235.1**	9.9	.20
B	3 Marathon	3619	(55)	233.2	10.4	
	4 Car	2907	(45)	234.4	10.3	.12
C	5 Target	3396	(72)	232.4	9.9	
	6 Prob Square	1347	(28)	235.8**	11.0	.34
D	5 Target	2325	(51)	232.2	9.8	
	3 Marathon	2198	(49)	235.2**	10.0	.31
E	3 Marathon	2339	(54)	234.9	9.6	
	1 Pizza	2019	(46)	232.5**	11.1	-.25
F	5 Target	3536	(56)	233.4	9.5	
	1 Pizza	2743	(44)	234.3**	11.3	.09

* Effect size equals the scale score mean for students choosing the second problem in a pair minus the mean for those choosing the first problem all divided by the standard deviation of students choosing the first problem.

** Means for students choosing the two problems are significantly different ($p < .05$).

however, the group choosing the first problem had a significantly higher mean (effect size = -.25). Although the effect sizes are of small to moderate size, it is clear that it cannot be safely assumed that the groups choosing different problems are equal in terms of the mathematics content knowledge.

Tables 4 and 5 report results parallel to those in Table 3 separately for male and female students. There are some differences in popularity of problems as a function of gender (e.g., males are somewhat more likely to choose the "Target" problem in all three forms in which it appears than females are), but the general tendency to choose the first problem rather than the second problem holds for both males and females. There are small differences between males and females in mean scores and in effect sizes, but the general finding that the content-based assessment scores have different means for students who choose different problems is common for both males and females.

Table 4

Number of Male Students Choosing Each Alternate Performance Problem and Scale Score Means, Standard Deviations and Problem Effect Size on Content-Based Assessment for Student Choosing Each Problem

Form	Problem	Number and (percent) selecting problem	Math scale score statistics		
			Mean	Standard deviation	Effect size*
A	1 Pizza	1349 (60)	234.2	11.0	
	2 Box	889 (40)	235.3	10.4	.10
B	3 Marathon	1557 (51)	233.8	10.8	
	4 Car	1474 (49)	235.4**	10.8	.15
C	5 Target	1641 (73)	233.0	10.1	
	6 Prob Square	615 (27)	237.9**	11.4	.49
D	5 Target	1188 (55)	232.7	10.4	
	3 Marathon	956 (45)	236.9**	10.4	.40
E	3 Marathon	1066 (52)	235.7	10.0	
	1 Pizza	1001 (48)	233.4**	11.7	-.23
F	5 Target	1730 (59)	233.5	9.9	
	1 Pizza	1213 (41)	236.0**	12.1	.25

* Effect size equals the scale score mean for students choosing the second problem in a pair minus the mean for those choosing the first problem all divided by the standard deviation of students choosing the first problem.

** Means for students choosing the two problems are significantly different ($p < .05$).

Students receiving scores of NE (no evidence) were excluded from the analyses involving dimension scores. For the Conceptual Understanding, Processes and Strategies, and the Communication dimension scores the number of students who received scores of NE ranged from 22 to 181 and was in most cases less than 5% of the students responding to a particular problem (see Table 6). As shown in Table 6, however, on the Verification dimension, a much larger fraction of the students (sometimes over half of the respondents to a problem) received scores of NE. Because of the large number of NE scores on that dimension, Verification was excluded from most of the analyses reported below. This made it possible to limit the analyses to students who had scores of 1 to 6 on the other three dimension scores without a great loss in students who were included in the analyses.

Table 5

Number of Female Students Choosing Each Alternate Performance Problem and Scale Score Means, Standard Deviations and Problem Effect Size on Content-Based Assessment for Student Choosing Each Problem

		Number and (percent) selecting problem		Math scale score statistics		
Form	Problem			Mean	Standard deviation	Effect size*
A	1 Pizza	1420	(65)	232.4	10.2	
	2 Box	772	(35)	235.3**	9.1	.28
B	3 Marathon	1791	(60)	233.0	9.9	
	4 Car	1215	(40)	233.8	9.6	.08
C	5 Target	1478	(70)	232.4	9.6	
	6 Prob Square	634	(30)	234.6**	10.2	.23
D	5 Target	944	(47)	232.2	9.1	
	3 Marathon	1079	(53)	234.1**	9.5	.21
E	3 Marathon	1100	(56)	234.3	8.8	
	1 Pizza	875	(44)	232.0**	10.4	-.26
F	5 Target	1575	(54)	233.7	9.0	
	1 Pizza	1332	(46)	233.3	10.3	-.04

* Effect size equals the scale score mean for students choosing the second problem in a pair minus the mean for those choosing the first problem all divided by the standard deviation of students choosing the first problem.

** Means for students choosing the two problems are significantly different ($p < .05$).

Table 6

Number of Students With Dimension Scores of NE (No Evidence) by Form and Problem Choice

Form-problem	*N*	Conceptual understanding	Processes and strategies	Communication	Verification
A1	2,840	89	183	138	1,507
A2	1,704	60	121	124	812
B1	3,429	96	137	149	1,696
B2	2,762	57	94	114	1,207
C1	3,212	91	148	149	1,427
C2	1,287	22	38	41	521
D1	2,202	61	91	107	1,027
D2	2,084	41	56	61	974
E1	2,234	33	50	88	1,035
E2	1,924	72	131	89	1,147
F1	3,382	76	118	181	1,480
F2	2,610	88	145	141	1,365

The mean Accuracy scores and the means on the Conceptual Understanding, Processes and Strategies, and Communication dimension scores for the alternative performance problems are compared in Table 7. Also shown in Table 7 are the standard deviations, results of the *t*-tests for differences between the pairs of means and the effect size for each score. As in Table 3, the effect size is defined as the mean for students choosing the second problem in a pair minus the mean for those choosing the first problem all divided by the standard deviation for students choosing the first problem. As can be seen in Table 7, with few exceptions, the scores obtained by students choosing different problems differ significantly. Moreover, the effect sizes on the dimension scores are frequently larger than those shown in Table 7 for the content scale scores. In absolute value, the effect sizes for the dimension scores range from .00 to .55. The mean of the absolute values of the effect sizes for the accuracy score is .22. The corresponding means for the three dimension scores are .29 for Conceptual Understanding, .30 for Processes and Strategies, and .24 for Communication.

Tables 8 and 9 report the results parallel to those in Table 7 separately for male and female students. There is considerable similarity between the results for males and females in terms of significant differences and the instances where the largest effects sizes are obtained.

M**ANCOVA results**. The differences between groups choosing different problems on both the content-based scale scores (which are derived from common items and thus do not depend on problem choice) and on the performance assessment dimension scores (which clearly are problem specific) raise the question of the degree to which differences on the latter might be explained in terms of differences in the overall mathematics achievement of groups that choose different problems. To explore this issue, a multivariate analysis of covariance (MANCOVA) was conducted for the accuracy and first three dimension scores for each form using the content-based scale score as the covariate. Gender and problem choice were used as the factors in the MANCOVAs. The multivariate *F* ratios for the tests of the significance of the covariate (i.e., the content-based assessment score), the two main effects (gender and problem choice), and the interaction of gender and problem choice are displayed in Table 10.

Table 7

Means, Standard Deviations and Effect Size on Performance Assessment Problems by Scoring Dimension, Form, and Problem Choice

		Means, (standard deviations), and [effect size]*			
Form	Problem (N)	Accuracy	Conceptual understanding	Processes and strategies	Communication
A	1 Pizza	.44	2.27	2.13	2.50
	(2,602)	(.75)	(1.33)	(1.38)	(1.15)
	2 Box	.71**	2.65**	2.50**	2.68**
	(1,523)	(.93)	(1.36)	(1.42)	(1.20)
		[ES=.36]	[ES=.29]	[ES=.27]	[ES=.16]
B	3 Marathon	.48	2.46	2.35	2.69
	(3,220)	(.76)	(1.29)	(1.34)	(1.09)
	4 Car	.68**	3.01**	2.88**	2.94**
	(2,606)	(.79)	(1.32)	(1.38)	(1.19)
		[ES=.26]	[ES=.43]	[ES=.40]	[ES=.23]
C	5 Target	.54	2.51	2.36	2.60
	(2,996)	(.84)	(1.07)	(1.14)	(0.97)
	6 Prob	.94**	3.06**	2.97**	3.09**
	Square	(.95)	(1.47)	(1.52)	(1.24)
	(1,228)	[ES=.48]	[ES=.51]	[ES=.54]	[ES=.51]
D	5 Target	.56	2.49	2.36	2.58
	(2,056)	(.85)	(1.05)	(1.13)	(0.96)
	3 Marathon	.56	2.66**	2.55**	2.80**
	(2,004)	(.79)	(1.30)	(1.35)	(1.08)
		[ES=.00]	[ES=.16]	[ES=.17]	[ES=.23]
E	3 Marathon	.52	2.58	2.49	2.77
	(2,158)	(.78)	(1.26)	(1.32)	(1.05)
	1 Pizza	.41**	2.24**	2.09**	2.47**
	(1,758)	(.72)	(1.35)	(1.38)	(1.15)
		[ES=-.14]	[ES=-.27]	[ES=-.30]	[ES=-.29]
F	5 Target	.60	2.57	2.45	2.68
	(3,201)	(.87)	(1.05)	(1.11)	(0.95)
	1 Pizza	.52**	2.46**	2.29**	2.64
	(2,418)	(.80)	(1.37)	(1.42)	(1.20)
		[ES=-.09]	[ES=-.10]	[ES=-.14]	[ES=-.04]

* Effect size equals the scale score mean for students choosing the second problem in a pair minus the mean for those choosing the first problem all divided by the standard deviation of students choosing the first problem.

** Mean for second problem significantly different ($p < .001$) than the mean for the first problem.

Table 8

Means, Standard Deviations and Effect Size on Performance Assessment Problems for Males by Scoring Dimension, Form, and Problem Choice

		Means, (standard deviations), and [effect size]*			
Form	Problem (N)	Accuracy	Conceptual understanding	Processes and strategies	Communication
A	1 Pizza	.51	2.41	2.27	2.55
	(1,221)	(.79)	(1.36)	(1.39)	(1.13)
	2 Box	.74**	2.73**	2.55**	2.70**
	(771)	(.94)	(1.38)	(1.44)	(1.23)
		[ES=.29]	[ES=.24]	[ES=.20]	[ES=.13]
B	3 Marathon	.55	2.53	2.43	2.64
	(1,426)	(.80)	(1.33)	(1.38)	(1.12)
	4 Car	.75**	3.06**	2.91**	2.93**
	(1,382)	(.80)	(1.34)	(1.40)	(1.21)
		[ES=.25]	[ES=.40]	[ES=.35]	[ES=.26]
C	5 Target	.59	2.56	2.41	2.59
	(1494)	(.86)	(1.10)	(1.17)	(0.97)
	6 Prob	1.10**	3.26	3.17**	3.15**
	Square	(.95)	(1.49)	(1.52)	(1.24)
	(583)	[ES=.59]	[ES=.64]	[ES=.65]	[ES=.58]
D	5 Target	.62	2.52	2.40	2.57
	(1,097)	(.87)	(1.27)	(1.15)	(0.96)
	3 Marathon	.67**	2.83**	2.71**	2.86**
	(912)	(.83)	(1.30)	(1.36)	(1.07)
		[ES=.06]	[ES=.24]	[ES=.27]	[ES=.30]
E	3 Marathon	.58	2.66	2.58	2.75
	(1,013)	(.79)	(1.28)	(1.34)	(1.06)
	1 P izza	.48**	2.37**	2.20**	2.49**
	(910)	(.76)	(1.38)	(1.41)	(1.14)
		[ES=-.13]	[ES=-.23]	[ES=-.28]	[ES=-.25]
F	5 Target	.67	2.63	2.50	2.64
	(1,597)	(.90)	(1.05)	(1.12)	(0.94)
	1 Pizza	.63	2.69**	2.50	2.70
	(1,115)	(.84)	(1.39)	(1.44)	(1.19)
		[ES=-.04]	[ES=-.06]	[ES=.00]	[ES=.06]

* Effect size equals the scale score mean for students choosing the second problem in a pair minus the mean for those choosing the first problem all divided by the standard deviation of students choosing the first problem.

** Mean for second problem significantly different ($p < .001$) than the mean for the first problem.

Table 9

Means, Standard Deviations and Effect Size on Performance Assessment Problems for Females by Scoring Dimension, Form, and Problem Choice

		Means, (standard deviations), and [effect size]*			
Form	Problem (N)	Accuracy	Conceptual understanding	Processes and strategies	Communication
A	1 Pizza	.38	2.16	2.03	2.45
	(1,318)	(.72)	(1.31)	(1.36)	(1.16)
	2 Box	.69**	2.58**	2.45**	2.55**
	(712)	(.92)	(1.35)	(1.40)	(1.13)
		[ES=.43]	[ES=.32]	[ES=.31]	[ES=.09]
B	3 Marathon	.42	2.53	2.29	2.73
	(1,713)	(.72)	(1.33)	(1.31)	(1.06)
	4 Car	.61**	2.96**	2.85**	2.96**
	(1,156)	(.77)	(1.30)	(1.35)	(1.17)
		[ES=.26]	[ES=.32]	[ES=.43]	[ES=.22]
C	5 Target	.50	2.46	2.33	2.61
	(1,414)	(.82)	(1.03)	(1.11)	(0.96)
	6 Prob	.79**	2.89**	2.80**	3.04**
	Square	(.92)	(1.42)	(1.49)	(1.23)
	(610)	[ES=.35]	[ES=.42]	[ES=.42]	[ES=.45]
D	5 Target	.48	2.46	2.32	2.60
	(892)	(.81)	(1.02)	(1.10)	(0.96)
	3 Marathon	.48	2.51	2.41**	2.76**
	(1,045)	(.75)	(1.27)	(1.32)	(1.09)
		[ES=.00]	[ES=.05]	[ES=.08]	[ES=.17]
E	3 Marathon	.47	2.51	2.42	2.79
	(1,078)	(.76)	(1.23)	(1.29)	(1.03)
	1 Pizza	.36**	2.12**	1.99**	2.45**
	(801)	(.69)	(1.32)	(1.35)	(1.17)
		[ES=-.14]	[ES=-.32]	[ES=-.33]	[ES=-.33]
F	5 Target	.53	2.52	2.41	2.71
	(1,530)	(.84)	(1.05)	(1.10)	(0.95)
	1 Pizza	.43**	2.29**	2.13**	2.60**
	(1,244)	(.75)	(1.33)	(1.39)	(1.20)
		[ES=-.12]	[ES=-.22]	[ES=-.25]	[ES=-.12]

* Effect size equals the scale score mean for students choosing the second problem in a pair minus the mean for those choosing the first problem all divided by the standard deviation of students choosing the first problem.

** Mean for second problem significantly different ($p < .001$) than the mean for the first problem.

Table 10

MANCOVA Results

Form (*df* error)	Source of variance	*F*	*P*
A (4,014)	Math content scores	903.25	.001
	Gender	4.51	.001
	Problem choice	20.06	.001
	Interaction	.62	.647
B (5,669)	Math content scores	1057.80	.001
	Gender	28.41	.001
	Problem choice	86.96	.001
	Interaction	3.52	.007
C (4,093)	Math content scores	624.23	.001
	Gender	15.33	.001
	Problem choice	31.97	.001
	Interaction	2.53	.039
D (3,938)	Math content scores	562.32	.001
	Gender	12.17	.001
	Problem choice	7.49	.001
	Interaction	1.76	.134
E (3,794)	Math content scores	783.79	.001
	Gender	11.67	.001
	Problem choice	21.05	.001
	Interaction	1.24	.293
F (5,478)	Math content scores	995.12	.001
	Gender	30.57	.001
	Problem choice	24.63	.001
	Interaction	3.54	.007

Note. *df* numerator equals 4 for all forms.

The highly significant *F* ratios for the math content scores were expected and simply reflect the fact that the content scores have substantial correlations with the performance assessment scores, which is the reason they were used as covariates. Although the interactions of gender and problem choice were statistically significant at the .05 level for three of the six forms, the magnitude of the interaction was quite small on all forms. This is reflected by the fact that with samples ranging in size from 3,784 to 5,669 none of the interaction *F* ratios exceeded 4.0. They ranged from .62 to 3.54 across the six forms. In contrast, the *F* ratios for the gender main effects, which is distributed on the same degrees of freedom as the interaction effects, ranged from a low of 4.51 to 30.57 (all significant at the .001 level). The corresponding range of *F* ratios was 7.49 to 86.96 for the problem-choice main effects, which also had the same degrees of freedom as the gender main effects and interaction effects. Thus, it is clear that these are highly significant differences in the scores obtained on the combination of Accuracy and dimension scores on the performance assessment problems as a function of problem choice even after adjusting for differences in the content-based scores obtained by students choosing different problems.

The adjusted means on the Accuracy and three dimension scores are shown in Table 11 for problem choice and gender main effects. As can be seen, the means adjusted for differences in the content-based scores are higher for students choosing problem 2 than for students choosing problem 1 on Forms A, B, and C, but the converse generally holds for Forms D, E, and F. From an inspection of the adjusted

Table 11

Adjusted Means From MANCOVA Analysis

		Problem choice		Gender	
Form	Dimension score	(1)	(2)	(F)	(M)
A	Accuracy	.48	.65	.55	.58
	Conceptual Und.	2.35	2.54	2.40	2.48
	Proc. and Strat.	2.22	2.38	2.27	2.33
	Communication	2.55	2.59	2.58	2.56
B	Accuracy	.50	.66	.54	.61
	Conceptual Und.	2.51	2.96	2.74	2.73
	Proc. and Strat.	2.40	2.84	2.63	2.60
	Communication	2.71	2.91	2.89	2.73
C	Accuracy	.58	.84	.66	.77
	Conceptual Und.	2.58	2.90	2.70	2.78
	Proc. and Strat.	2.45	2.80	2.59	2.65
	Communication	2.66	2.95	2.85	2.76
D	Accuracy	.61	.52	.51	.61
	Conceptual Und.	2.59	2.56	2.54	2.61
	Proc. and Strat.	2.47	2.45	2.43	2.48
	Communication	2.67	2.72	2.73	2.66
E	Accuracy	.49	.46	.45	.50
	Conceptual Und.	2.52	2.32	2.40	2.45
	Proc. and Strat.	2.43	2.18	2.29	2.32
	Communication	2.72	2.53	2.68	2.57
F	Accuracy	.63	.50	.51	.61
	Conceptual Und.	2.62	2.42	2.46	2.58
	Proc. and Strat.	2.50	2.25	2.32	2.42
	Communication	2.71	2.60	2.70	2.61

means for male and female students, it can be seen that males have higher adjusted means on Accuracy than females on all six forms. However, females have higher means than males on the Communication dimension for all six forms, and the differences are mixed across forms on the other two dimension scores.

The magnitude of the performance assessment mean differences between groups based on problem choice is generally smaller after adjusting for differences in the content-based scores. This can be seen in Table 12 where the problem-choice effect sizes for the performance assessment accuracy and dimension scores are presented before and after adjustments for the content-based scores. The mean absolute value of the effect sizes across the six forms is reduced by roughly 20% on the Accuracy score and approximately 30% to 40% on the three dimension scores. Nonetheless, the effect size favoring problem 2 is large enough on Forms A, B, and C to undermine the comparability of the scores of students choosing different problems on those three forms.

Table 12

Effect Sizes for Problem Choice on Performance Assessment Accuracy and Dimension Scores Before and After Adjustments for Differences on the Content-Based Mathematics Scale Score

	Performance assessment problems by scoring dimension							
	Accuracy		Conceptual understanding		Processes and strategies		Communication	
Form	Before	After	Before	After	Before	After	Before	After
A	.36	.24	.29	.14	.27	.12	.16	.03
B	.26	.21	.43	.36	.40	.33	.23	.18
C	.48	.31	.51	.30	.54	.31	.51	.30
D	.00	-.10	.16	-.02	.17	-.02	.23	.06
E	-.14	-.04	-.27	-.16	-.30	-.19	-.29	-.18
F	-.09	-.15	-.10	-.19	-.14	-.23	-.04	-.12
Mean absolute value	.22	.18	.29	.20	.30	.18	.24	.15

Note. The *Before* effect size equals the scale score mean for students choosing the second problem in a pair minus the mean for those choosing the first problem all divided by the standard deviation of students choosing the first problem. The *After* effect size equals the adjusted scale score mean for students choosing the second problem in a pair minus the adjusted mean for those choosing the first problem all divided by the standard deviation of students choosing the first problem.

Based on the accuracy and dimension scores obtained by students on the first operational administration of the performance assessment problems, the combined conjunctive rule for meeting the standard appears quite stringent. The rule for passing on the performance assessment part of the assessment is that a student obtain a score of 1 or higher on accuracy *and* a score of 4.0 or higher on *each* of the dimension scores. Even if the Verification dimension score is ignored because of the large percentage of students who received scores of NE for Verification (see Table 6), the passing rate would still be quite small when the combined passing rule is applied to the Accuracy and first three dimension scores. This can be seen in Table 13 where the number and percentage of students meeting the combined passing standard is presented for each problem and form. It should be noted that the total number of students for each form is somewhat larger for Table 13 than for Tables 6 and 10. The larger number of students is due to the inclusion of students with scores

Table 13

Number and Percentage of Students Meeting Combined Conjunctive Rule of an Accuracy Score of at Least 1.0 and Scores of At Least 4.0 on All Three Performance-Assessment Dimensions by Form and Problem Choice

Form	Problem	Number of students	Number meeting standard	Percent meeting standard
A	1 Pizza	3021	325	10.8
	2 Box	1801	300	16.7
	Form Total	4822	625	13.0
B	3 Marathon	3628	512	14.1
	4 Car	2927	677	23.1
	Form Total	6555	1189	18.1
C	5 Target	3427	236	6.9
	6 Prob Square	1356	403	29.7
	Form Total	4783	639	13.4
D	5 Target	2347	139	5.9
	3 Marathon	2222	334	15.0
	Form Total	4569	473	10.4
E	3 Marathon	2354	338	14.4
	1 Pizza	2045	229	11.2
	Form Total	4399	567	12.9
F	5 Target	3566	246	6.9
	1 Pizza	2764	407	14.7
	Form Total	6330	653	10.3

of NE in computing the percent passing, on the grounds that a student with a score of NE does not meet the criterion of a 1.0 or higher in Accuracy and a 4.0 or higher on each of the three dimension scores.

Across the six forms, the percentage of students meeting the combined passing criteria on the accuracy of dimensions scores ranged from 10.3% for Form F to 18.1% for Form B. Regardless of the passing standard for the dimension scores, questions may be raised both about the comparability of the performance-assessment passing rates across forms that present different pairs of problems and within forms based on the problem that a test taker chooses to answer. There was also a substantial difference in percent passing using these criteria as a function for problem choice on some of the forms. On Form C, for example, 29.7% of students responding to problem 2 met the passing criteria whereas only 6.9% of the students responding to problem 1 met the passing criteria.

Discriminant analyses. For each form, the groups for the discriminant analysis were formed by the performance assessment problem students chose to answer. Thirteen variables were used in identifying functions that discriminated between the two problem-choice groups on each form. These variables consisted of five content-based assessment mathematics subscores (calculation/estimation, measurement, statistics/probability, algebra, and geometry), three background variables (parent education, student racial/ethnic group, and gender), and four questionnaire items. The four questionnaire items asked students to indicate the mathematics class in which they were enrolled, whether they were "good at doing this kind of problem," whether they had used a scoring guide like the Oregon guide in responding to this kind of mathematics problem, and how often they had "practiced this kind of problem."

The discriminant function analysis chi-square statistics, canonical correlations of group membership with the performance assessment scores, and the discriminant function means at the group centroids are displayed in Table 14. Although the chi-square statistics are all significant at the .001 level, the canonical correlations range from .11 to .19 across the six forms. Thus, while there are significant differences in scores for students responding to the different problems on each form, there is considerable overlap in the scores.

Table 14

Chi Square Statistics, Canonical Correlations and Group Centroids for Discriminant Analyses

Statistic		Form A	Form B	Form C	Form D	Form E	Form F
Chi-square		59.75	59.81	129.78	130.81	90.69	64.08
Canonical correlation		0.13	0.11	0.19	0.19	0.16	0.12
Discriminant function means at group centroids	Problem 1	-0.10	-0.10	-0.12	-0.19	0.16	-0.10
	Problem 2	0.16	0.12	0.30	0.20	-0.18	0.13

The standardized discriminant function coefficients and correlations of the observed variables with the discriminant functions are reported in Table 15. Correlations greater than .40 in absolute value are shown in bold-face type in the bottom half of Table 15.

On Form A with two geometry problems, seven variables have correlations greater than .40 with discriminant function. These are the five content-based mathematics subscores, the level of the mathematics course in which the student was enrolled, and whether the student reported being good at this kind of problem. Students who chose the Box problem had a higher mean on all seven of these variables than students who chose the Pizza problem. On Form B with two algebra problems, students who chose the Car problem were more likely to be male (correlation of .73 with discriminant function), tended to be enrolled in a higher-level mathematics course, and tended to have relatively higher algebra, geometry, and measurement subscores (correlations of .50 to .57 with discriminant function) than students who chose the Marathon problem. The comparison of results for two probability problems on Form C shows that students who chose the Probability Square problem tended to have parents who had more education, were more likely to be enrolled in a higher level mathematics course, considered themselves good at this kind of problem, and had higher scores on all the content-based mathematics subscores than students who chose the Target problem.

Forms D, E, and F provided students with a choice of problems from two different content areas. Students who chose the algebra problem (Marathon) tended to have higher Algebra subscores (correlation of .71 with discriminant function) than students who chose the probability problem (Target). The discriminant function also has substantial correlations with each of the other four content-based mathematics

Table 15

Discriminant Coefficients and Combined Correlations of Observed Variables With Discriminant Functions

Variable	Form A	Form B	Form C	Form D	Form E	Form F
Standardized coefficients						
Parent Education	-0.11	-0.02	0.21	-0.14	0.06	0.11
Math Course	0.03	-0.27	0.04	-0.12	-0.54	0.19
Good at Kind of Problem	0.36	0.01	0.35	0.25	-0.01	0.63
Have Used Scoring Guide	0.01	-0.13	-0.03	0.01	-0.06	0.10
Practiced Similar Problems	0.05	0.11	0.01	-0.16	-0.03	-0.01
Calculations/Estimation	-0.09	-0.42	0.32	-0.01	0.42	-0.16
Measurement	0.11	0.14	-0.12	0.33	-0.30	0.18
Statistics/Probability	-0.16	0.04	0.09	0.00	0.38	-0.48
Algebra	0.48	0.25	0.22	0.35	0.13	0.37
Geometry	0.48	0.36	0.25	0.19	-0.22	0.43
Race	0.09	0.01	-0.04	-0.04	0.29	-0.16
Gender	0.26	0.73	-0.31	-0.56	-0.27	-0.55
Correlations with discriminant functions[a]						
Parent Education	0.20	0.19	**0.47**	0.14	0.33	0.24
Math Course	**0.44**	**0.47**	**0.52**	**0.57**	**0.74**	-0.19
Good at Kind of Problem	**0.63**	0.33	**0.56**	0.38	0.17	**0.64**
Have Used Scoring Guide	0.11	0.20	0.19	**0.71**	-0.18	-0.19
Practiced Similar Problems	0.11	0.15	0.09	**0.66**	0.01	0.07
Calculations/Estimation	**0.59**	0.28	**0.78**	**0.57**	**0.67**	0.24
Measurement	**0.66**	**0.50**	**0.64**	**0.67**	0.37	0.36
Statistics/Probability	**0.48**	0.36	**0.66**	**0.54**	**0.67**	0.05
Algebra	**0.78**	**0.52**	**0.76**	**0.71**	**0.62**	**0.44**
Geometry	**0.79**	**0.57**	**0.77**	**0.66**	**0.46**	**0.47**
Race	0.14	0.03	0.06	0.03	0.36	0.11
Gender	0.36	**0.73**	-0.19	**-0.45**	-0.24	**-0.41**

[a] Correlations greater than 0.40 in bold.

subscores as well as with the questionnaire items and with gender (females were more likely to choose the Marathon problem than the Target problem). On Form E, students who chose the geometry problem (Pizza) are distinguished from students who chose the algebra problem (Marathon) by a combination the mathematics course in which they were enrolled (correlation of .74 with discriminant function) and four of the content-based mathematics subscores. Finally, as might be expected, students who chose the geometry problem (Pizza) on Form F tended to have relatively higher Geometry subscores than students who chose the probability problem (Target).

Structural equation analyses. A total of 9 variables were used in most of the structural equation analyses. Included were (a) the five content-based mathematics assessment subscores (Calculation and Estimation, Measurement, Statistics and Probability, Algebra, and Geometry); and (b) the four performance-assessment mathematics scores (Accuracy, Conceptual Understanding, Processes and Strategies, and Communication).

The first comparison for each form was between the single-factor model depicted in Figure 1 and the two-factor model depicted in Figure 2. As can be seen in the figures, the two-factor model allowed for separate, but correlated, factors for the five content-based scores and the four performance assessment scores. For the single-factor model, the content-based scores and the performance assessment scores were forced to load on the same factor.

The one- and two-factor models were evaluated on a total of 12 samples formed by the combination of six forms and performance assessment problem choice for each form. The fit of the two-factor model was significantly better than the one-factor model in all cases. The one-factor model is a special case of the two-factor model with one less parameter corresponding to the correlations between the two common factors. Thus, the difference between the chi-square statistics for the two models is a chi-square with 1 degree of freedom. The values of those chi-squares for the differences were significant in all cases and ranged from 1,516.2 to 4,291.3.

The two-factor model shown in Figure 2 formed the basis for all subsequent analyses of the 9 mathematics scores. Given the two-factor model, the questions of primary interest for this study were (a) to what degree are the same constructs assessed when students choose different performance assessment problems to answer, and (b) to what degree are those constructs assessed with equal validity. These questions were addressed by comparing the loadings in the two-factor model for the two groups of students formed by their problem choice on a given form. The following analyses were replicated for each of the six forms of the mathematics assessment. In Model I, the Equal Weights Model, the unstandardized factor loadings for the two groups formed by the performance assessment problem students chose to answer were constrained to be equal. Model I has 59 degrees of freedom.

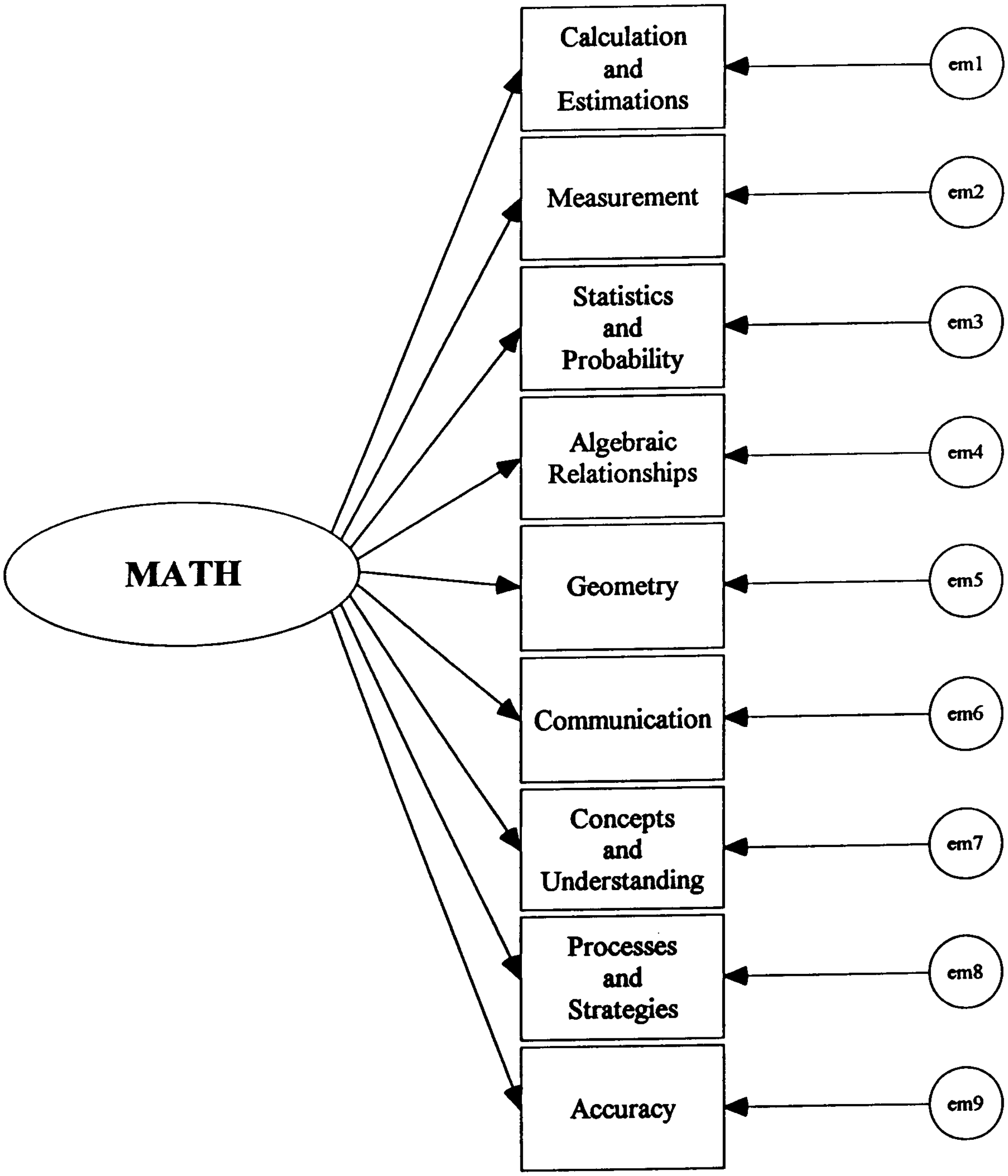

Figure 1. One-factor model for nine math scores.

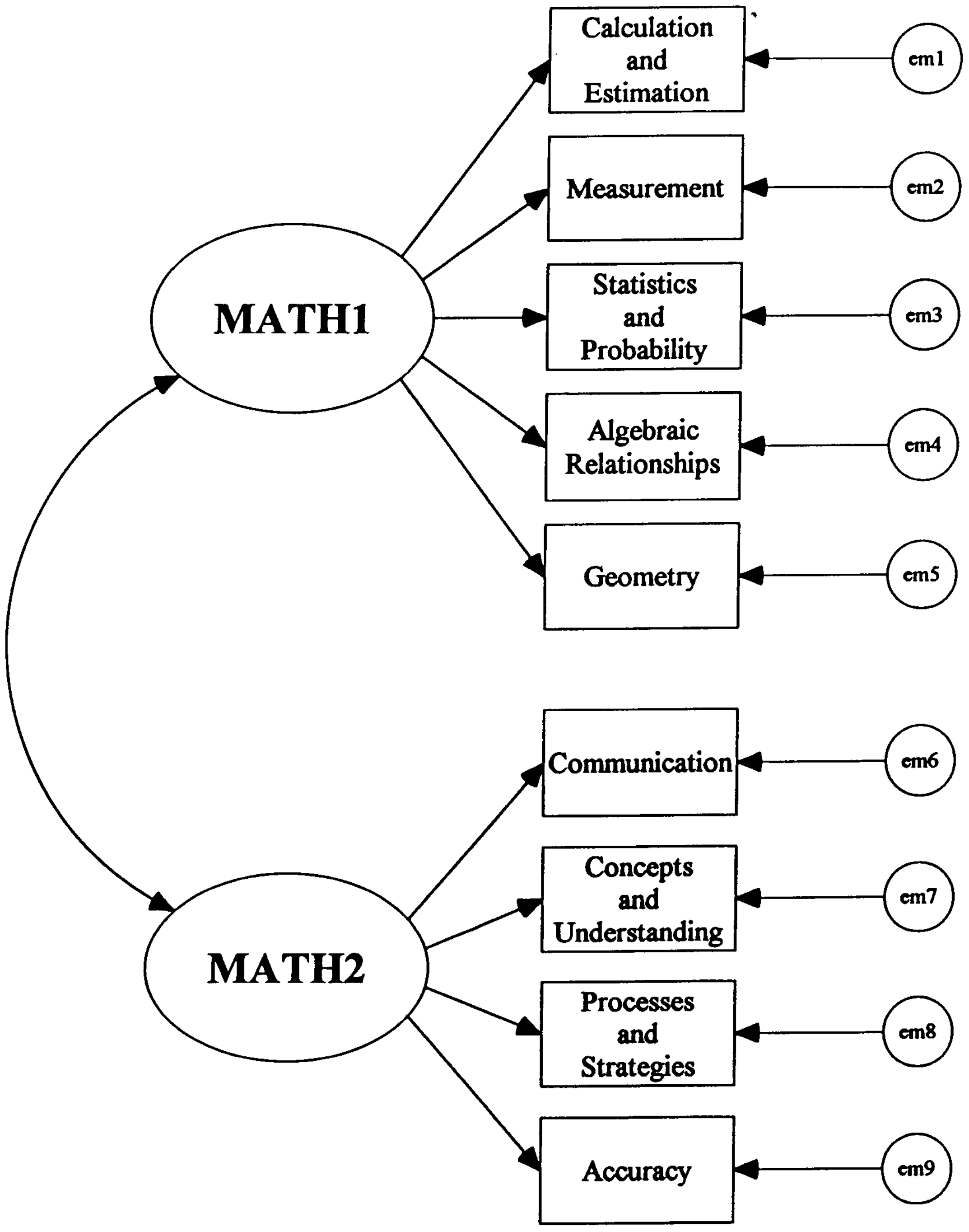

Figure 2. Two-factor model for nine math scores.

Model II, the Weights Free by Group Model, used the same two-factor structure for each problem-choice group, but allowed for different factor weights for all the observed variables as well as different factor variances and covariances. Model II has 52 degrees of freedom and is nested within Model I. Thus, the difference between the chi-square statistics of Models I and II is a chi-square statistic with 7 (59-52) degrees for freedom.

The chi-square statistics, chi-square values divided by their degrees of freedom, the generalized fit statistics (GFI), and the change in chi-square values for the comparisons of Models I and II are listed in Table 16 for each of the six forms. Chi-square and fit statistics (GFI and χ^2 / df are also reported for Models III and IV, which are simply the separate models for the two problem-choice groups that when combined form Model II.

The chi-square statistics are significant ($p < .01$) for all four models on each of the six forms. These significant chi-square values are hardly surprising given the large samples of students for each problem choice. As is shown in the first column of Table 16, the sample sizes range from 1,287 to 3,382 across the 12 form-by-problem-choice groups. Despite the significant chi-square values for all models, the fit is quite good. The GFI indices for Model I, the most constrained model, range from .986 to .990 across the six forms. The GFI for the separate models for each problem choice (Models III and IV) range from .984 to .992. Chi-square values divided by degrees of freedom also suggest generally good fit.

The standardized factor weights and factor intercorrelations for Models III and IV for Form A are reported in Table 17. A comparison of the standardized factor weights across problems 1 and 2 shows that the weights on both factors are quite similar. The largest difference in standardized factor weights for an observed variable is only .04 (obtained for the calculation/estimation score). There is, however, a larger difference in the correlation between the two factors. The two math factors correlated .79 for problem 1 compared to only .66 for problem 2.

Table 16

Chi-Square Statistics and Fit Indices for Two-Factor Structural Equation Models With Different Constraints Across Groups Formed by Performance-Assessment Problem Choice (Replicated by Form)

Form (N)	Model	df	χ^2	χ^2 / *df*	GFI	Change in *df*	Change in χ^2
A	I. Equal Weights	59	251.4	4.26	.990		
	II. Weights Free by Group	52	199.3	3.83	—	7	52.1
2,602	III. Problem 1 Group Only	26	91.1	3.50	.992		
1,523	IV. Problem 2 Group Only	26	108.2	4.16	.984		
B	I. Equal Weights	59	302.5	5.13	.989		
	II. Weights Free by Group	52	264.6	5.09	—	7	37.9
3,220	III. Problem 1 Group Only	26	162.1	6.23	.989		
2,606	IV. Problem 2 Group Only	26	102.5	3.94	.989		
C	I. Equal Weights	59	202.1	3.43	.990		
	II. Weights Free by Group	52	180.0	3.46	—	7	22.1
3,212	III. Problem 1 Group Only	26	115.0	4.42	.992		
1,287	IV. Problem 2 Group Only	26	65.0	2.50	.989		
D	I. Equal Weights	59	244.3	4.14	.987		
	II. Weights Free by Group	52	203.7	3.92	—	7	40.6
2,202	III. Problem 1 Group Only	26	86.0	3.31	.991		
2,084	IV. Problem 2 Group Only	26	117.7	4.53	.991		
E	I. Equal Weights	59	221.0	3.75	.987		
	II. Weights Free by Group	52	197.8	3.80	—	7	23.2
2,234	III. Problem 1 Group Only	26	122.0	4.69	.987		
1,924	IV. Problem 2 Group Only	26	75.8	2.92	.991		
F	I. Equal Weights	59	345.5	5.86	.986	7	
	II. Weights Free by Group	52	297.7	5.73	—		47.8
3,382	III. Problem 1 Group Only	26	212.1	8.16	.985		
2,610	IV. Problem 2 Group Only	26	85.6	3.29	.992		

Tables 18 through 22 report the Model III and Model IV confirmatory factor analysis results for Forms B through F, respectively, in a manner parallel to that used in Table 17 for Form A. From an inspection of Tables 18 through 22, it can be seen that the differences between standardized coefficients are always less than or equal to .05 for 8 of the 9 observed scores. The one variable that is an exception is the Accuracy score, where the weights differ by .19 (.70 vs. .89) on Form C, by .14 (.66 vs. .80) on Form D and by .17 (.69 vs. .89) on Form F. The correlations between the two math factors also show some variability with differences of .13, .01, .10, .07, .16, and .18 on Forms A through F, respectively.

Table 17

Form A Standardized Factor Weights and Factor Intercorrelations for the Separate Two-Factor Models for Each Problem Choice

Variable		Problem 1 (Pizza) factors		Problem 2 (Box) factors	
		Math 1	Math 2	Math 1	Math 2
Calculation/Estimation		.82	.00	.78	.00
Measurement		.79	.00	.76	.00
Statistics/Probability		.75	.00	.72	.00
Algebra		.82	.00	.80	.00
Geometry		.81	.00	.78	.00
Conceptual Understanding		.00	.96	.00	.96
Processes & Strategies		.00	.97	.00	.96
Communication		.00	.84	.00	.87
Accuracy		.00	.84	.00	.82
Factor	Math 1	1.00		1.00	
Inter-	Math 2	.79	1.00	.66	1.00

Table 18

Form B Standardized Factor Weights and Factor Intercorrelations for the Separate Two Factor Models for Each Problem Choice

Variable		Problem 1 (Pizza) factors		Problem 2 (Box) factors	
		Math 1	Math 2	Math 1	Math 2
Calculation/Estimation		.82	.00	.80	.00
Measurement		.78	.00	.77	.00
Statistics/Probability		.74	.00	.72	.00
Algebra		.81	.00	.81	.00
Geometry		.79	.00	.80	.00
Conceptual Understanding		.00	.97	.00	.97
Processes & Strategies		.00	.96	.00	.97
Communication		.00	.82	.00	.86
Accuracy		.00	.82	.00	.83
Factor	Math 1	1.00		1.00	
Inter-	Math 2	.70	1.00	.71	1.00

Table 19

Form C Standardized Factor Weights and Factor Intercorrelations for the Separate Two Factor Models for Each Problem Choice

		Problem 1 (Pizza) factors		Problem 2 (Box) factors	
Variable		Math 1	Math 2	Math 1	Math 2
Calculation/Estimation		.78	.00	.82	.00
Measurement		.76	.00	.79	.00
Statistics/Probability		.73	.00	.74	.00
Algebra		.80	.00	.85	.00
Geometry		.79	.00	.82	.00
Conceptual Understanding		.00	.94	.00	.97
Processes & Strategies		.00	.95	.00	.98
Communication		.00	.83	.00	.86
Accuracy		.00	.70	.00	.89
Factor	Math 1	1.00		1.00	
Inter-	Math 2	.63	1.00	.73	1.00

Table 20

Form D Standardized Factor Weights and Factor Intercorrelations for the Separate Two Factor Models for Each Problem Choice

		Problem 1 (Pizza) factors		Problem 2 (Box) factors	
Variable		Math 1	Math 2	Math 1	Math 2
Calculation/Estimation		.81	.00	.80	.00
Measurement		.77	.00	.78	.00
Statistics/Probability		.73	.00	.72	.00
Algebra		.80	.00	.78	.00
Geometry		.77	.00	.78	.00
Conceptual Understanding		.00	.94	.00	.96
Processes & Strategies		.00	.94	.00	.96
Communication		.00	.82	.00	.82
Accuracy		.00	.66	.00	.80
Factor	Math 1	1.00		1.00	
Inter-	Math 2	.62	1.00	.69	1.00

Table 21

Form E Standardized Factor Weights and Factor Intercorrelations for the Separate Two Factor Models for Each Problem Choice

		Problem 1 (Pizza) factors		Problem 2 (Box) factors	
Variable		Math 1	Math 2	Math 1	Math 2
Calculation/Estimation		.78	.00	.78	.00
Measurement		.75	.00	.76	.00
Statistics/Probability		.72	.00	.72	.00
Algebra		.79	.00	.80	.00
Geometry		.76	.00	.78	.00
Conceptual Understanding		.00	.96	.00	.97
Processes & Strategies		.00	.97	.00	.96
Communication		.00	.84	.00	.81
Accuracy		.00	.84	.00	.80
Factor	Math 1	1.00		1.00	
Inter-	Math 2	.79	1.00	.63	1.00

Table 22

Form F Standardized Factor Weights and Factor Intercorrelations for the Separate Two Factor Models for Each Problem Choice

		Problem 1 (Pizza) factors		Problem 2 (Box) factors	
Variable		Math 1	Math 2	Math 1	Math 2
Calculation/Estimation		.78	.00	.82	.00
Measurement		.73	.00	.80	.00
Statistics/Probability		.71	.00	.75	.00
Algebra		.81	.00	.83	.00
Geometry		.77	.00	.83	.00
Conceptual Understanding		.00	.95	.00	.96
Processes & Strategies		.00	.94	.00	.97
Communication		.00	.80	.00	.85
Accuracy		.00	.69	.00	.86
Factor	Math 1	1.00		1.00	
Inter-	Math 2	.61	1.00	.79	1.00

One final set of structural equation analyses was conducted to take advantage of the design that repeated three of the performance assessment problems by pairing them with different problems on different forms. Recall that the Pizza problem appeared as the first problem on Form A, and the second problem on Forms E and F (see Table 2). That is, problems A1, E2, and F2 were the same Pizza problem. Similarly, the Marathon problem appeared as problems B1, D2 and E1, and the Target problem appeared as problems C1, D1, and F1. This design allowed three additional confirmatory factor analyses for simultaneous groups using the same three-factor model shown in Figure 2. Three analyses, one for each repeated problem, were completed. The model in each run fixed the unstandardized factor weights to be the same for all observed variables across the three forms in which a common problem appeared (e.g., A1, E2, and F2 for the Pizza problem). For a fixed problem, this constrained equal-weights-across-forms model is called Model V. The degrees of freedom, chi-square statistics, and fit statistics for Model V are listed in Table 23. Also shown in Table 23 are degrees of freedom and chi-square values for Model VI, which is the two-factor model where weights are free to vary across form for each of the repeated problems. The last two columns of Table 23 report the change in degrees of freedom and the change in chi-square statistics for the comparison of Models V and VI.

As can be seen, the fit as judged by the GFI or by the ratio of the chi-square values to the degrees of freedom is reasonably good for the constrained Model VI. Although the changes in chi-square statistics are all significant (p < .05)—indicating

Table 23

Chi-Square Statistics and Fit Indices for Two-Factor Structural Equation Models With Different Constraints Across Groups Formed by Performance-Assessment Problem Choice (Replicated Across Forms for Common Problems)

Problem	Model	df	χ^2	χ^2 / *df*	GFI	Change in *df*	Change in χ^2
Pizza	VI. Equal Weights	92	277.2	3.01	.991		
	VII. Weights Free by Form	78	252.5	3.24		14	24.7
Marathon	VI. Equal Weights	92	413.5	4.49	.987		
	VII. Weights Free by Form	78	401.9	5.15		14	11.6
Target	VI. Equal Weights	92	441.7	4.80	.988		
	VII. Weights Free by Form	78	413.1	5.30		14	28.6

that Model VI provides a significantly better fit than Model V—the improvement is modest. Indeed, the ratio of the chi-square statistic for the model to the model degrees of freedom is smaller in all three cases for Model V than Model VI.

Discussion and Conclusions

The analyses reported above support a number of conclusions about the effects of allowing students to choose which of a pair of problems to answer.

1. It is clear that problems differ in popularity. On all six forms there was a tendency for students to be more likely to choose the first problem presented than the second problem. The strength of this tendency varied across form, presumably due to differences in the relative appeal of problems regardless of problem position. It is worth recalling, however, that the directions students were given regarding problem choice were quite brief and did not include advice on strategies students might use in choosing which problem to answer. Thus, it is unclear how much differences in problem popularity might change if students were given advice on strategies to use in choosing a problem to answer.

2. It is clear that groups choosing different problems differ systematically. Problems are not equally attractive to boys and girls or to groups formed on the basis of race/ethnicity or parents' education. Furthermore, the groups of student choosing different problems differ in terms of their mean content-based math performance.

3. The accuracy and dimension scores that students obtain on the performance assessment problems differ as a function of the problem students choose to answer. Those differences in mean scores on the four dimensions remain after adjusting for differences in content-based math scores.

4. The discriminant analyses show some consistency with between problem choice and relative levels of performance in different content areas. However, the variables that best discriminate between groups choosing different problems do not correspond in a straightforward fashion to differences in student strengths and weaknesses on different subdomains of mathematics (e.g., algebra, geometry).

5. Overall, the confirmatory factor analysis models tested indicate two things. First, the models that tested form by problem choice pairwise (i.e., Models I-II) indicate that despite students choosing different problems, sometimes from different content strands, the two-factor structure holds well across the two groups. Or, more

roughly, doing different problems did not alter the factor structure. Second, the models that tested form by problem choice in groups of three (i.e., Model V) indicate that despite the ordering of a particular problem (i.e., whether, for example, the Pizza problem appeared first or second in the list), confounded with the choice of alternate problems, the two-factor structure holds well across the three groups. Or, more roughly, the ordering combined with the problem that a problem is paired with did not alter the factor structure.

Given only the results of the structural equation analyses, an argument may be made for providing choice since similar validity is obtained for measuring the underlying constructs. On the other hand, the results focusing on mean differences in performance make it evident that it would be unwise to ignore problem choice in reporting scores, particularly for potentially high-stakes purposes such as the determination of whether or not a student meets the standard of initial mastery. The percentages of students meeting a combined standard on the performance-assessment dimension vary too much as a function of problem to be ignored. Thus, it seems clear that fairness requires that some equating adjustments would be needed before making high-stakes decisions based on performances of students on problems where choice is allowed.

References

Fitzpatrick, A. R., & Yen, W. M. (1995). The psychometric characteristics of choice items. *Journal of Educational Measurement*, *32*, 243-259.

Gulliksen. H. O. (1950). *A theory of mental tests*. New York: Wiley. (Reprinted 1987, Hillsdale, NJ: Lawrence Erlbaum Associates)

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Oregon Department of Education. (1996, May). *Assessment update* (Vol. 1, No. 1). Salem, OR: Oregon Department of Education, Office of Assessment and Evaluation.

Oregon Department of Education. (1996-97). *Mathematics problem solving: Student directions*. Salem, OR: Oregon Department of Education, Office of Assessment and Evaluation.

Oregon Department of Education. (1997). *Mathematics: Teacher support package.* Salem, OR: Oregon Department of Education, Office of Assessment and Evaluation.

Wainer, H., & Thissen, D. (1994). On examinee choice in educational testing. *Review of Educational Research*, *64*, 159-195.

Wainer, H., Wang, X., & Thissen, D. (1994). How well can we compare scores on test forms that are constructed by examinees' choice? *Journal of Educational Measurement*, *31*, 183-199.

Wang, X., Wainer, H., & Thissen, D. (1995). On the viability of some untestable assumptions in equating exams that allow examinee choice, *Applied Measurement in Education, 8, 211-225.*