College Board Report No. 95-3
ETS RR No. 96-20

Differences in Strategies Used to Solve Stem-Equivalent Constructed-Response and Multiple-Choice SAT®-Mathematics Items

IRVIN R. KATZ, DEBRA E. FRIEDMAN,
RANDY ELLIOT BENNETT, AND ALIZA E. BERGER

College Entrance Examination Board, New York, 1996

Acknowledgments

This research was supported in part by the College Board and Educational Testing Service. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the College Board or ETS. We thank Ann Gallagher, Carol Jackson, Anil Kanjee, Margaret Redman, and Susan Wilson for their assistance with this project, and Brent Bridgeman, Drew Gitomer, Jacqueline Jones, and William Ward for their comments on earlier drafts of this report.

Irvin R. Katz is a Research Scientist and Debra E. Friedman is a Research Associate for the Division of Cognitive and Instructional Science at Educational Testing Service. Randy Elliot Bennett is a Principal Research Scientist for the Division. Aliza E. Berger is with the Department of Education at Ben Gurion University.

Researchers are encouraged to freely express their professional judgment. Therefore, points of view or opinions stated in College Board Reports do not necessarily represent official College Board position or policy.

Additional copies of this report may be obtained from College Board Publications, Box 886, New York, New York 10101-0886. The price is $15.00.

Printed in the United States of America.

Contents

Figures

Tables

Abstract

This study investigated the strategies subjects adopted to solve stem-equivalent SAT-Mathematics (SAT-M) word problems in constructed-response (CR) and multiple-choice (MC) formats. Parallel test forms of CR and MC items were administered to subjects representing a range of mathematical abilities. Format-related differences in difficulty were more prominent at the item level than for the test as a whole. At the item level, analyses of subjects' problem-solving processes appeared to explain difficulty differences as well as similarities.

Differences in difficulty derived more from test-development than from cognitive factors: On items in which large format effects were observed, the MC response options often did not include the erroneous answers initially generated by subjects. Thus, the MC options may have given unintended feedback when a subject's initial answer was not an option or allowed a subject to choose the correct answer based on an estimate.

Similarities between formats occurred because subjects used similar methods to solve both CR and MC items. Surprisingly, when solving CR items, subjects often adopted strategies commonly associated with MC problem solving. For example, subjects appeared adept at estimating plausible answers to CR items and checking those answers against the demands of the item stem.

Although there may be good reasons for using constructed-response items in large-scale testing programs, multiple-choice questions of the sort studied here should provide measurement that is generally comparable to stem-equivalent constructed-response items.

Introduction

Researchers have frequently noted that some items are more difficult in the constructed-response format than in the multiple-choice format, while performance on other items appears to be unaffected by format (Traub, 1993). For example, Ward, Dupree, and Carlson (1987) classified reading comprehension items according to the cognitive demands the researchers assumed the items placed on examinees. Factor analyses did not support the notion that performance differences between formats were related to the cognitive demands of the items. Similarly, researchers examining the results of performance on computer science items were unable to document performance differences even when formats appeared to make very different cognitive demands (Bennett, Rock, & Wang, 1991). What causes such results remains unclear. One reason for failing to explain the presence and unexpected absence of format differences may be that such investigations have focused almost exclusively on the *results* of examinees' performance, neglecting the methods used to solve items—a potentially important source of format-related differences.

A process-oriented approach, as a complement to the traditional result-oriented approach, was taken by Martinez and Katz (1996) in their analysis of the problem-solving requirements of stem-equivalent, architecture figural-response items. The researchers identified three types of items distinguished by the general processes needed to solve the items: (1) items that test for knowledge of the definition of architectural symbols (declarative); (2) items that require examinees to apply a standard procedure, often one learned in the classroom (learned procedure); and (3) items that require examinees to apply their knowledge in a novel way (discovered strategy). Psychometric and process analyses (using "think aloud" protocols) agreed that there were few format differences on items requiring the examinees to apply a learned procedure, whereas the puzzle-like discovered-strategy problems tapped different skills depending on format.

In the current study, we investigated the different strategies subjects adopted to solve items in stem-equivalent constructed-response (CR) and multiple-choice (MC) formats in which the formats differed only in that the MC problems contained response options (Traub & MacRury, 1990). Although there are many other forms of CR items that differ more widely from MC (Bennett, 1993), we focused our analyses on stem-equivalent items in order to reduce the potential sources of differences in performance, thus making the task of identifying format-related differences more tractable.

How could an understanding of problem-solving processes shed light on format effects? We offer the following conjecture (also discussed by Martinez & Katz, 1996, and Traub, 1993): to the extent that the processes involved in solving the CR and MC versions of an item are the same, there should be no format effects (e.g., in terms of difficulty). Note that the converse—different problem-solving processes leading to different levels of difficulty—might not always occur even if our conjecture is correct. It is possible for different problem-solving processes to result coincidentally in similar levels of difficulty.

Because the only difference between stem-equivalent CR and MC items is that the latter contain response options, it has been suggested that whether an examinee uses the response options in solving an item determines whether format effects will occur (Traub, 1993). This claim implies a process-based explanation of format ef-

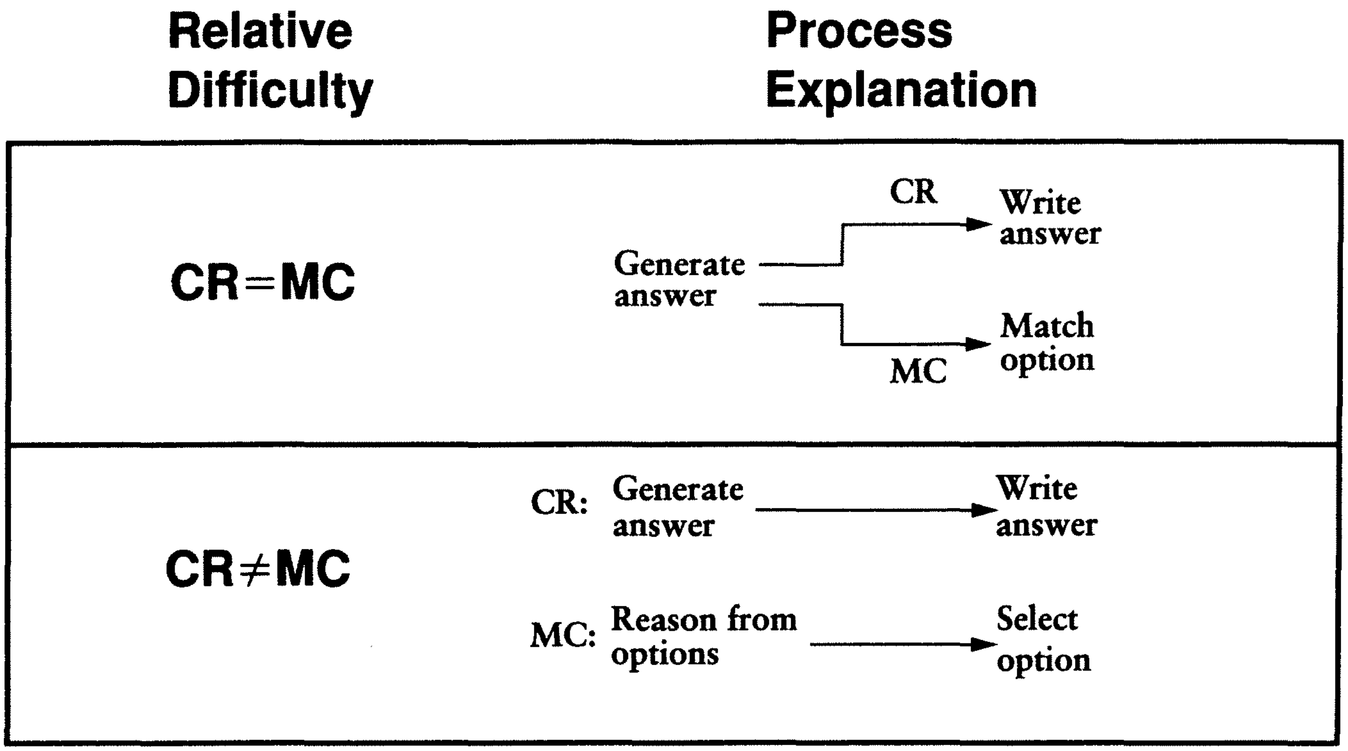


FIGURE 1. Possible explanation for format differences.

fects. As an example, consider Figure 1. When performance on the MC and CR counterparts of an item is equivalent, an examinee generates an answer in the "traditional" manner, perhaps by writing and solving equations, regardless of format. In the case of a CR item, the answer produced is simply written down; for MC items, the examinee matches the generated answer with the correct response option (Snow, 1980). When performance on MC and CR counterparts is not equivalent, examinees solve MC items by using the response options as aids in selecting the correct answer.

To investigate the relationship between item format and problem-solving processes, we created parallel test forms of CR and MC items. The items presented in MC format on one form were presented as CR items on the other, and vice-versa. These forms were evenly distributed among students representing a range of mathematical ability. In keeping with past studies, our analyses first focused on format differences in terms of accuracy on the test as a whole. We then investigated the processes underlying any format differences (or lack thereof) via a comparative analysis of problem-solving strategies between formats.

Method

Subjects

A list of 672 high school students taking the June 1991 administration of the SAT and living in the greater Princeton area was obtained from ETS program files. Letters were mailed to these students describing the project and inviting each to either mail back a stamped, self-addressed postcard or telephone us directly. Of the 672 contacted, 208 students responded. Each student was telephoned and invited to come to ETS. Fifty-five of the 208 students took part in the study.

The subjects were grouped into three ability levels: low ability (n = 17), defined as having recent SAT-M scores of 375–450 (26th–42nd percentile); medium ability (n = 18), defined as scores of 475–550 (50th–71st percentile); and high ability (n = 20), defined as scores of 600–800 (82nd–99th percentile). The three ability levels represented the top three quartiles, narrowed somewhat to accentuate differences between groups. The subjects were selected so that at each ability level there were approximately equal numbers of males and females. Table 1 shows the mean SAT-M score for each ability group.

Table 1

Mean SAT-M Scores by Ability Level

	Ability level			
	Low	*Medium*	*High*	*Overall*
Male	413	520	653	533
	(8)	(9)	(9)	(26)
Female	418	513	655	538
	(9)	(9)	(11)	(29)
Overall	415	517	655	535
	(17)	(18)	(20)	(55)

Note: Values enclosed in parentheses represent *n* per cell.

Instruments

Ten multiple-choice (MC) items were selected from disclosed forms of the SAT-M. These items represented the general content categories of the test: three algebra items, five arithmetic items (including three "percent" questions and two other arithmetic items), and two geometry items. The items were selected subject to the following three constraints:

1. All the items could be converted into the constructed-response format by deleting their response options;
2. Isomorphic items could be created for all questions in order to develop a new item that would require the same problem-solving strategy as the original, but have a different enough cover story that subjects would not notice the similarity;
3. A range of difficulty could be represented. Six of the items chosen were of medium difficulty (equated Δ[1] $= 11–13$) and four were easier (equated $\Delta < 10$). The easier items were included to maintain the motivation of the lower-ability subjects.

Tickets version (original):
If 70 tickets to a play were bought for a total of $50.00 and if tickets cost $1.00 for adults and $0.50 for children, how many children's tickets were bought?

Tickets version (isomorph):
Jenna won a total of 90 red tokens and yellow tokens while playing a board game. Each red token is worth 1 point and each yellow token is worth 4 points. If the total value of Jenna's red and yellow tokens is 120 points, how many yellow tokens does she have?

FIGURE 2. Example of isomorphic items.

Figure 2 shows an SAT-M item and its isomorph. Note that both items may be classified as "simultaneous equations" problems. Also, the items involve the same number of quantities and these quantities are in the same qualitative relation to one another. The only differences are in the story used to describe the quantities and the provided values.

We used item isomorphs to alleviate one of the more difficult problems encountered in studying differences between item formats, that of the contamination induced by asking subjects to solve the same item in two formats. Using isomorphs is a reasonable approach because there is considerable evidence suggesting that individuals fail to recognize equivalent problems, even if the problems differ only in the details of their cover stories (Gick & Holyoak, 1983).

The selection procedure resulted in a set of 40 items:[2] 10 original MC items, 10 isomorphic MC items, 10 CR versions of the original MC items, and 10 CR versions of the MC isomorphs (see Figures A-1 to A-10 in the Appendix). These items were compiled into two tests of 20 items each, with each test consisting of 10 MC items and the CR counterparts of their isomorphs. For each test, two counter-balanced orders of format presentation were created, resulting in four test forms (Figure 3). Approximately equal numbers of subjects from each ability group were assigned to take each test form.

	Problem Set 1 (k=20)	Problem Set 2 (k=20)
Administration order A	Form 1A MC–original items CR–isomorph items	Form 2A MC–isomorph items CR–original items
Administration order B	Form 1B CR–isomorph items MC–original items	Form 2B CR–original items MC–isomorph items

FIGURE 3. Contents of the four test forms.

Procedures

The test forms were administered individually in sessions lasting 1.5–2 hours. The subjects worked alone in a room separate from the experimenter, although the experimenter was available to clarify the task, if necessary.

When they arrived, the subjects were informed that they would be taking a test similar to the SAT-M, consisting of multiple-choice and constructed-response questions. Subjects were asked to work as quickly but as accurately as possible and to complete the problems one at a time, in order, and without going back. Approximately half of the subjects were told there would be a four-minute time limit on each problem; the other subjects were allowed to take as long as needed. This manipulation was introduced in the anticipation that some degree of speededness would accentuate format differences.

The subjects provided concurrent verbal protocols (cf., Ericsson & Simon, 1984) as they solved the items. Subjects were instructed to say aloud anything that they would normally "say" to themselves while solving a problem. Subjects' verbalizations were recorded on videotape. The videotape also recorded any notes or calculations made by the subjects.

Results

Test-Level Analyses

The first question to be addressed was, simply, does format affect accuracy for isomorphic MC and CR items? We ran a format (CR, MC) by ability (high, medium, low) by format-order (MC first, CR first) by timing (whether a time limit was given) repeated-measures ANOVA, with item format (CR, MC) as a within-subjects factor and with ability, format-order, and timing as between-subjects factors. The dependent measures for the ANOVA were the total number correct on each section (CR and MC) of the test. Note that because the number of subjects was small relative to the number of factors, the power of the statistical test was considerably limited.

Table 2

Analysis of Variance Results for Test-Level Effects

Source	df	F
	Between subjects	
Timing (T)	1	2.78
Ability (A)	2	24.86***
Order (O)	1	2.17
T × A	2	.93
T × O	1	.15
A × O	2	.74
T × A × O	2	.39
S within-group error	43	(2.72)
	Within subjects	
Format (F)	1	6.53*
F × T	1	.31
F × A	2	.50
F × O	1	7.40**
F × T × A	2	.22
F × T × O	1	.36
F × A × O	2	.32
F × T × A × O	2	.75
F × S within-group error	43	(1.67)

Note: Values enclosed in parentheses represent mean square errors.
S = subjects.

$^*p < .05$. $^{**}p < .01$. $^{***}p < .0001$.

Results from the ANOVA are presented in Table 2; the corresponding means and standard deviations are shown in Table 3.[3] Significant effects were found for format, ability, and the format-by-format-order interaction. The main effect of format was the smallest of the three significant sources of variation ($F(1,43) = 6.53$, $p<.02$) and might be partially explained by subjects correctly guessing on some MC items. The main effect of ability ($F(2,43) = 24.86$, $p<.0001$) was more substantial, but expected, with the lower-ability subjects performing worst, the high-ability subjects best, and the medium-ability subjects in between. The significant interaction between format and format-order ($F(1,43) = 7.40$, $p<.01$) stemmed primarily from the performance on the CR items of subjects who answered the CR items first. The subjects' performance on these items was worse than their performance on the MC items and worse than the performance of the MC-first subjects on items in either format. One explanation for this effect is that the subjects learned something while solving the MC items that helped them solve the CR counterparts. Finally, there was no significant ability-by-format or timing-by-format interaction.

Table 3

Means and Standard Deviations for Test-Level ANOVA

Format-Order:	MC-first		CR-first		Overall	
	MC	CR	MC	CR	MC	CR
Ability level						
Low						
M	6.3	6.2	6.0	4.6	6.2	5.5
SD	1.7	1.9	1.2	1.3	1.4	1.8
Medium						
M	7.4	7.6	8.2	6.4	7.9	6.8
SD	1.0	1.3	1.7	1.5	1.6	1.5
High						
M	8.4	8.6	8.9	8.1	8.7	8.4
SD	1.2	1.4	1.4	1.6	1.3	1.5
Overall						
M	7.4	7.5	7.8	6.5	7.6	7.0
SD	1.6	1.8	1.9	2.0	1.7	2.0

Did all items contribute equally to the test-level effects? Previous research suggests that individual items may be more or less sensitive to response format (Bridgeman, 1992; Martinez & Katz, 1996). Figure 4 shows the difference between proportion correct on the MC versus CR versions of each item type. Each set of bars represents a different original-isomorph item pair; the items are ordered from greatest to least in terms of

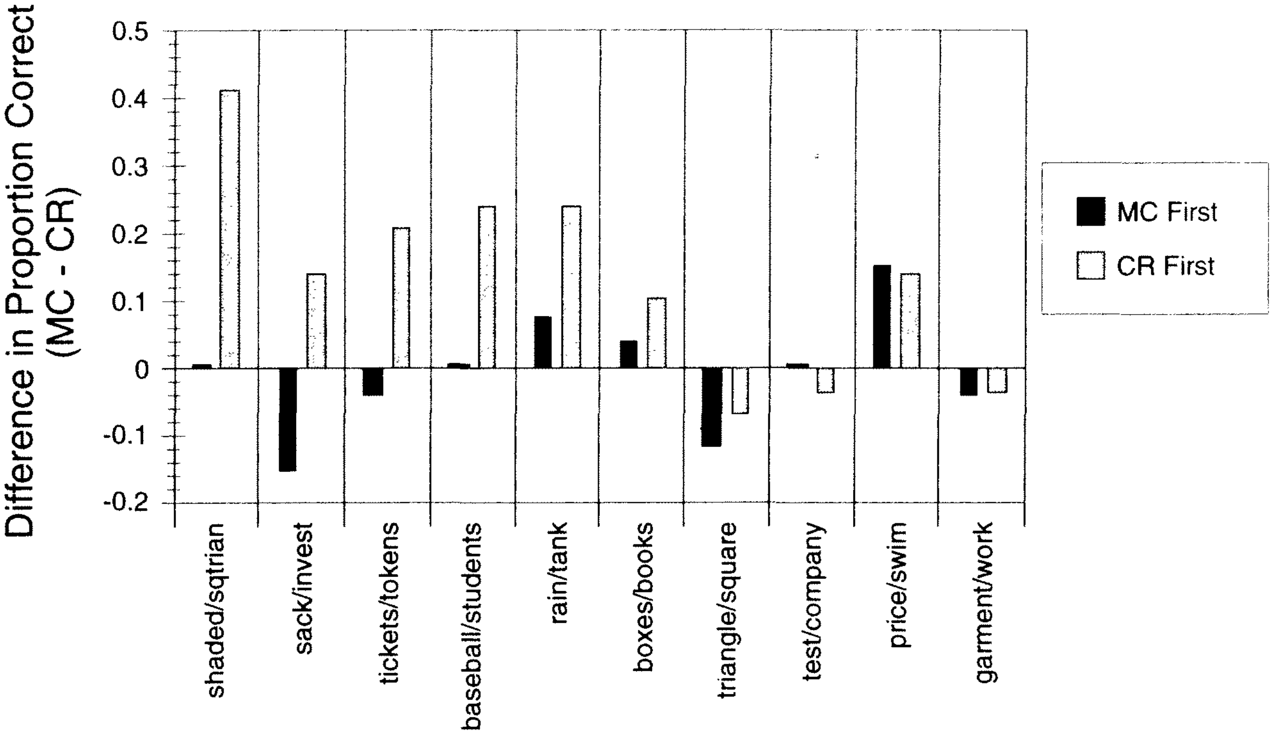


FIGURE 4. Difference in proportion correct between formats (MC–CR) separated by item pair and format-order.

the size of an item pair's format-by-format-order interaction. Bars that extend above the zero point represent items for which the MC version was easier, while bars below the zero point indicate that the CR version of an item was easier. Solid bars represent scores for subjects answering the MC items first; shaded bars represent scores for the CR-first subjects.

Even though the CR and MC items used in this study were very similar, there was a wide range of differences in difficulty. Items ranged from being much easier in the MC format, to having approximately equal difficulty regardless of format, to being slightly easier in the CR format. In addition, the degree of the format-by-format-order interaction varied across item types: for the first five item types shown in Figure 4, the CR versions were much more difficult when that format was administered first; for the last five item types, the relative difficulty of the items presented in the two formats seems independent of format-order. These results suggest that a test-level analysis may be missing important item-level differences.

What is the source of these differences in difficulty? We predicted that format-related differences in difficulty would appear when subjects use different problem-solving processes to solve CR and MC versions of the same item. In particular, whether subjects used the response options when solving MC items should determine whether format differences occur.

Item-Level Analyses Based on Problem-Solving Strategies

To address the issue of how the subjects used the MC options, it was necessary first to identify the different strategies (both correct and faulty) subjects used to solve items, which yielded a set of strategy categories unique to each pair of items. We then combined the subjects' problem-solving approaches into two groups: "traditional" strategies, which are commonly associated with CR problem solving (e.g., writing and solving algebraic equations), and "nontraditional" strategies that involve estimation or reasoning from potentially correct answers—strategies commonly associated with MC problem solving. A third category, "unknown/other," indicated that a subject's problem-solving approach could not be identified or that a subject's incorrect approach was not a variant of one of the traditional or nontraditional strategies.

The strategy categories were initially identified by viewing the videotaped protocols of 12 randomly selected subjects. In analyzing the remaining subjects, because the problem-solving strategies were quite distinct, there was little difficulty in unambiguously assigning to one of the categories a particular subject's approach. One researcher classified all 55 subjects' responses, while another researcher classified 20 percent of the responses. Inter-rater agreement for this subsample was 93 percent, and conflicts were resolved through discussion. Occasionally a problem-solving approach not represented in the performance of the initial 12 subjects was encountered, and the categorization scheme was appropriately augmented.

For this analysis, the item pairs were divided into those that showed relatively larger format-related differences in difficulty (hereafter, "format differences") and those that showed smaller differences (median split on maximum MC–CR difference). The CR and MC versions of the shaded/sqtrian, baseball/students, and rain/tank item pairs (Figures A-1, A-3, and A-4, respectively, in the Appendix) differed in difficulty more than did other items. The sack/invest, tickets/tokens, and price/swim item pairs (Figures A-2, A-5, and A-8, respectively, in the Appendix) exhibited smaller format differences. The latter three item pairs were chosen because they could be solved using either traditional or nontraditional strategies and because their lack of format differences could not be attributed to ceiling effects (the percentage correct for each item pair was less than .87). The remaining four item pairs did not meet these criteria and so were not included in the analysis.

For the three item pairs that showed larger differences in difficulty, the expectation was that those differences resulted because subjects used the response options (i.e., used nontraditional strategies) to solve the MC versions of the items, but used traditional strategies to solve the CR items. Table 4 shows the distribution of strategies by format for the items exhibiting format differences. As expected, nontraditional strategies were used more often by subjects when solving the MC items. Thirty percent of solutions to the MC items demonstrated nontraditional strategies compared with 19 percent of solutions to the CR items. Furthermore, when solving the CR items, subjects were less successful (19 percent correct) in their use of nontraditional strategies compared to when they solved the MC items (53 percent).

This latter effect stemmed primarily from the shaded/sqtrian item (Figure A-1 in the Appendix), which many subjects solved through visual estimation based on the figure provided. We can speculate that estimation methods were more likely to work in the MC version simply because subjects could use an "estimate-and-choose-closest-option" strategy, which was unavailable for the CR items. Of course, incorrect estimations could have occurred even with MC options if those options were numbers that were close to each other (i.e., the distinctions required to select the correct option exceeded the subject's estimation ability).

Table 4

Distribution of Strategies Used for Items Showing Larger Format-Related Differences in Difficulty

	Strategy categories					
	Traditional		*Nontraditional*		*Unknown*	
	Correct/Incorrect		*Correct/Incorrect*		*Correct/Incorrect*	
MC	81	24	26	23	2	9
CR	75	43	6	25	0	16

For the items functioning similarly across formats, the expectation was that subjects would *not* use strategies normally associated with reliance on MC options (i.e., nontraditional strategies), but instead would primarily rely on traditional strategies to the same degree for both formats. Table 5 shows the results. Whereas traditional strategies predominate and were used about equally across formats (MC: 61 percent; CR: 68 percent), nontraditional strategies were also used with approximately equal frequency on the MC and CR items (MC: 35 percent; CR: 30 percent). This finding is consistent with the results based on item difficulty, but unexpected because such popular nontraditional strategies as "plug-in" and estimation are typically associated only with MC items. (In the plug-in strategy, the subject generates potential answers by selecting response options and checking those answers against the item stem.) Similar distributions of traditional and nontraditional strategies thus explain the similar functioning of these items across response formats.

Table 5

Distribution of Strategies Used for Items Showing Smaller Format-Related Differences in Difficulty

	Strategy categories					
	Traditional		*Nontraditional*		*Unknown*	
	Correct/Incorrect		*Correct/Incorrect*		*Correct/Incorrect*	
MC	76	25	44	14	2	4
CR	72	41	35	14	2	1

Recall that all three of the item pairs with large format differences contributed to the format-by-format-order interaction in overall performance (Figure 4). In particular, the MC version of these items tended to be easier than the corresponding CR version when the CR version was presented first. The reason for the interaction can be seen by splitting Table 4 into the two format-order conditions (Table 6). The percentage correct for the MC items was similar irrespective of whether this format was presented first (62 percent) or second (70 percent). The interaction was focused on the CR items. The percentage of correct responses was low when these items were presented first (40 percent), but increased when the CR items were preceded by their MC counter-

Table 6

Format-by-Format-Order Interaction for Items Showing Larger Format-Related Differences in Difficulty

		Strategy categories					
		Traditional		*Nontraditional*		*Unknown*	
Format order:	Format	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
MC-first	MC	35	11	12	14	1	5
	CR	41	18	5	7	0	7
CR-first	MC	46	13	14	9	1	4
	CR	34	25	1	18	0	9

parts (63 percent).

At least a portion of this interaction may be attributed to feedback the MC options provided to subjects—feedback that may, in turn, have aided subjects in solving the CR counterpart items. That is, if a subject generated an answer that was not among the MC alternatives, the subject may have been cued to reexamine his or her problem-solving method or to try an alternative method. This feedback may have prodded the subject to correct his or her faulty problem-solving method and later to apply the correct procedure to counterpart items. Although a few of the most common errors were represented in the MC options, subjects often made arithmetic and other errors (e.g., estimation) that resulted in idiosyncratic answers not included among the MC options.

Unfortunately, even with videotaped protocols, it was difficult to determine whether subjects were considering the MC options. Thus, if a subject generated an answer and then continued problem solving until eventually selecting a result from among the options, it was difficult to tell whether the subject continued problem solving solely because the answer originally generated was not among the MC options. However, we can estimate the influence of feedback from the MC options by observing how many solutions to CR items were not among the alternatives in the MC versions of the items.

For the CR-first subjects, 26 errors were made while solving MC items and 52 while solving CR items. Nine of the 52 incorrect responses represented a failure to provide a response to the item. Of the remaining 43 errors, approximately half (22) were not among the alternatives offered in the MC version.[4] If these 22 subjects had been given the MC version of the items, at least some of them might have ended up responding correctly, reducing the format-by-format-order interaction.

One of the item pairs (sack/invest) showing small format differences nevertheless contributed substantially to the overall format-by-format-order interaction. For the CR-first subjects, approximately the same number of errors were made on the MC (11) and CR (14) versions. Consistent with the results on other items, approximately half (6) of the erroneous responses to the CR version were not included among the MC version's options. Thus, had these subjects solved the MC version of the item, some of them might have answered it correctly.

Summary of Item-Level Effects

The problem-solving strategy analyses presented above resulted in two main findings. First, consistent with expectations, for the items showing format-related differences in difficulty, unequal use of nontraditional strategies between the two formats was observed. However, at least a portion of the format differences could be attributed to inadvertent feedback from the MC options. That is, the largest format differences occurred when (a) the MC options allowed use of an "estimate-and-choose-closest" (or "calculate-and-choose-closest") strategy or (b) the MC options did not contain the erroneous answers that subjects generated. Thus, for MC items, subjects were cued that their initial answer was incorrect—feedback that they did not receive from the CR versions.

The second result was that nontraditional methods (e.g., estimation, plug-in) were used with equal frequency in solving MC and CR items when those items showed small differences in difficulty. This result augments the process explanation of format similarities implied by the literature (Figure 1). When there were no format differences in accuracy, it was not necessarily because subjects used traditional CR methods to solve MC items. Instead, similar rates of accuracy for CR and MC items indicated that subjects used the same processes (whether traditional or nontraditional) when solving items in both formats. Note that use of nontraditional strategies may indicate that an item is tapping constructs different than those tapped when subjects used a traditional approach.

Plug-in Strategy

That a strategy commonly associated with the MC format was used with equal frequency to solve CR items was unexpected. How, exactly, does this strategy work and how is it possible to plug in potential answers with the CR format? In this section, we describe a process model of plug-in behavior that may help to answer these questions.

Our process model was created through qualitative analysis of the videotaped recordings of subjects using the plug-in strategy on the tickets/tokens item pair (Figure A-5 in the Appendix). The majority of subjects solved this item by plugging in potential values for one of the unknown quantities and then seeing whether, given the relations among quantities, a particular value satisfied the constraints of the item stem. Before giving a detailed account of the model, we present an example of a subject's plug-in behavior while solving the CR form of the tickets item:

If 70 tickets to a play were bought for a total of $50.00 and if tickets cost $1.00 for adults and $0.50 for children, how many children's tickets were bought?

This example is from a subject chosen at random from among those using the plug-in strategy. After reading the item stem in its entirety, the subject said, "Well, what if there are 70 adults?" He then reasoned that the cost of the adult tickets would be $70, which is greater than the allotted amount ($50). The subject next decided to set the number of children's tickets at 25. He calculated that the number of adult tickets would be 45 (70 − 25 = 45) and that the cost of the adult tickets would be $45. While the subject calculated the cost of the children's tickets (25 × $.50), he realized that the amount would not be a whole dollar amount, so tried a different value for the number of children's tickets. After trying 20 children's tickets, propagating the value through the other quantities, and rejecting that answer (20 children's tickets would result in 50 adult tickets, which would be $50, so the money would be "all used up"), the subject tried 30 children's tickets. He next calculated the number of adult tickets (70 − 30 = 40), and their cost ($40). He then made a simple math error, incorrectly calculating the cost of the children's tickets as $10. Because $10 + $40 = $50, the desired cost, the subject stated that 30 children's tickets is the correct answer. This cycle of behavior (select value, propagate to other values, evaluate) was observed in practically all subjects using the plug-in strategy.

An outline of the process model is given in Figure 5.

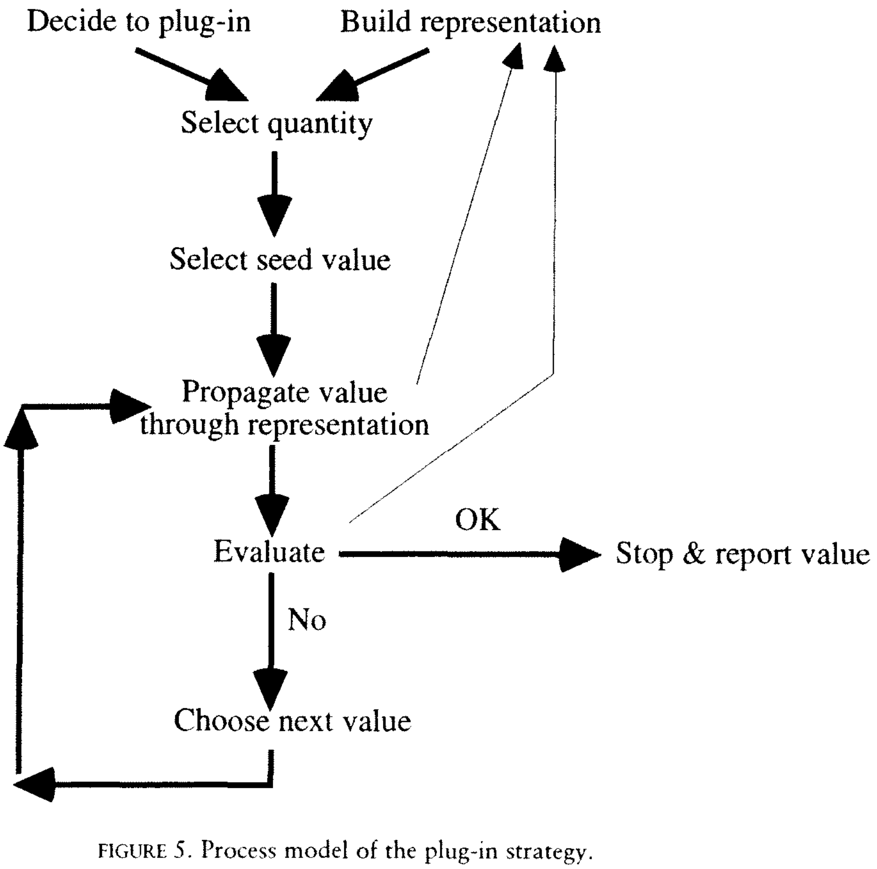
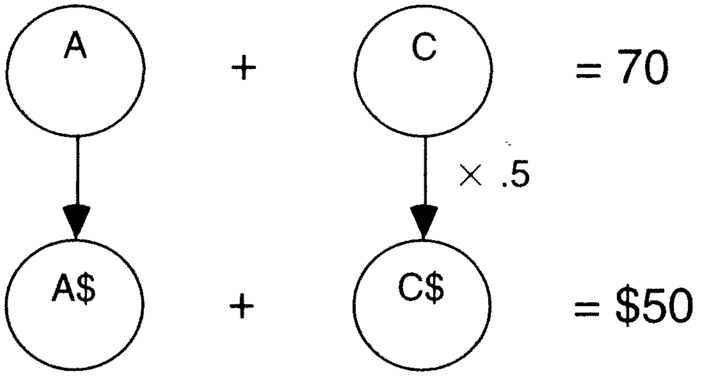


FIGURE 5. Process model of the plug-in strategy.



"A" and "C" represent the number of adult and children's tickets, respectively.
"A$" and "C$" represent the total cost of adult and children's tickets, respectively.

FIGURE 6. Possible mental representation of the tickets item.

Arrows depict the order of the various processes. The *decide to plug in* and *build representation* processes are left unordered, recognizing that some subjects may decide to plug in before developing a detailed mental representation of the item, while others may build a mental representation while attempting to solve the item using another strategy (e.g., algebra). The lighter arrows between *build representation*, *propagate*, and *evaluate* depict the observation that during these latter two processes, subjects might build, add to, or correct their representation of the item.

Each of the processes is described below:

Decide to use plug-in strategy. As suggested earlier, not all subjects decided to use the plug-in strategy immediately; some used the strategy in reaction to their failure with other approaches. Some of those who used it immediately might have done so because they realized that the standard approach (e.g., algebra) would be too difficult or too time-consuming.

Build representation of item. When using the plug-in strategy, a subject's mental representation of the item might contain three types of information: (1) the quantities involved in the item (e.g., number of adult tickets, number of children's tickets, cost of each adult ticket, total cost of all tickets); (2) the mathematical relations among the different quantities; and (3) the constraints on the values that the quantities might take (e.g., cost of adult tickets must be less than $50). Figure 6 depicts the information that might be represented in a subject's mind. Of course, subjects might create an incorrect representation of the item, perhaps leaving out one or more quantities. Note that the same representation may be used for solving the problem algebraically. This correspondence supports our observations that subjects sometimes began solving a problem algebraically, then switched to the plug-in strategy without difficulty.

Select quantity. With the correct representation of an item, a subject can plug values into any of the unknown quantities and propagate those values to the other quantities. However, the solution to an item is easiest if a subject plugs a value into the goal variable (here, the number of children's tickets).

Select seed value. Subjects differed on the first value they selected to plug into their representation of the item. On the MC items, some subjects began with the first option. On the CR items, some subjects chose values seemingly at random, while others worked within a range they had defined.

Propagate value. This process was often used in conjunction with evaluation, as when a value was propagated and the subject recognized that the calculated result was outside the expected range or did not conform to the desired form of the value (e.g., the result was not a whole dollar amount). Generally speaking, most subjects could propagate values successfully, although their representation of the item might have been faulty.

Evaluate; choose next value. After propagating the values, subjects judged whether the original value correctly met the constraints of the item. If the value was deemed correct, the subject responded with that value as the answer. (Although two subjects did continue to plug in even after deciding that a particular value was correct.) For values deemed incorrect, we observed two types of judgments. Some subjects stated merely that the answer was incorrect. Other subjects judged whether the value was too high or too low and then used that judgment as a basis for selecting the next value to be plugged in.

This process model suggests that the plug-in strategy can work similarly with both the MC and CR formats. The processes involved are essentially the same, the primary difference lying in the presence of a small, fixed set of seed values in the MC format, one of which is known to be the correct answer. Because subjects appear to be proficient at estimating appropriate seed values anyway, the absence of response options seems to pose little impediment to the effective use of this strategy with CR items.

Discussion and Conclusions

This study investigated the strategies subjects adopt to solve stem-equivalent SAT-M word problems in MC and CR formats. Consistent with previous analyses, format-related differences in difficulty were more prominent at the item level than for the test as a whole (Bridgeman, 1992). At the item level, subjects' problem-solving approaches appeared to explain format differences as well as similarities.

Differences in difficulty derived more from test development than from cognitive factors: For items in which large format effects were observed, the MC options often did not include the erroneous answers generated by subjects, a result consistent with previous research on the quantitative section of the GRE General Test (Bridgeman, 1992) and on the SAT-M (Braswell, 1990). Thus, the MC options gave unintended feedback when a subject's initial answer was not an option or when they allowed a subject to choose the correct answer based on an estimate. These uses of response options appeared to account for the major performance differences observed between CR and MC items.

Similarities between formats occurred because subjects solved some CR and MC items using similar methods. A typical MC approach is to plug in the response options, looking for one that satisfies the constraints of the item stem. Surprisingly, subjects used this strategy with CR items as frequently as with MC items. Subjects appeared adept at estimating plausible answers to CR items and checking those answers against the demands of the item stem. In other words, subjects frequently generated their own values to plug in.

What are the implications of these findings for the SAT? First, that MC items provide unintended hints suggests a potential source of construct-irrelevant variance (Messick, 1989). How serious a source of irrelevant variance these hints constitute depends on what help they provide and to whom they provide it. High-ability examinees may occasionally pose a wrong answer because of a calculation error, but recover when they discern that the response is not among the options. Because the SAT-M is intended to be a measure of reasoning ability, hints that help to reduce variance due to low-level procedural errors ought to increase construct validity. On the other hand, the response options may occasionally alert lower-ability examinees that their conceptual approach is wrong and that another should be tried. To the extent that subsequent attempts hit on a correct approach, construct validity is diminished. This source of irrelevant variance might be reduced by keying distractors to the conceptual errors that examinees make. Information on such errors might be obtained by initially pretesting items in the CR format, choosing appropriate distractors based on the response data, and then pretesting the MC versions. Whereas CR formats such as the SAT-M grid-in[5] eliminate the potential for hints, they also increase the importance of calculation skills (unless procedural requirements are kept to a minimum or examinees have access to calculators).

In deriving these implications, we assume that the construct intended to be assessed by the SAT-M includes examinees' ability to solve problems without externally provided feedback (such as when the answer generated is not among the response options). However, some educators have suggested that using feedback successfully to correct one's conceptual approach is part of mathematical problem solving. This suggestion, while completely valid, poses challenges for large-scale testing as it requires methods to distinguish between situations in which feedback causes examinees to correct their conceptual approach versus situations in which feedback causes examinees to fall back on an alternative, construct-irrelevant approach (e.g., guessing). Indeed, Gallagher (1992) found gender differences in high-ability students' problem solving subsequent to discovering that their response was not among the MC options. Females were more likely to correct their approach, perhaps discovering where their low-level procedural error had been made; males were more likely to switch to alternative problem-solving strategies (e.g., guessing).

When the MC options permit an examinee to choose the correct answer based on an estimate from a figure or from calculations, the potential effects are more complex. If the estimate is derived from the mathematical principle underlying the item, then construct validity may again be improved. This potential improvement results from avoiding the low-level procedures needed to compute the response exactly, the execution of which only serves to trip up some examinees familiar with the fundamental idea. (The quantitative comparison item type, in fact, rests on this notion of estimation from underlying principles.) However, if the estimate can be generated from factors unrelated to the intended mathematical concept (e.g., a rudimentary figural relation), construct validity will be negatively affected. Thus, the MC options introduce construct-irrelevant variance only when estimates emanate from problem-solving methods tangential to the reasoning ability measured by the SAT-M. In these instances, narrowing the distances between adjacent distractors will discourage estimation, as will changing to the grid-in format.

Does the plug-in strategy constitute a source of irrelevant variance? In the context of items intended to measure algebraic symbol manipulation skills, it clearly does. The student who works $x^2 + 2x - 14 = 10$ by plugging in has avoided solving the problem algebraically. However, when the task is intended to assess mathematical reasoning, as in the word problems used in this study, the situation may differ. Many of the item pairs can be solved in several ways, including by using algebra and the plug-in approach. Regardless of strategy, the student must represent the problem mentally to solve it successfully. This representation phase is arguably the central mathematical reasoning step in the problem-solving process, and how that representation is expressed and executed may be incidental to measuring the intended ability (see, e.g., Bennett & Sebrechts, in press). This is especially the case if mathematical reasoning is recognized as expressible through both formal and informal means, a position consistent with current trends in public education (Working Groups for the Commission on Standards for School Mathematics of the National Council of Teachers of Mathematics, 1989). If, on the other hand, the intention is to measure the ability to represent and solve problems *formally*, plug-in approaches constitute a threat to validity. As we have shown, the CR format does not preclude such alternatives. Consequently, there may be no simple means of eliminating plug-in strategies short of requiring examinees to show their work and grading their problem-solving approaches. Automatic methods for scoring the steps involved in mathematical problem solving are under development (e.g., Sebrechts, Bennett, & Katz, 1993; Sebrechts, Bennett, & Rock, 1991), and may eventually permit the evaluation of problem-solving processes.

Several limitations of this study should be noted. First, these findings are most properly applied to mathematical word problems of the type found on the SAT. Second, the sample was a small, geographically restricted one that was probably not representative of the population of SAT examinees. Third, standard psychometric concerns, such as the effects of guessing on relative difficulty and measurement reliability between the two formats (i.e., guessing makes MC items easier, but adds noise to measurement), were not directly addressed. These facts suggest that the results be replicated with other tests of mathematical ability and with larger, more geographically diverse samples before being generalized. All the same, the consistency of our findings with those of Bridgeman (1992) on the GRE-Q, as well as with those of Braswell (1990) on the SAT-M, suggests that the current findings have some degree of generalizability.

Future research might concentrate on several questions. One question concerns how frequently MC options cue examinees to correct procedural versus reasoning errors in their initial problem-solving efforts. Procedural errors may be the more frequently corrected because they are easier to locate and eradicate. On the other hand, conceptual errors may be more common on the SAT-M given the nature of the examination. Empirical data on this question should have direct relevance for construct validity: If the response options tend to cue the correction of procedural mistakes, the construct validity of MC items as a measure of reasoning should be strengthened.

A second question relates to how radically item formats must differ before process differences occur regularly. The current study was conducted with stem-equivalent items—that is, items identical in every respect except for the presence or absence of response options. The results suggest that format-related process differences can probably be eliminated through such operational measures as narrowing the distances between MC distractors and writing distractors that are more closely keyed to common examinee errors. However, it is possible to write CR items that differ more radically from MC than the stem-equivalent versions used in this study. These differences in item format may occur along several dimensions including the complexity of the response (e.g., in terms of the number of elements), the degree of judgment required in scoring, the number of correct forms that the response might take, and whether the problem even has a correct answer. Which of these dimensions causes examinees to use problem-solving processes that are different from those used for MC items—and, thus, opens the possibility for measuring different abilities—is not clear. Research might proceed by manipulating such item dimensions in large samples. Follow-up protocol studies might then be conducted to zero in on the problem-solving processes underlying any detected differences. For example, recent work in our laboratory (Berger, 1995; Berger & Katz, 1994) suggests that subtle changes in an item's stem can predictably affect subjects' choice of traditional versus nontraditional strategies.

CR items are preferred over MC by many in the education community because the former are believed to measure more important skills, be more relevant to applied decision making, better reflect changing social values, and have more positive social consequences (Bennett, 1993). With respect to measurement differences, CR formats are argued to measure higher-order reasoning abilities, while the MC format taps lower-level procedural skills or factual knowledge. Our analysis of mathematical word problems suggests that the situation is more complex. In many instances, the same reasoning procedures are used regardless of format. Even when differences do occur, the problem-solving approaches encouraged by MC may improve the measurement of reasoning skills as often as they detract from it.

Although there may be good reasons for using CR items in large-scale testing programs, from a cognitive perspective, MC items of the sort studied here should provide measurement that is generally comparable to stem-equivalent CR items. Simply removing the options from a MC item to create a CR counterpart will probably lead neither to better measurement of mathematical skill nor to the assessment of new skills. Our results suggest that more radical changes to existing tests may be required to achieve improved measurement.

References

Bennett, R. E. (1993). On the meanings of constructed response. In R. E. Bennett & W. C. Ward (Eds.), Construction versus choice in cognitive measurement. Hillsdale, NJ: Erlbaum.

Bennett, R. E., Rock, D. A., & Wang, M. (1991). Equivalence of free-response and multiple-choice items. Journal of Educational Measurement, 28, 77-92.

Bennett, R. E., & Sebrechts, M. M. (in press). Measuring the representational component of quantitative proficiency. Journal of Educational Measurement.

Berger, A. E. (1995). Solving mathematics word problems: Problem features affect strategy choice. Unpublished doctoral dissertation, Teachers College, Columbia University.

Berger, A. E., & Katz, I. R. (1994, November). Effects of problem features on strategy selection in mathematics word problems. Poster presented at the thirty-fifth annual meeting of the Psychonomic Society, St. Louis, MO.

Braswell, J. (1990, April). An alternative to multiple-choice testing in mathematics for large-volume examination programs. Presented at the annual meeting of the American Educational Research Association, Boston, MA.

Bridgeman, B. (1992). A comparison of quantitative questions in open-ended and multiple-choice formats. Journal of Educational Measurement, 29, 253-271.

Ericsson, K. A., & Simon, H. A. (1984). Protocol analysis: Verbal reports as data. Cambridge, MA: MIT Press.

Gallagher, A. (1992). Strategy use on multiple-choice and free-response items: An examination of differences among high scoring examinees on the SAT-M (ETS Research Rep. No. 92-54). Princeton, NJ: Educational Testing Service.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. Cognitive Psychology, 15, 1-38.

Martinez, M. E., & Katz, I. R. (1996). Cognitive processing requirements of constructed figural response and multiple-choice items in architecture assessment. Educational Assessment, 3(1), 83-98.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement. New York: Macmillan.

Sebrechts, M. M., Bennett, R. E., & Katz, I. R. (1993). A research platform for interactive performance assessment in graduate education (ETS Research Rep. No. 93-08). Princeton, NJ: Educational Testing Service.

Sebrechts, M. M., Bennett, R. E., & Rock, D. A. (1991). Agreement between expert system and human raters' scores on complex constructed-response quantitative items. Journal of Applied Psychology, 76, 856-862.

Snow, R. E. (1980). Aptitude processes. In R. E. Snow, P. A. Federico, & W. E. Montague (Eds.), Aptitude, learning, and instruction, Vol. 1: Cognitive process analyses of aptitude. Hillsdale, NJ: Erlbaum.

Traub, R. E. (1993). On the equivalence of the traits assessed by multiple-choice and constructed-response tests. In R. E. Bennett & W. C. Ward (Eds.), Construction versus choice in cognitive measurement. Hillsdale, NJ: Erlbaum.

Traub, R. E., & MacRury, K. (1990). [Multiple-choice vs. free-response in the testing of scholastic achievement]. In K. Ingenkamp & R. S. Jäger (Eds.), Tests und trends 8: Jahrbuch der Pädagogischen Diagnostik. Weinheim und Basel: Beltz Verlag.

Ward, W. C., Dupree, D., & Carlson, S. B. (1987). A comparison of free-response and multiple-choice questions in the assessment of reading comprehension (ETS Research Rep. No. 87-20). Princeton, NJ: Educational Testing Service.

Working Groups for the Commission on Standards for School Mathematics of the National Council of Teachers of Mathematics. (1989). Curriculum and evaluation standards for school mathematics. Reston, VA: National Council of Teachers of Mathematics.

Endnotes

1. Delta is a linear transformation of percent correct. It is standardized over an item pool, and it has a mean of 13 and a standard deviation of 4. Higher values indicate greater difficulty.

2. The actual item pool contained an additional four items, which were excluded from all analyses. These items represented a separate subexperiment and so did not conform to the design and materials specifications discussed above.

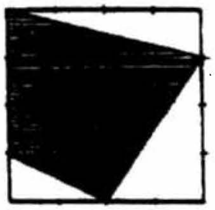
3. The timing factor is omitted from this table because that factor did not contribute to any of the significant effects.

4. All three of the item pairs contributed similarly to this effect. For each item pair, approximately half of the CR erroneous responses did not match any of the corresponding MC alternatives.

5. The grid-in format is an alternative to the typical selection of an answer by filling in a bubble (A-E) on the answer sheet. Instead, examinees indicate their numerical answers to mathematics items by writing the answer directly on the answer sheet, then filling in the corresponding bubbles on a numeric grid. Variations on the grid-in format are described by Braswell (1990).

Appendix

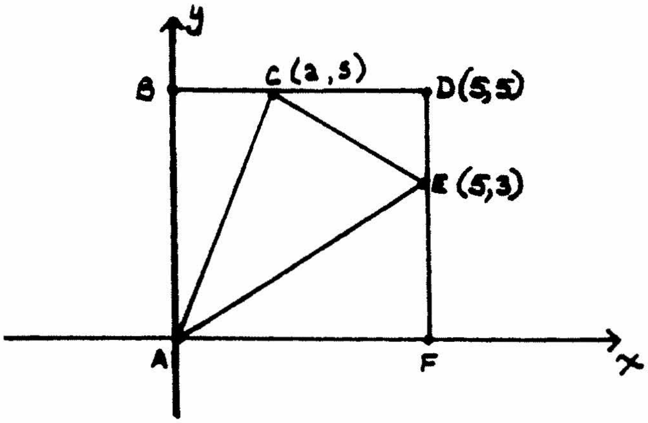
Shaded Item



Each of the sides of the square above is divided into four equal segments. $\frac{\text{Area of shaded region}}{\text{Area of square}} =$

(A) $\frac{1}{4}$ (B) $\frac{3}{8}$ (C) $\frac{1}{2}$ (D) $\frac{5}{8}$ (E) $\frac{3}{4}$

SqTrian Item



What percent of the area of the square ABDF is the area of triangle ACE?

(A) 30% (B) 38% (C) 40% (D) 42% (E) 50%

FIGURE A-1: Shaded/Sqtrian Item Pair

Sack Item

The weight of a sack of grain plus 1/3 of this weight is equal to 24 pounds. What is the weight of the sack of grain, in pounds?

(A) 18 (B) 16 (C) 14 (D) 8 (E) 6

Invest Item

Alan earned 3/4 as much on his investment as Sheila earned on her investment. If together they earned a total of $140, how much did Sheila earn on her investment?

(A) $105 (B) $80 (C) $60 (D) $50 (E) $35

FIGURE A-2: SACK/INVEST ITEM PAIR

Baseball Item

Out of a total of 154 games played, a ball team won 54 more games than it lost. If there were no ties, how many games did the team win?

(A) 94 (B) 98 (C) 100 (D) 102 (E) 104

Students Item

In the senior class at McKinley High School there are 36 more females than males. If this class contains 136 students, how many are female?

(A) 86 (B) 92 (C) 100 (D) 104 (E) 106

FIGURE A-3: BASEBALL/STUDENTS ITEM PAIR

Rain Item

August had 3 times as much rain as July and 1/9 of the rain occurred in July. If there were 28.8 inches of rain during the summer growing season, how many inches of rain fell during June?

(A) 9.6 (B) 12.8 (C) 16.0 (D) 19.2 (E) 25.6

Tank Item

A tank contains 33.6 liters of gasoline. The tank is emptied in 3 days. If 1/7 of the gasoline is used the 1st day, twice that quantity is used on the 2nd day, and the rest is used on the 3rd day, how many liters were used on the 3rd day?

(A) 9.6 (B) 11.2 (C) 19.2 (D) 24.0 (E) 28.8

FIGURE A-4: RAIN/TANK ITEM PAIR

Tickets Item

If 70 tickets to a play were bought for a total of $50.00 and if tickets cost $1.00 for adults and $0.50 for children, how many children's tickets were bought?

(A) 20 (B) 25 (C) 30 (D) 35 (E) 40

Tokens Item

Jenna won a total of 90 red tokens and yellow tokens while playing a board game. Each red token is worth 1 point and each yellow token is worth 4 points. If the total value of Jenna's red and yellow tokens is 120 points, how many yellow tokens does she have?

(A) 10 (B) 18 (C) 30 (D) 60 (E) 80

FIGURE A-5: TICKETS/TOKENS ITEM PAIR

Boxes Item

How many more boxes would be needed to package 1,200 magazines in boxes of 10 than in boxes of 12?

(A) 2 (B) 10 (C) 20 (D) 100 (E) 200

Books Item

In a certain elementary school library, books are stacked 20 to a shelf, while in a nearby high school library, they are stacked 30 to a shelf. How many more shelves would be required to stack 600 books in the elementary school than in the high school?

(A) 5 (B) 10 (C) 15 (D) 30 (E) 40

FIGURE A-6: Boxes/Books Item Pair

Test Item

On a mathematics test, Anita solved 36 out of the 40 problems correctly. What percent of the problems did she solve correctly?

(A) 9% (B) 10% (C) 76% (D) 90% (E) 96%

Company Item

At a certain company, 24 out of the 60 employees have worked there since 1988. What percent of the employees have worked there since 1988?

(A) 4% (B) 10% (C) 30% (D) 40% (E) 84%

FIGURE A-7: Test/Company Item Pair

Price Item

If a department store offers an item originally priced at \$8.00 at a reduced price of \$6.00, by what percent is the original price reduced?

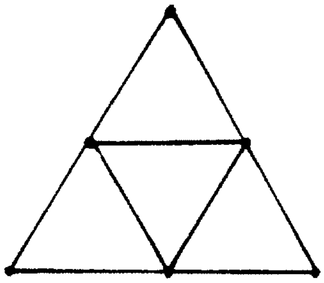
(A) 10% (B) $12\frac{1}{2}$% (C) 20% (D) 25% (E) $33\frac{1}{3}$%

Swim Item

At the beginning of the school year, Mr. Blake had 15 students in his advanced swimming class. By June, he had only 12 students in that class. What percent of the students dropped the class?

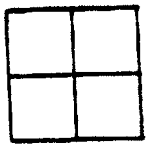
(A) $6\frac{2}{3}$% (B) $8\frac{1}{3}$% (C) 10% (D) 20% (E) 25%

FIGURE A-8: PRICE/SWIM ITEM PAIR



The figure above is an equilateral triangle divided into four equal equilateral triangles. If the perimeter of the large triangle is 3, what is the perimeter of one of the smaller triangles?

(A) $\frac{1}{3}$ (B) $\frac{3}{8}$ (C) $\frac{3}{4}$ (D) 1 (E) $1\frac{1}{2}$



The figure above is a square divided into four equal smaller squares. If the perimeter of the large square is 1, then the perimeter of a small square is

(A) $\frac{1}{16}$ (B) $\frac{1}{8}$ (C) $\frac{1}{6}$ (D) $\frac{1}{4}$ (E) $\frac{1}{2}$

FIGURE A-9: TRIANGLE/SQUARE ITEM PAIR

Garments Item

Lewis has 12 garments in his closet, all of which are shirts and trousers. If 75 percent of the garments are shirts, how many of the garments are trousers?

(A) 3 (B) 4 (C) 5 (D) 9 (E) 10

Work Item

In one 30-day month, David was at work 80 percent of the days. How many days was David not at work that month?

(A) 6 (B) 7 (C) 8 (D) 24 (E) 28

FIGURE A-10: Garments/Work Item Pair