GRE

RESEARCH REPORT

Transfer Between Variants of Quantitative Items

Mary Morley
Brent Bridgeman
René Lawless

October 2004

GRE Board Report No. 00-06R
ETS RR-04-36

Transfer Between Variants of Quantitative Items

Mary E. Morley

College Board, NY

Brent Bridgeman and René R. Lawless

ETS, Princeton, NJ

GRE Board Research Report No. 00-06R

October 2004

The report presents the findings of a
research project funded by and carried
out under the auspices of the
Graduate Record Examinations Board.

Educational Testing Service, Princeton, NJ 08541

****

Researchers are encouraged to express freely their professional judgment. Therefore, points of view or opinions stated in Graduate Record Examinations Board reports do no necessarily represent official Graduate Record Examinations Board position or policy.

****

The Graduate Record Examinations and Educational Testing Service are dedicated to the principle of equal opportunity, and their programs, services, and employment policies are guided by that principle.

Educational Testing Service
Princeton, NJ 08541

Abstract

This study investigated the transfer of solution strategies between close variants of quantitative reasoning questions. Pre- and posttests were obtained from 406 college undergraduates, all of whom took the same posttest; pretests varied such that one group of participants saw close variants of one set of posttest items while other groups saw close variants of other sets. Some participants also saw items that looked similar to some posttest questions but were mathematically different. Between pre- and posttests, some participants viewed solution rationales explaining their incorrect answers. Students at all ability levels performed better on close variants, suggesting there was some positive transfer, but the presence of appearance variants interfered with this transfer. This study suggests an important first step toward understanding how tests and item vats can be designed to capitalize on the extent to which students set up test-taking strategies.

Key words: Mathematical transfer, item models, item variants, automated item generation, quantitative reasoning

Acknowledgements

We would like to acknowledge the involvement of several individuals whose efforts ensured a successful data collection:

- Mike Wagner designed our test delivery and item review systems and managed all technical issues related to this research study.
- Special thanks go to Ed Wolfe of Michigan State University and Elizabeth Baron of Xavier University of Louisiana, whose assistance made this data collection not only possible, but successful as well. Both obtained the computer laboratory space required to administer the study and provided us with the support that we needed to ensure that the computer hardware met our specifications. They also lined up staff to assist us with different aspects of the study, such as computer laboratory configuration, and to administer the study.
- Several test developers reviewed and edited the items in our tests, providing us with valuable quality assurance. Specifically, we would like to thank Beth Brownstein, Darryl Ezzo, Jutta Levin, Judy Smith, and Sheng Wang.

We would also like to acknowledge the hard work and important contributions of:

- Alyson Tregidgo and Venus Mifsud, for their work in subject recruitment
- Lixiong Gu, for managing the scheduling and proctoring at Michigan State University (MSU)
- Ola Rostant and Benita Barnes, for proctoring and design of the MSU instrument
- Yamlak Tsega and Cornelius Marshall, for ensuring that all of our technical needs were met at Xavier University of Louisiana
- Matt Bridgeman, Dan Keys, and Travis Potter, our proctors in Princeton, for administering our study

Table of Contents

List of Tables

List of Figures

Introduction

Problem and Rationale

The introduction of computer adaptive testing (CAT) by ETS® during the 1990s was a bold move, designed to take advantage of the burgeoning information revolution. While model-based psychometrics (e.g., Lord, 1980) and technological foundations had been laid previously to operationalize adaptive testing, test development practices had not yet changed to take advantage of these methodologies. In particular, test development remained "item-centric" rather than model-based.

Nevertheless, some research intended to rethink test development practices took place—primarily in connection with incorporating cognitive principles into the modeling of item difficulty (see Bejar, Embretson, & Chaffin, 1991). In addition, considerable research on the generation of items according to principles and with control of their psychometric attributes was conducted at ETS and elsewhere (see Bejar, 1993, for a summary). In particular, the notion that items could be viewed as instances of a more general class or "model" emerged. Specifically, for mathematics, schema theory was the basis of the Math Test Creation Assistant—one of the first incarnations of the use of item modeling to aid the ETS mathematics test development process. Similar work followed in support of analytical reasoning and verbal item types.

An item model can be thought of as a means of generating close variants with the intention that the isomorphs will be psychometrically and otherwise exchangeable and equivalent (see Bejar et al., 2002). Item writers can create item models manually or through the use of item generation software. In this study, we used existing Graduate Record Examinations® (GRE®) items as the base (or "parent") items for our item models, and hence for producing close variants. We also used each of these base items to produce another kind of variant: "appearance variants"—items that are only superficially similar to their base items. For control purposes, we also included existing GRE items that were matched to the base items only in terms of test specifications (e.g., item difficulty, content category, item type, etc.); we refer to these items as "matched items." Figure 1 depicts the relationship between these different types of items.

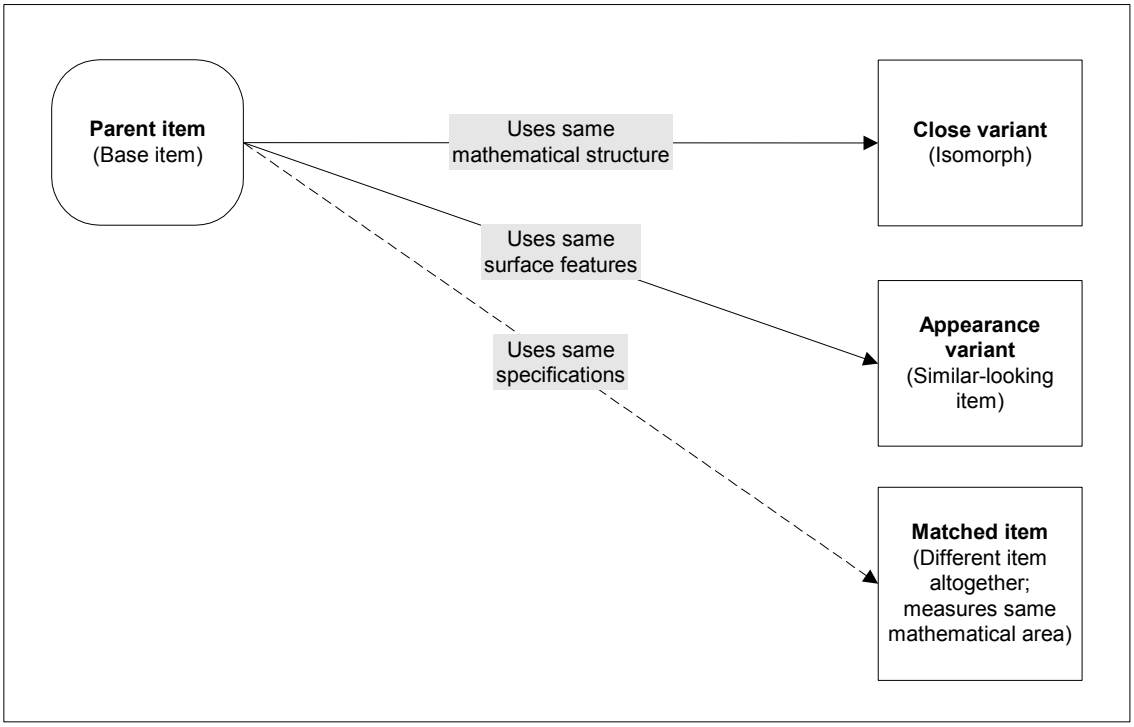

Figure 1. **Relationship between types of items.**

Because close variants are produced from a limited palette of possibilities, they all necessarily share structural and content similarities and therefore cannot be thought of as independent items. It may be possible to use different close variants with different examinees and effectively lengthen the "shelf life" of an item model beyond that of a single item by making it less likely that "pre-knowledge" can improve scores inappropriately. However, what this study seeks to understand is whether that "shelf life" has *actually* been extended, or if in fact the possibility exists that construct-irrelevant strategies (that is, using strategies to derive the correct solution without using the appropriate problem-solving skills) can compromise that item model.

The potential use of construct-irrelevant strategies still raises concerns about the management of variants with respect to automated test assembly. The basis for these concerns is possible similarities among variants that have the same parent. Increased overlap among items in a test pool or vat is a potential threat to test security and score validity. A number of policies have been instituted to minimize the potential negative impact of using item variants. These policies restrict the number of variants that can be created from the same parent as well as the number of variants that may appear in any given CAT pool. Item models are also useful when

many unique linear forms must be produced a year; similar restrictions might apply to the reuse of variants on these forms.

The policies that have already been established are based on assumptions about the similarity of item variants and their potential impact on performance. They assume that examinees are likely to recognize problem variants as similar, and that transfer effects between variants may exist. For instance, if a variant is used in test preparation materials or otherwise exposed, performance on another variant from the same family in a test situation might be better than otherwise expected, thereby compromising the validity of the item's operating characteristics.

There are times when transfer can be valid—such as when an examinee solves a rate problem by first recognizing that it is a rate problem. The problem that item modeling poses is that examinees may memorize rules that are only useful for answering questions belonging to a narrow class and that these memorized rules may lead to correct performance for construct-irrelevant reasons. If such a rule (for example, "If the question involves jelly-beans, just add the two numbers given in the problem to get the answer.") were to work for all items in a variant family, it would compromise the validity of the item's operating characteristics.

A model-based approach to test development requires the formulation of abstract descriptions from which families of similar test questions that are capable of producing many variants may be generated. These descriptions are written such that the psychometric attributes of the instances are well estimated once the model itself is calibrated. While this approach has obvious implications for the efficiency of the test development process, it also has implications for item and test security. The ability to rapidly produce psychometrically equivalent items allays item exposure concerns but does not fully address the issue of the security of the item models. What must be understood is how well students transfer information about test items. This information can then be used to inform how best to author item models, and can also contribute to our understanding of the mathematical attributes that make an item more or less difficult.

To fully appreciate the security implications, it is useful to understand how the modeling process is carried out. A highly discriminating parent item (at a targeted level of difficulty) is selected and used as the basis for an item model. In the resulting item model, the surface features (those item features that do not contribute to difficulty) are varied to produce many possible instances, called isomorphs.

The use of item models might enhance item and test security by making it possible to have the equivalent of a large number of items in an item vat. But perhaps more importantly, through the use of item models it may be possible to forestall score gains due to "backdoor response strategies"[1] and "pre-knowledge"[2] that are independent of abilities we wish to measure. By varying the features of generated items so that there is much less dependency between item appearance and item key, it would be possible to prevent these construct-unrelated gains in performance. Reducing such dependency would enhance validity to the degree that it eliminates the effect of construct-irrelevant transfer on test scores.

Through the examination of performance on different types of variants, this investigation has sought to better understand how well students transfer knowledge about different types of variants. If construct-irrelevant transfer between item variants does not occur or can be forestalled, then restraints on their use could be relaxed, resulting in significant savings for any testing program that must maintain secure item pools.

Review of the Literature

Research on the impact of problem categorization on mathematical problem-solving (Gliner, 1991; Hinsley, Hayes, & Simon, 1977) and on differences in problem categorization related to mathematical ability (Bennett, Sebrechts, & Rock, 1995; Gliner, 1989; Silver, 1979) only partially supports the notion that score increases will result from previous exposure to item models. Students with higher mathematical proficiency tend to recognize structural similarities among problems, while students with less mathematical proficiency tend to base similarity judgments on surface features. Furthermore, students use information about problem categories to guide problem solving. However, it isn't clear how much or what type of training is necessary for students to develop the expertise that would permit transfer among similar problems (Marshall, 1995; Schoenfeld & Herrmann, 1982; Singley, Anderson, & Gevins, 1990).

In order to solve mathematics problems, students must draw an analogy between a newly presented problem and one that they have successfully solved in the past. Early work by Gick and Holyoak (1980) suggests that, "One of the major blocks to successful use of analogy may be failure to spontaneously notice its pertinence to the target problem." Subsequent research has concluded that the surface features of newly presented problems impact students' ability to choose analogous source problems and to activate specific strategies that were used to solve

those problems (Holyoak & Koh, 1987; Bernardo, 2001; Phye, 2001).

One body of research has found that students' expertise is also a critical factor in the ability of students to achieve analogous mathematical transfer. Reed (1987) concluded, "Previous work has shown that experts are better than novices in recognizing isomorphic problems … This distinction is important … because the perceived transparency of the mapping determines the recognition of isomorphs, but the detection that the mapping is nonisomorphic determines the discrimination between problems within a category. The recognition of isomorphs depends on the students' ability to recognize that the concepts in two different problems correspond to one another. The discrimination between two similar problems depends on their ability to find concepts in one problem that do not map onto concepts in the other problem." In the current study, we attempted to replicate these findings through the use of isomorphs and appearance variants.

Novick (1988) demonstrated that students with a deeper understanding of the mathematics behind a presented problem are able to recognize the structural features and are better equipped to correctly solve the problem. Conversely, novices are reliant on the surface features of the new problem, as they are more salient. Further, Novick and Holyoak (1991) best summed up the role of expertise when they stated, "the best predictors of analogical transfer for … [mathematics] problems were mathematical expertise and knowledge of the numerical correspondences required for successful procedure adaptation."

The results of these cognitive studies have several implications in terms of item/test security. On one hand, if students are provided with explanations about the actual mathematics contained in the quantitative section of the GRE General Test, and they subsequently perform well on items as well as item models, we can conclude that the they have recognized structural features of the items and have learned the mathematics—that is, the actual constructs being measured. On the other hand, if after seeing an item students try to apply construct-irrelevant tricks—such as choosing a distractor based on the surface features of an item and the likely position of the key—but they do not succeed in choosing the key, it can be argued that they have not learned the mathematics. In the latter case, the notion of developing item models is further validated and can be used to inform decisions about the use of item models in test development and test administration. Further, if students with higher mathematics ability respond correctly to items created by item models, but students of lower ability do not, again it argues that higher

ability students are more expert at recognizing the structural features of items and that lower ability students rely more heavily on the surface features of items, as suggested by Novick (1988) and Novick and Holyoak (1991). This outcome again supports both the validity of the construct being measured and the security of the item model.

Research Questions

The main goal of this investigation was to answer the following research questions:

1. Will transfer occur between close variants (i.e., items that vary in their surface features but have the same mathematical structure)?
2. When problem solving rationales are provided following an incorrect response, will more transfer occur between close variants than between items that are related only incidentally—"matched" items (i.e., items that measure the same mathematic processes but are expressed differently from their corresponding "base" items)?
3. Will the presence of appearance variants (i.e., items that superficially appear to be similar to their corresponding base items but structurally require different mathematic processes to solve) influence student performances on close variants in the same test?
4. Is transfer related or associated with any student characteristics (e.g., ability level, ethnicity, etc.) or item characteristics (e.g., item format)?

Method

Participants

The target population was comprised of undergraduate college students who had not previously taken the GRE General Test. A variety of methods were used to recruit participants, including college newspaper advertisements, flyers, and referrals from other students. Data were collected at Michigan State University in East Lansing, Xavier University of Louisiana in New Orleans, a CompUSA training center in Philadelphia, and at ETS in Princeton, New Jersey. A total of 406 participants were tested: 283 participants in East Lansing, 41 in New Orleans, 79 in the Princeton area, and three in Philadelphia. At each location, testing was conducted using existing computer laboratories reserved solely for this study, and all test administrations were supervised by trained proctors.

We recruited at schools of different sizes and included an historically Black university. In addition, students attending a variety of colleges and universities participated from the Princeton area. In exchange for their participation, participants were paid $40. In an effort to motivate students to put forth their best effort, they were paid an additional $10 if they met or exceeded their (self-reported) ACT or SAT scores. The extra $10 was used as a tool to motivate the students to render their best performance; scores on the ACT and SAT were not checked. All students were paid the extra $10.

Of the 417 recruits who originally agreed to participate, the records of six participants were lost due to unrecoverable computer problems. Proctors recorded all computer errors on the subject rosters at the testing sites, and examinee performance records were then reviewed to ascertain whether they were usable. Five records were eliminated from the test analyses because proctors observed that these subjects appeared not to take their participation seriously. This assumption was confirmed in three ways: through the examination of the participants' total testing time (which was less than 7 minutes per test); their total number of correct solutions (which was less than 11 per test); and on a plot of the residuals, their data points were determined to be outliers). After eliminating subjects due to the previously mentioned problems, data remained intact for 406 participants. Table 1 describes the sample.

Testing Conditions

A final objective of the experiment was to examine whether or not the provision of solution rationales (i.e., explanations of the mathematics behind test items) after a pretest would impact participants' ability to transfer that information to subsequent problems on a posttest, and to compare this result with simple exposure to items without the provision of rationales. Participants were randomly assigned to one of two conditions to make this determination:

- participants who received rationales for items they answered incorrectly on the pretest.
- participants who did not receive any feedback for items they answered incorrectly on the pretest

Table 1

Demographic Characteristics of Participants

Attribute	Total (n = 406)	Percentage of total
Gender		
Male	146	36%
Female	260	64%
Ethnicity		
Black or African American	59	15%
Asian, Asian American, or Pacific Islander	33	8%
White (non-Hispanic)	293	72%
Other[a]	21	5%
Educational Status		
Freshman	66	16%
Sophomore	99	24%
Junior	133	33%
Senior	101	25%
Other	7	2%
Undergraduate major		
Non-science-based majors	251	62%
Science-based majors	155	38%

[a] Includes American Indian, Alaskan Native, Mexican, Mexican American, Chicano, Puerto Rican, Latin American, and Other Hispanic classifications.

The Experimental Instrument

Six pretest forms and one posttest form—each consisting of 27 items—were developed in linear, computer-based test formats. Each form was designed to span the levels of difficulty typically found on linear, paper-and-pencil GRE test forms. GRE mathematics test developers reviewed all of the close variants and appearance variants for content and correspondence to their respective base items.

The experimental manipulation was accomplished by administering different pretests; all participants took the same posttest. An item in the posttest was classified as a close variant if participants had taken a pretest that contained a close variant of that item. Similarly, a posttest

item was classified as an appearance variant if participants had taken a pretest that included an appearance variant of that item. Six pretest forms were administered to observe whether the presence of different types of item variants would forestall construct-irrelevant transfer. (A detailed explanation of the different types of item variants follows.) Each of these forms contained a different combination of close variants and other items. Because all of the students took the same posttest, any group differences found on the posttest would best be explained by the types of items to which participants were exposed on different pretests.

The Item Pool

The posttest consisted of retired GRE items disclosed in the *GRE Big Book* (currently out-of-print). They were chosen to approximately meet the specifications for an actual, paper-and-pencil GRE quantitative test, except that no data interpretation sets were used.[3] Items selected for the posttest served as base (i.e., parent) items upon which the variants in the pretest were based. The pretest forms were comprised of three types of items: matched items, close variants, and appearance variants, each of which had a close correspondence to a particular base item in the posttest. Figure 2 presents a sample base item. The variants that were developed from the posttest items are defined the paragraphs that follow.

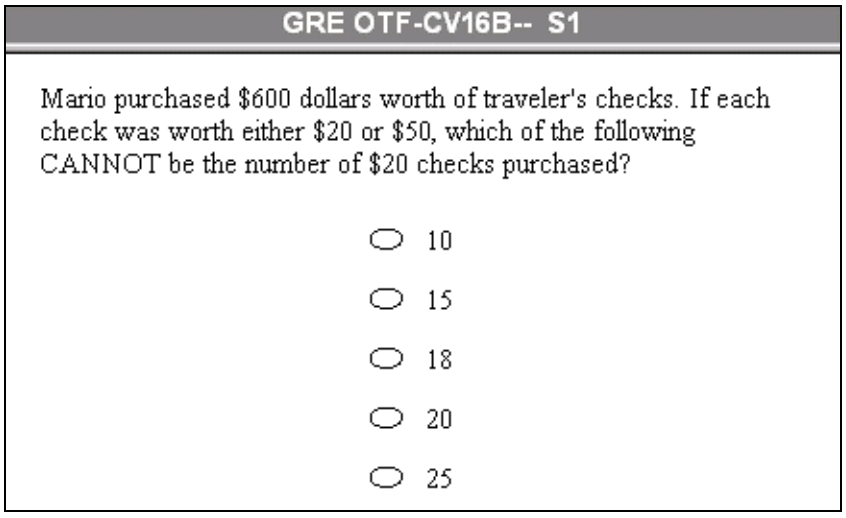

Figure 2. **Sample base (parent) item from posttest.**

Matched items. Matched items, which were also selected from disclosed items in the *GRE Big Book,* were determined to be analogous with parent items based on their

correspondence with each of the following four characteristics:

- GRE item type: either quantitative comparison (in which examinees must decide whether one of two quantities is larger, they are equal, or there is not enough information to decide) or problem solving (five-choice multiple-choice items)
- mathematical area (arithmetic, algebra, geometry, or data interpretation)
- context (pure mathematical item versus word problem)
- difficulty

Full item-response-theory statistics for *GRE Big Book* items were not published; however, the book did provide P^+s (i.e., the percentage of examinees answering each item correctly), which could be used as a measure of each item's difficulty. Matched items were chosen so that the difference between the P^+ of a matched item and its corresponding base item was less than or equal to 0.1. The mean P^+ for all of the base items and matched items was 0.67. Figure 3 shows a sample matched item.

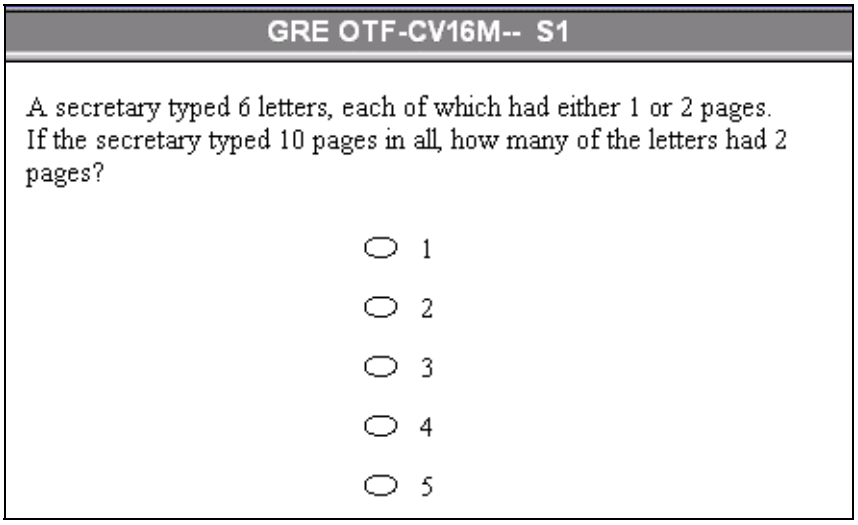

Figure 3. **Matched item to the base item in Figure 2 (matched in terms of mathematical area and difficulty).**

Close variants. Surface features (i.e., names, numbers, and contexts) of base items were altered to create close variants, taking care to preserve the "friendliness" of the base item. These isomorphs were meant to target the same difficulty level as the corresponding base item. Figure 4 presents a close variant of the base item found in Figure 2.

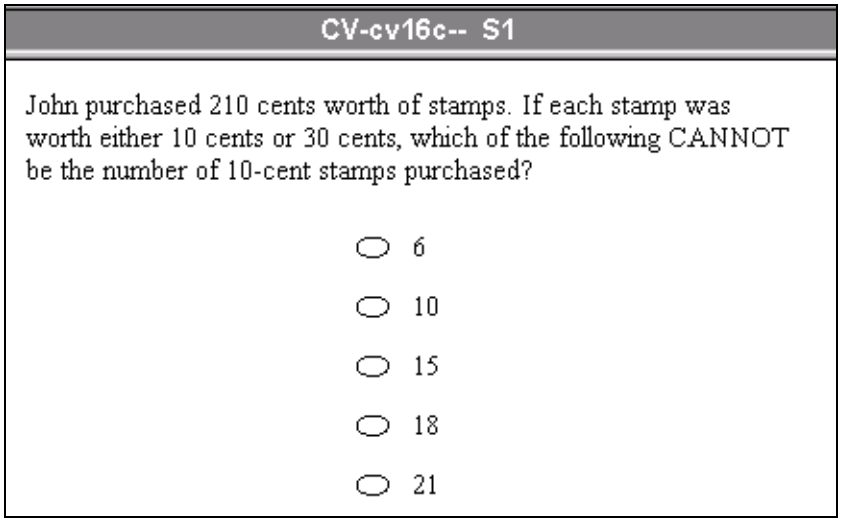

Figure 4. **Close variant (isomorph) of the base item in Figure 2.**

Appearance variants. An appearance variant was also written for each of the base items on the posttest. Appearance variants were designed to look like their corresponding base items, but to differ in some important way. For instance, if the base item showed a figure of a triangle, then the appearance variant showed the same figure. If the base item was a word problem about Mary and her bike, then the appearance variant was also a word problem about Mary and her bike. However, the similarity was only superficial; the underlying mathematics required to solve the appearance variant was changed so that it differed from the mathematics required to solve the base item. Appearance variants were included in the design of the instrument to help reveal whether participants recognized the underlying mathematics required to correctly solve problems or whether they were instead attending to surface features. Figure 5 shows an appearance variant of the base item found in Figure 2.

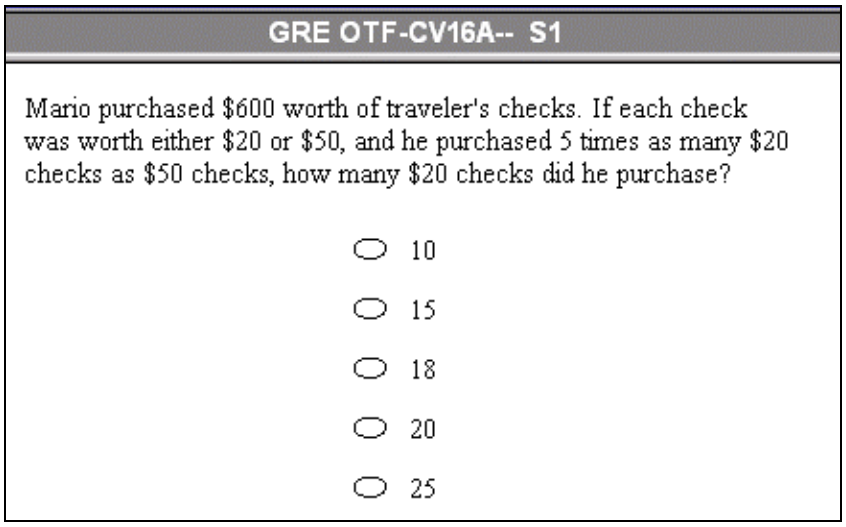

Figure 5. **Appearance variant of the base item in Figure 2.**

Pretest Forms

The 27 base items from the posttest were randomly divided into three nine-item sets (A, B, and C) upon which the item variants were generated. The 27 items were randomly sequenced such that the order of the 27 items in each of the pretest forms was parallel. Participants therefore encountered a variety of item difficulties as well as a mixture of quantitative-comparison and problem-solving items. Table 2 illustrates the configuration of variant types within sets for each particular test form. Thus, the nine items in Set A were close variants for participants who took pretest Forms 1 or 4; the same nine items were matched items for participants who took pretest Forms 2, 3, or 5; and they were appearance variants for participants who took pretest Form 6. Participants were randomly assigned to each of the six-pretest forms.

Table 2

Pretest Form Configurations

Form	Set A	Set B	Set C
1	Close	Matched	Matched
2	Matched	Close	Matched
3	Matched	Matched	Close
4	Close	Appearance	Matched
5	Matched	Close	Appearance
6	Appearance	Matched	Close

This clustering (of close variants, appearance variants, and matched items) was undertaken to determine whether participants performed better on close variants of questions they had previously seen than on completely new questions (matched items). Only half of the forms contained appearance variants so that we could determine whether the presence of such items forestalled construct-irrelevant transfer (i.e., interfered with the student's ability to pick out which items were close variants). For example, as Figure 6 illustrates, the item in position 1 of Form 2 was a matched item (in terms of difficulty) of base item 10 from the posttest; similarly, an appearance variant of base item 10 from the posttest appeared in position 1 of Form 5. Item 2 in both forms was identical: a matched item of base item 11 from the posttest.

Administrative Procedure

At the beginning of the pretest, participants' computers displayed general directions, along with this message: "You may notice that some of the problems in the second test resemble problems in the first test either visually or mathematically, so please read them carefully. The similarity may or may not help you solve the problem." In addition, they were told that they would have 45 minutes to complete each of two 27-item tests.

At the completion of the pretest, participants who had been designated by random assignment to receive rationales were administered feedback for the items they answered incorrectly. At the completion of this section, these participants were administered the posttest. Participants who had not been designated to receive rationales were administered the posttest immediately after completion of the pretest. Upon completion of the posttest, all participants were administered a short survey regarding their attitudes about computerized testing and test fairness.[4] Upon completion of the survey, all participants were given their raw scores for the pre- and posttests.

Pretest Form 2

Position Number	Accession Number
1	CV10**M**
2	CV11**M**
3	CV18**M**
4	CV08**M**
5	CV13**C**
6	CV24**C**
7	CV16**M**
8	CV23**C**
9	CV04**M**
10	CV09**M**
11	CV20**C**
12	CV12**M**
13	CV22**M**
14	CV03**C**
15	CV15**M**
16	CV14**C**
17	CV02**M**
18	CV17**C**
19	CV21**M**
20	CV05**C**
21	CV25**M**
22	CV07**M**
23	CV26**M**
24	CV27**M**
25	CV01**M**
26	CV06**C**
27	CV19**M**

Pretest Form 5

Position Number	Accession Number
1	CV10**A**
2	CV11**M**
3	CV18**A**
4	CV08**A**
5	CV13**C**
6	CV24**C**
7	CV16**A**
8	CV23**C**
9	CV04**M**
10	CV09**A**
11	CV20**C**
12	CV12**A**
13	CV22**M**
14	CV03**C**
15	CV15**M**
16	CV14**C**
17	CV02**M**
18	CV17**C**
19	CV21**M**
20	CV05**C**
21	CV25**M**
22	CV07**M**
23	CV26**A**
24	CV27**A**
25	CV01**M**
26	CV06**C**
27	CV19**A**

Close Variants are marked with a bold **C** and are shown with white boxes

Appearance Variants are marked with a bold **A** and are shown with black boxes

Matched Items are marked with a bold **M** and are shown with gray boxes

Figure 6. ***Example of item sequencing in two pretest forms.***

Results and Discussion

Covariate

Because of the similarity in content between the pretest and the posttest, the pretests would seem to be an ideal covariate. However, there was not a single pretest, but rather a set of six related pretests that varied by the different experimental conditions. Although these tests were intended to be of roughly equal difficulty, they were not identical. Making the additional

assumption that raw scores across forms were related in a linear fashion, we equated them by computing standard scores separately on each pretest and used these standard scores as the ability covariate. The correlations of the ability standard scores with the number of correct responses on the posttest for Sets A, B, and C were .71, .73, and .71, respectively. The purpose of this covariate was simply to reduce the within-group error variance; it could not adjust for ability differences between groups.

Mean Score Differences by Variant Type and Presence of Rationale

The posttest number correct was analyzed separately for each item set. Although everyone was administered tests comprised of all three sets, they were randomly assigned to different experimental conditions for each set. For example, the participants who were administered appearance variants in Set A were administered matched items in Set B and close variants in Set C (see Table 2). A general linear model analysis of variance (GLM ANOVA) was run that included ability (i.e., standardized pretest scores), gender, school major, and ethnicity variables to ensure that there were no significant interactions between these variables and the experimental manipulation (rationales and variant type). The model was then rerun excluding all of the interactions.

For each set, the combined interactions accounted for less than two percent of the variance, so we selected the simpler main-effects-only models. Providing rationales did not have a significant effect in any of the item sets. In Set A, the rationales' F (1, 396) = 2.76, p = .10, and in the other two sets the Fs were less than one (to view the complete ANOVA tables, see the appendix). However, the variants condition was significant at $p < .01$ for all three sets: the Fs (1, 396) for Sets A, B, and C were 5.12, 13.67, and 7.74, respectively.

Table 3 presents the means for the three variant types on the three item sets. For all three sets, Fisher's least significant difference test indicated that appearance variants were more difficult than close variants, and for Sets B and C, close variants were also significantly easier than matched items. For Sets A and B, matched items were significantly easier than appearance variants; in Set C this difference fell just short of the significance criterion ($p = .051$).

As a reminder, lower mean scores imply increased difficulty of an item or item set. Despite the variations from set to set, the general pattern is clear. Having previously seen an item with the same mathematical structure appears to enhance performance, but having seen an item

that *appears* to be similar, but that actually has a different underlying mathematical structure, degrades performance.

Table 3

Mean Number Correct and Sample Size by Variant Type for Three Item Sets

Variant type	Mean correct for three item sets[a]			Sample size for three item sets		
	A	B	C	A	B	C
Appearance	6.32_a *(6.20)*	5.04_a *(4.96)*	4.83_a *(4.67)*	65	69	67
Matched	6.92_b *(6.74)*	5.72_b *(5.60)*	5.23_a *(5.13)*	205	200	206
Close	6.73_b *(6.69)*	6.19_c *(5.98)*	5.65_b *(5.49)*	136	137	133

Note. Complete ANOVA tables can be found in the appendix.

[a] Means shown without parentheses have been adjusted to the mean of the ability covariate; unadjusted means appear in parentheses. Means in the same column that do not share subscripts differ at $p < .05$ (and at $p < .01$ in the B and C sets) by the Fisher least significant difference test. Mean square errors are 1.72, 2.19, and 2.05 for Sets A, B, and C, respectively.

Item Difficulties for Close Variants, Appearance Variants, and Matched Items

Figures 7-9 show item difficulties (P^+) for each posttest item by pretest group in Sets A, B, and C, respectively. Within sets, the position in which each item was administered is specified, and items are identified as being one of two item types: problem solving (PS: five-choice multiple-choice items) or quantitative comparison (QC: in which examinees must decide which of two quantities is larger, whether they are equal, or whether there is not enough information to decide). Across item types, appearance variants were clearly the most difficult. For seven out of nine items in both Set A and Set B, the lowest P^+ was for appearance variants; in Set C appearance variants were most difficult in five out of the nine items. Close variants were easier than matched items for four items in Set A and for six items in each of Sets B and C.

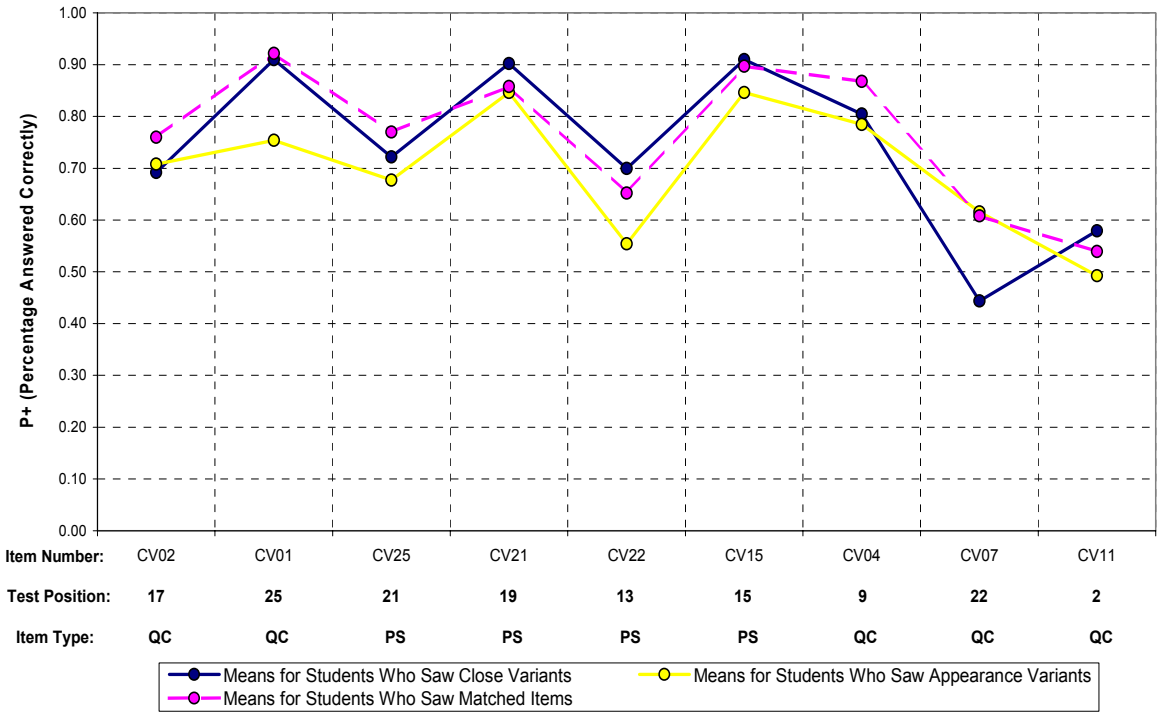

Figure 7. **Participant performance on posttest, grouped by type of item seen first (Set A).**

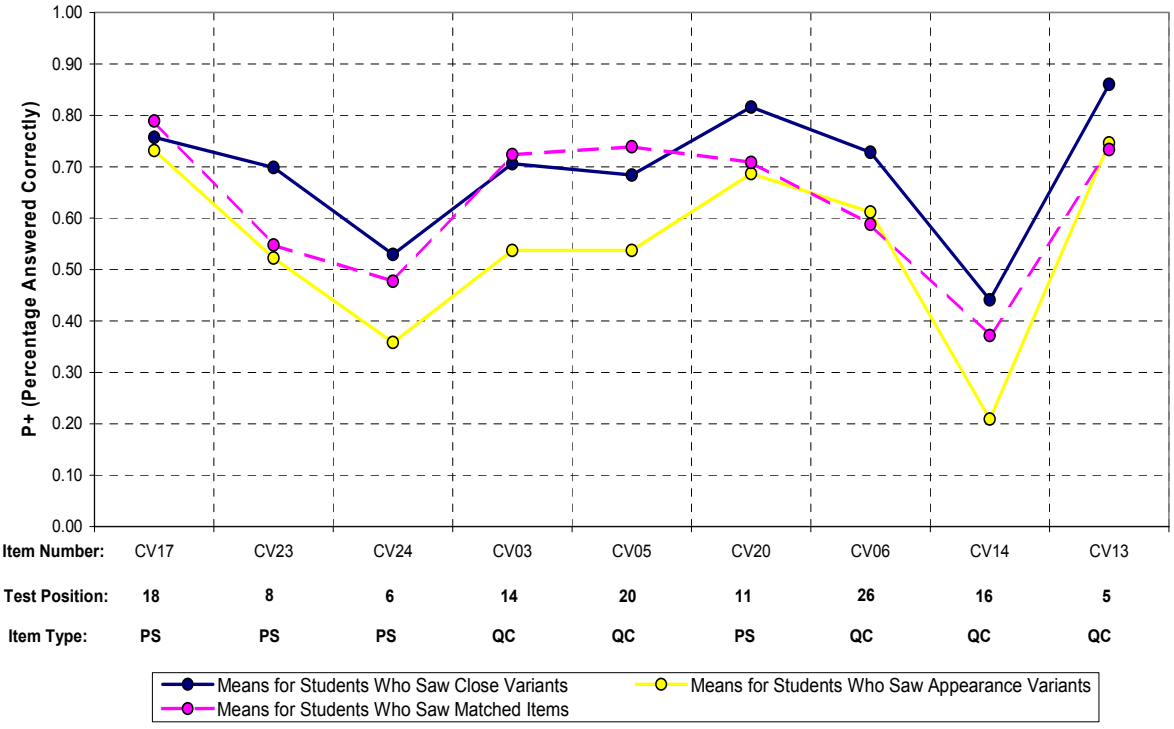

Figure 8. **Participant performance on posttest, grouped by type of item seen first (Set B).**

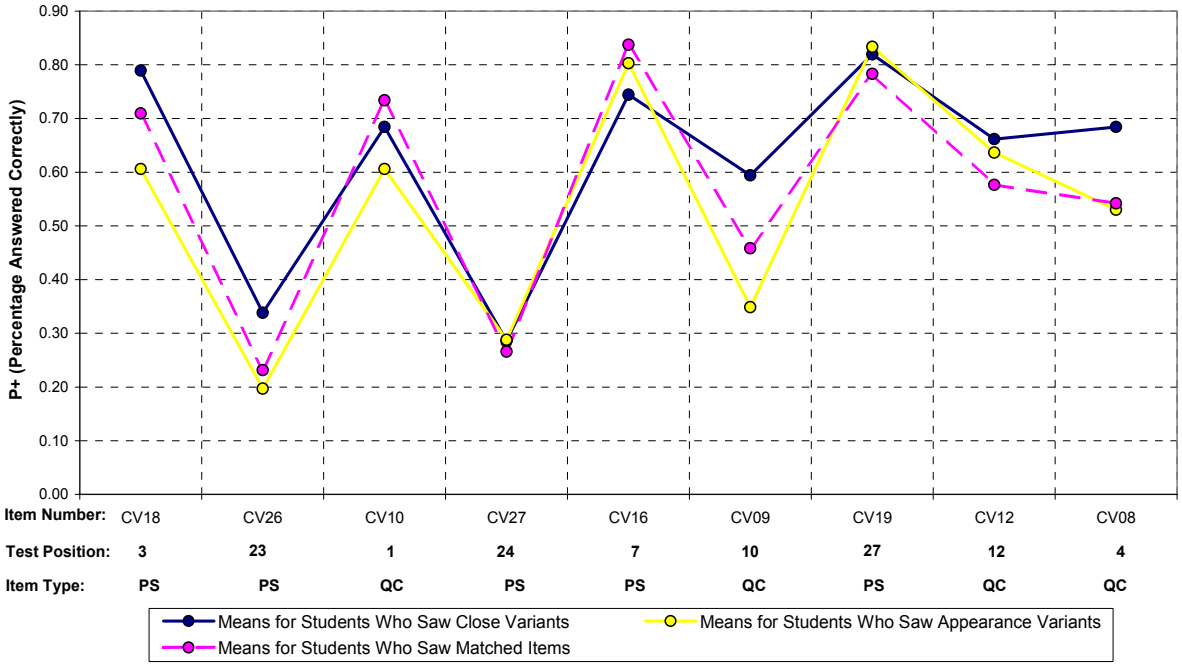

Figure 9. **Participant performance on posttest, grouped by type of item seen first (Set C).**

Scores on Close Variants in Test With or Without Appearance Variants

The above analysis compared performance on the three variant types separately. An additional issue is whether the presence of appearance variants in a test affects performance on close variants. Half of the participants took pretests that contained appearance variants and half had no appearance variants in their pretests. Mean scores on the three sets of close variants for these groups are presented in Table 4. Unlike Table 3, there is no overlap in these groups; each of the six means comes from a different group of participants. ANOVAs (gender x ethnicity x major x ability x rationale x presence of appearance variants) in each item set indicated no significant interactions, so the simple model was retained. As indicated in the table, presence of appearance variants had a significant effect in two of the three item sets [in Set *A*, $F(1, 127) = 0.109$, $p = .741$; in Set *B*, $F(1, 128) = 7.36$, $p = .008$; in Set *C*, $F(1, 124) = 5.33$, $p = .023$]. We do not know why there was no effect in Set *A*, though it should be noted that this was the easiest set. It is possible that appearance variants had an effect only on relatively difficult items.

Table 4

Mean Number Correct on Close Variants When Appearance Variants Are or Are Not Included in the Test for Three Item Sets

Appearance variants in test	Mean correct on close variants for three item sets			Sample size for three item sets		
	A	B	C	A	B	C
Yes	6.61 *(6.55)*	5.63 *(5.28)*	5.44 *(5.13)*	67	67	65
No	6.68 *(6.88)*	6.25 *(5.92)***	5.94 *(5.33)**	69	70	68

Note. Means adjusted to the mean of the ability covariate (unadjusted means appear in parentheses). Mean square errors are 1.82, 7.76, and 1.55 for sets *A*, *B*, and *C,* respectively.

*$p < .05$ (comparing Yes and No within set).

**$p < .01$ (comparing Yes and No within set).

Summary and Conclusion

On the tests assembled for this study, participants performed better on close variants, indicating some positive transfer from their pretest experience; however, appearance variants appeared to interfere with this transfer. This result demonstrates that the presence of related appearance variants in tests containing close variants causes interference with student performance. This discovery suggests that tests and item vats can be designed to capitalize on the extent to which students set up test-taking schemas. By administering tests containing both of these types of variants, constructs can be better measured and students' use of construct-irrelevant strategies may be forestalled.

By the very nature of mathematics, many items appear to be similar but differ mathematically. Thus, it is not surprising that many "accidental" isomorphs and appearance variants occur in the current GRE item vat. These "accidental" families of isomorphs can be grouped together to serve as parent items for item models. An economical approach to item modeling might involve first writing item models that produce close variants, paired with writing item models that generate appearance variants of the first model. In essence, the two models would produce appearance variants of each other. This approach could be used to prevent construct-irrelevant transfer. In addition, by producing item models in this fashion, item shelf life

may be extended.

Because participants showed improvement on the posttest, the question of student learning still arises. Perhaps it may be explained as a warm-up effect. It is clear from this study that exposure to solution rationales had no significant impact on performance. Although we realize that the presentation of rationales has no parallel with formal test preparation, it can tell us whether brief exposure to correct solutions can impact student performances. Perhaps students who received rationales were unable to transfer that information to posttest items because they were not familiar enough with the mathematical concepts presented in the rationales, or perhaps the brief explanations were not meaningful enough.

These results raise the question of whether more prolonged or extended teaching or coaching could impact student performance and thus compromise item and test security. Taking the prudent approach, it should be assumed that, if an item modeling approach is adopted, test preparation schools may soon alter their curricula to include instruction in item modeling. These schools could not only teach item models, but could also make students aware of the existence of appearance variants and help them discriminate between appearance variants and close variants. This, in turn, could force coaching schools to focus more on teaching the mathematics underlying items, leading to construct-*relevant* improvements in student performance, which would not be a security concern.

The accessibility of the Internet opens up an additional arena for potential security compromises: Web sites devoted to the unauthorized disclosure of items. It is strongly suggested, therefore, that further research be conducted to ascertain the vulnerability of item models and empirically address such security concerns. Whether coaching can help students differentiate between close variants and appearance variants remains to be seen. Policymakers and test developers can use the results of such research to guide the use of an item modeling approach when developing tests and as a safeguard against item theft.

Because item security is at the very heart of test validity and fairness, it is paramount to ensure that construct-irrelevant strategies do not influence students' test scores. The use of "backdoor" strategies such as item memorization and construct-irrelevant associations between items and keys may be hard to detect and may compromise item security; they should be examined in future research. One such avenue to explore is whether these results are consistent across other item types, such as the numerical-response, which may be added to the GRE

General Test in the future. This research could be combined with a study that explores the impact of prolonged instruction in multiple-choice solution strategies. The results of such a study, when considered with the results of the current study, would provide important empirical evidence of how item modeling can be used to help ensure test security.

References

Bassok, M., & Holyoak, K. J. (1989). Interdomain transfer between isomorphic topics in algebra and physics. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(1), 153-166.

Bejar, I. I. (1993). A generative approach to psychological and educational measurement. In N. Frederiksen, R. J. Mislevy & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (1st ed., pp. 323-357). Hillsdale, NJ: Lawrence Erlbaum.

Bejar, I. I., Embretson, S., & Chaffin, R. (Eds.). (1991). *Cognitive and psychometric analysis of analogical problem solving.* New York: Springer-Verlag. (Also published as GRE Board Report No. 84-19. Princeton, NJ: ETS.)

Bejar, I. I., Lawless, R. R., Morley, M. E., Wagner, M. E., Bennett, R. E., & Revuelta, J. (2002). *A feasibility study of on-the-fly item generation in adaptive testing* (GRE Board Professional Report No. 02-23). Princeton, NJ: ETS.

Bennett, R. E., Sebrechts, M. M., & Rock, D. A. (1995). *A task type for measuring the representational component of quantitative proficiency* (GRE Board Professional Report No. 92-05P). Princeton, NJ: ETS.

Bernardo, A. B. I. (2001). Analogical problem construction and transfer in mathematical problem solving. *Educational Psychology, 21*(2), 137-150.

Carroll, J. B. (1976). Psychometric tests as cognitive tasks: A new "structure of intellect." In L. Resnick (Ed.), *The nature of intelligence* (pp. 27-56). Hillsdale, NJ: Lawrence Erlbaum.

ETS. (1998). *GRE: Practicing to take the General Test big book.* Princeton, NJ: Author.

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology, 12*, 306-355.

Gliner, G. S. (1989). College students' organization of mathematics word problems in relation to success in problem solving. *School Science and Mathematics, 89*(5), 392-404.

Gliner, G. (1991). College students' organization of mathematics word problems in terms of mathematical structure vs. surface structure. *School Science and Mathematics, 91*(3), 105-110.

Hinsley, D. A., Hayes, J. R., & Simon, H. A. (1977). From words to equations: Meaning and representation in algebra word problems. In M. A. Just & P. A. Carpenter (Eds.), *Cognitive processes in comprehension* (pp. 89-106). Hillsdale: Lawrence Erlbaum.

Holyoak, K. J., & Koh, K. (1987). Surface and structural similarity in analogical transfer. *Memory and Cognition, 15*(4), 332-340.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Marshall, S. P. (1995). *Schemas in problem solving*. New York: Cambridge University.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice Hall.

Novick, L. R. (1988). Analogical transfer, problem similarity, and expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*(3), 510-520.

Novick, L. R., & Holyoak, K. J. (1991). Mathematical problem solving by analogy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*(3), 398-415.

Phye, G. D. (2001). Problem-solving instruction and problem-solving transfer: The correspondence issue. *Journal of Educational Psychology, 93*(3), 571-578.

Reed, S. K. (1987). A structure-mapping model for word problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*(1), 124-139.

Reed, S. K., Dempster, A., & Ettinger, M. (1985). Usefulness of analogous solutions for solving algebra word problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*(1), 106-125.

Schoenfeld, A. H., & Herrmann, D. J. (1982). Problem perception and knowledge structure in expert and novice mathematical problem solvers. *Journal of Experimental Psychology, 8*(5), 484-494.

Silver, E. A. (1979). Student perceptions of relatedness among mathematical verbal problems. *Journal for Research in Mathematics Education, 10*(3), 195-210.

Singley, M. K., Anderson, J. R., & Gevins, J. S. (1990). *Promoting abstract strategies in algebra word problem solving* (IBM Research Report No. 15861). Yorktown Heights, NY: IBM T. J. Watson Research Center.

Sternberg, R. J. (1977). *Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities*. New York: Wiley.

Whitely, S. E. (1980). Latent trial models in the study of intelligence. *Intelligence*, *4*(2), 97-132.

Notes

[1] A "backdoor response strategy" is a simple procedure that does not require a high level of ability and leads to the key or at least the elimination of distractors.

[2] Pre-knowledge refers to advanced knowledge as to which items will appear on the test.

[3] Because data interpretation sets are comprised of multiple items using the same stimulus, there are two difficulties with modeling. First, the amount of data they contain requires data structures not currently available in the Math Test Creation Assistant. Second, many data sets are based on real data that cannot be varied (e.g., U. S. unemployment rates in the year 1998). Another reason for not including data interpretation sets was that ETS research is currently exploring the use of item modeling in discrete items and not item sets.

[4] This survey was administered on behalf of Edward Wolfe of Michigan State University as part of a collaboration that provided free computer laboratory space for the administration of this study at that testing site.

Appendix

Complete Analysis of Variance Tables

Table A1

Analysis of Variance for Reduced Model for Set A (Items 1-9)

Source	*df*	*F*	η^2	*p*
Gender	1	0.31	.00	.58
Test condition (Rationale/No rationale)	1	2.76	.01	.10
Ethnicity	3	0.63	.01	.60
Major	1	0.01	.00	.93
Variant type	2	5.12**	.03	.01
Ability	1	322.88**	.45	.00
Error	396	(1.72)		
Total	406			

Note. R^2 = .522 (Adjusted R^2 = .512)

* p < .05 ** p < .01

Table A2

Analysis of Variance for Reduced Model for Set B (Items 10-18)

Source	*df*	*F*	η^2	*p*
Gender	1	2.90	.01	.09
Test condition (Rationale/No rationale)	1	0.47	.00	.49
Ethnicity	3	2.71*	.02	.05
Major	1	7.65**	.02	.01
Variant type	2	13.67**	.07	.00
Ability	1	339.75**	.46	.00
Error	396	(2.19)		
Total	406			

Note. R^2 = .581 (Adjusted R^2 = .571).

* p < .05 ** p < .01

Table A3

Analysis of Variance for Reduced Model for Set C (Items 19-27)

Source	*df*	*F*	η^2	*p*
Gender	1	0.65	.00	.42
Test condition (Rationale/No rationale)	1	0.58	.00	.45
Ethnicity	3	2.56	.02	.06
Major	1	13.22**	.03	.00
Variant type	2	7.74**	.04	.00
Ability	1	292.82**	.43	.00
Error	396	(2.05)		
Total	406			

Note. $R^2 = .542$ (Adjusted $R^2 = .532$).

* $p < .05$ ** $p < .01$

GRE-ETS
PO Box 6000
Princeton, NJ 08541-6000
USA

To obtain more information about GRE programs and services, use one of the following:
Phone: 1-866-473-4373
(U.S., U.S. Territories*, and Canada)
1-609-771-7670
(all other locations)
Web site: www.gre.org

* America Samoa, Guam, Puerto Rico, and US Virgin Islands

I.N. 726113