GRE®
RESEARCH

Factors Affecting Difficulty in the Generating Examples Item Type

Irvin R. Katz
Audrey W. Lipps
J. Gregory Trafton

April 2002

GRE Board Professional Report No. 97-18P

ETS Research Report 02-07

ETS® Educational Testing Service
Princeton, NJ 08541

Factors Affecting Difficulty in the Generating Examples Item Type

Irvin R. Katz
Educational Testing Service

Audrey W. Lipps
George Mason University, Fairfax, Virginia

J. Gregory Trafton
Naval Research Laboratory, Washington, DC

GRE Board Report No. 97-18P

April 2002

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

Educational Testing Service, Princeton, NJ 08541

Educational Testing Service
Princeton, NJ 08541

Abstract

This paper investigates the predictive validity of various features of Generating Examples (GE) test items – algebra problems that pose mathematical constraints and ask examinees to provide example solutions meeting those constraints. Selection of item features was motivated by a cognitive model of how examinees solve GE items using informal solution strategies such as generate-and-test. Experiment 1 examined the extent to which examinee performance can be explained by features predicted to affect difficulty, and Experiments 2 and 3 investigated the generality and cognitive bases of the difficulty model. The factors studied accounted for approximately 55% of the variance among item difficulty levels, and this predictive power was maintained on a more heterogeneous set of items. Cognitive strategies underlying the difficulty factors were also examined.

Key words: Quantitative reasoning, underdetermined problems, assessment of quantitative skills, new item types, constructed response, Generating Examples, generate-and-test

Acknowledgements

This research was supported by the Graduate Record Examination Board and Educational Testing Service, with in-kind contributions from George Mason University and the Naval Research Laboratory. We thank Margaret Redman and Finis Collins for providing research and administrative assistance, Mary Morley for creating the experimental items, Dennis Quardt for crafting and applying the automatic scoring routines, and Karla Hoffman for accessing and compiling information from the Graduate Record Examinations (GRE®) database. We thank Erik Altmann, Malcolm Bauer, Randy Bennett, Brent Bridgeman, Patrick Kyllonen, Don Powers, Christian Schunn, Lawrence Stricker, and Susan Trickett for their insightful comments on previous presentations of this work. Finally, we thank the GRE graduate assistants and Adisack Nhouyvanisvong for their assistance with data collection.

Table of Contents

List of Tables

Page

List of Figures

Page

Introduction

The Generating Examples (GE) item type asks examinees to supply example solutions that satisfy a set of mathematical constraints. These constraints do not determine a unique solution, so GE items can have multiple correct solutions. Previous research on this item type has investigated the types of skills it elicits – compared with the types of skills elicited by more standard algebra items – both through cognitive (Nhouyvanisvong, Katz, & Singley, 1997) and psychometric analyses (Bennett, Morley, Quardt, Rock, & Katz, 1999a; Bennett et al., 1999b). Investigations of item factors that affect the difficulty of GE items have focused largely on simple structural characteristics (e.g., number of solutions requested or instructions on the solvability of the items). Investigations of item factors motivated by cognitive models of how examinees solve GE items (e.g., the type of constraints posed) have been limited to post hoc task analyses of items. These items were created based on the intuitions of test development staff rather than using a systematic approach designed to test the cognitive models.

In order to examine the predictive validity of difficulty factors derived from a cognitive modeling of solving GE items, we conducted three experiments. In the first experiment, we systematically manipulated difficulty factors to create a set of experimental items, then examined the extent to which examinee performance (both accuracy and solution time) can be explained by the factors. In the next two experiments, we investigated the generality and cognitive bases of the difficulty model. Experiment 2 is a partial replication of Experiment 1, using a less systematically created set of items (i.e., more closely duplicating GE items as they might appear in a test of quantitative reasoning). As a means to explain the mechanisms by which difficulty factors affect performance, Experiment 3 explicates some of the heuristics examinees use in solving GE items.

In the next two sections of this report, we provide examples of GE items and the solution strategies examinees were observed to use in previous work. We then describe the difficulty factors investigated in the current work.

Underdetermined Problems

Underdetermined problems, such as GE items, do not provide all the information necessary to find a unique solution to a problem. That is, it is impossible to apply a *standard* algebraic solution strategy to determine a correct solution to such a problem (i.e., a strategy in which the problem solver isolates and solves for one unknown variable, such as X = 22). Some underdetermined problems may involve inequalities that allow the problem solver to use algebraic manipulation to isolate a variable with an inequality (e.g., X > 20). However, this final derived inequality does not constitute a solution. The problem solver must still generate a solution based on that inequality. Thus, underdetermined problems require problem solvers to generate examples of *potential* solutions.

Figure 1 provides an example of an underdetermined problem. In this case, the problem solver must produce example values for the number of standard and commemorative stamps that satisfy the explicitly stated constraints shown below the prompt (an actual item would show just the story prompt).

Jamal wants to buy some standard-issue stamps at $0.25 each and some commemorative stamps at $0.40 each. He wants more than 25 stamps, including at least two of each, but can spend no more than $10.00. What is one possibility for the number of standard-issue stamps and commemorative stamps Jamal could buy?

S, C are integers > 1
S + C > 25
0.25S + 0.40C = 10.00

Figure 1. **A Generating Examples (GE) item.**

The three constraints shown in Figure 1 do *not* determine a unique solution to the problem. Several values for the number of each type of stamp satisfy these constraints. A problem solver might proceed by first identifying the explicit constraints given in the problem. Taking into consideration all the constraints that have been identified, both explicit and implicit, the problem solver must propose a candidate solution and evaluate it against the constraints. A successful example is found if the proposed solution does not violate any problem constraint.

It is this reasoning with regard to constraints that makes GE items seemingly different from the more traditional, well-determined word problems. Unless the examinee detects, prioritizes, and sets the constraints by specifying appropriate values, the problem cannot be solved. The underdetermined nature of these problems makes them potentially more realistic than many other item types, as most real problems are themselves not well formulated (Frederiksen, 1984).

Generate-and-Test

Underdetermined items such as GE are solved through a combination of algebraic simplification (i.e., combining the posed constraints) and a strategy called "generate-and-test" (Nhouyvanisvong, Katz, & Singley, 1997). Generate-and-test is a heuristic that at first blush may seem hopelessly unlikely to lead to a solution. Yet, in the domain of algebra word problems, many students use generate-and-test effectively. In the generate-and-test strategy, a student chooses a possible answer, then checks whether that answer satisfies the constraints of the problem. If the candidate answer does not fit, the student generates a new one.

Until recently, informal problem-solving strategies such as generate-and-test received little attention by cognitive or psychometric researchers. Much of the research on algebra word-problem solving instead focused on formal representations and procedures (i.e., equation posing and solving) used by students when solving traditional, single-solution word problems (e.g., Kintsch & Greeno, 1985; Mayer, Larkin, & Kadane, 1984; Paige & Simon, 1966). Yet, generate-and-test is more prevalent than the previous research implies. A form of this heuristic (called "guess and check") is even taught by some algebra instructors. In several studies of algebra problem solving, Katz and Berger (1995) found that participants (high school students) used generate-and-test on approximately 50% of problems presented by the researchers. Generate-and-test is not necessarily a "fallback" strategy for weaker students, but is used effectively by students with good math skills (Katz, Bennett, & Berger, 2000; Tabachneck, Koedinger, & Nathan, 1995). Other researchers have similarly noted the prevalence of generate-and-test and other "informal" strategies (e.g., Hall, Kibler, Wenger, & Truxaw, 1989; Koedinger & Tabachneck, 1994).

The underdetermined nature of GE items requires that, to arrive at an answer using generate-and-test, an examinee must generate a potential solution and then test it against the constraints in the

problem. Thus, in order to understand GE items and the factors that affect their difficulty, one should start with a cognitive model of the generate-and-test strategy.

A Model of the Generate-and-Test Strategy

Nhouyvanisvong, Katz, and Singley (1997) developed a computational cognitive model of examinee application of the generate-and-test strategy. Their model was based on analyses of verbal protocols generated while solving both well-determined (i.e., standard algebra) and underdetermined (i.e., GE) algebra problems. Research participants used the skills associated with generate-and-test (e.g., estimating a solution, updating an estimate, inferring implicit constraints from an item stem) when solving both standard algebra and GE items. However, the researchers assert that, by virtue of their underdetermined nature, GE items might place greater emphasis on these skills, which are not well represented in existing assessments. Thus, GE items "… might not require unique skills, but skills that do not often come to the fore (or perhaps come to the fore inappropriately) in well-determined problems" (p. 8).

Figure 2 presents a simplified version of the generate-and-test model (adapted from Nhouyvanisvong & Katz, 1998). The model begins by estimating a value for one of the desired variables (e.g., the number of standard-issue stamps), using half of the potential maximum value for that variable (here either 12 or 13). This heuristic of using half the maximum value was derived by Nhouyvanisvong and Katz from analyses of verbalizations that students made while solving GE items. Using this estimate, the model derives the value of any other variables (e.g., number of commemorative stamps) and then tests whether these values satisfy the remaining constraints in the model (i.e., propagating the estimate). If an inconsistency is found, a new estimate is generated; if all constraints are satisfied, a solution is reported.

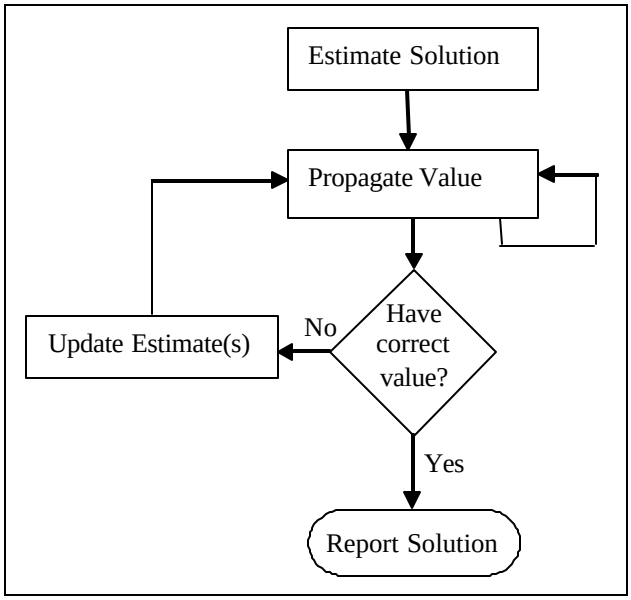


Figure 2. **A flowchart model of the generate-and-test strategy.**

Difficulty Factors

Based on this cognitive model of the generate-and-test strategy, we can derive factors predicted to affect performance. For the purposes of the research presented in this report, we focused on two of the model's processes – estimating a solution and propagating an estimate – to derive two difficulty factors: (a) the types of constraints posed by the problem and (b) the density of correct solutions.

Type of Constraint

According to the model, an item should be difficult to the extent that it requires a potential solution to satisfy more constraints – that is, difficulty should increase with more "testing" in a generate-and-test model. A greater number of tests should be associated with greater difficulty both because of the greater opportunity to make an error and because of the increased load on cognitive resources. However, as noted in previous work (Nhouyvanisvong & Katz, 1998; Bennett et al., 1999a), not all constraints increase the amount of testing. Nhouyvanisvong and Katz distinguish between *generator* and *verifier* constraints. Generator constraints are used by the model to estimate a potential solution, whereas verifier constraints are used to test the accuracy of an estimate.

According to the model, generator constraints merely constrain the initial estimate – simplifying an already simple "constrained guessing" process – and so have relatively little effect on difficulty. For example, given the constraint that X < 30, any estimates generated would be less than 30. Because the constraint was used to generate the estimate, it would be redundant to again check that this constraint is satisfied by the estimate. Figure 3 shows the item in Figure 1 with one type of additional generator constraint; the items in Figure 1 and Figure 3 are predicted to be equal in difficulty.

Jamal wants to buy some standard-issue stamps at $0.25 each and some commemorative stamps at $0.40 each. He wants more than 25 stamps, including at least two of each, but can spend no more than $10.00. **If he wants more than 14 of the commemorative stamps**, what is one possibility for the number of standard-issue stamps and commemorative stamps Jamal could buy?

S, C are integers > 1
S + C > 25
0.25S + 0.40C = 10.00
C > 14

Figure 3. **Previous GE item with one additional generator constraint.**

As noted above, verifier constraints are used to check the correctness of an estimated solution. Thus, more verifier constraints should translate to a greater number of tests, resulting in a more difficult item. Each test of whether a constraint is satisfied represents an opportunity for an examinee to make a mathematical error and either wrongly accept an incorrect solution or mistakenly reject a correct one. Figure 4 shows the item in Figure 1 with one type of additional verifier constraint; the item in Figure 4 is predicted to be the more difficult of the two.

Jamal wants to buy some standard-issue stamps at $0.25 each and some commemorative stamps at $0.40 each. He wants more than 25 stamps, including at least two of each, but can spend no more than $10.00. **If the number of standard-issue stamps must be greater than 10 times the number of commemorative stamps**, what is one possibility for the number of standard-issue stamps and commemorative stamps Jamal could buy?

S, C are integers > 1
S + C > 25
0.25S + 0.40C = 10.00
S > 10C

Figure 4. **Previous GE item with one additional verifier constraint.**

The particular forms of the generator and verifier constraints added above are not meant to be all-inclusive. For example, another of the constraints in the example item might also be considered a type of generator constraint: S + C > 25. With little effort, an examinee might make the assignment of S = 5 and C = 20, effectively splitting the constant across the two variables (incorrectly, as it turns out in this case, because of the inequality). Other apportionings of the constant are also possible (e.g., 13 and 13). This type of constraint (with only unit coefficients) may be considered a generator constraint because examinees use it to come up with potential solutions rather than to verify a potential solution. Experiment 3 provides evidence that these types of constraints do indeed serve as generators.

Similarly, the other inequality in the example item, 0.25S + 0.40C = 10.00, is clearly a type of verifier constraint, in that examinees would use it to check the correctness of a potential solution. Consider the item in Figure 5, which has two verifier constraints. According to the model, because this item has more verifier constraints than the item in Figure 1, it should be more difficult.

An organization is arranging transportation for 960 people, by van or by bus. Each bus can transport 32 people and costs $160, each van can transport 8 people and costs $48. The organization can pay no more than $5280 for transportation. What is one possibility for the number of vans and buses the organization could use?

B, V are positive integers
$32B + 8V \geq 960$
$160B + 48V = 5280$

Figure 5. **GE item with two verifier constraints.**

These items demonstrate the power of using a cognitive model to guide difficulty modeling. A cognitive model allows the researcher to identify seemingly different features of an item that nevertheless affect difficulty in the same way. Thus, even though $C > 14$ and $S + C > 25$ are dissimilar, they serve the same function in the cognitive model; they are generator constraints, and so should not directly affect difficulty. Similarly, the two forms of verifier constraints serve the same role according to the theory, and so should both contribute to the difficulty of an item.

Two previous studies of GE items provide support for the predictions that (a) adding a verifier constraint to an item will increase difficulty, whereas (b) an additional generator constraint will not. Nhouyvanisvong and Katz (1998) investigated GE items designed at the level of the quantitative section of the Graduate Record Examinations (GRE®) General Test, whereas Bennett et al. (1999a) investigated the use of GE items in the context of the mathematics portion of the SAT® I: Reasoning Test. However, neither study set out initially to systematically manipulate the type of constraints. Rather, added constraints were identified by way of post hoc analysis of the items, and so there is the possible alternative explanation that other, confounded differences in the items led to the results. In addition, neither study provided an opportunity to investigate the number of verifier constraints as a continuous variable. In both studies, the key comparisons were between a base item and variants of that item (the original item with an added generator constraint or an added verifier constraint) which leaves open the question of how to alter item difficulty without explicitly adding constraints (e.g., by changing a constraint from a verifier to a generator). Experiment 1 explicitly sets out to systematically manipulate both number

of verifier constraints and solution density (discussed next) by crafting items that contain the desired features.

Solution Density

The solution density of an item is the ratio of correct solutions to all possible answers. The greater the density of correct solutions relative to all possible answers, the easier it should be for an examinee to locate one of those correct solutions. Conversely, if correct solutions are relatively rare (lower density), examinees should have greater difficulty finding them. Thus, examinees would need to verify a greater number of potential solutions, which would increase item difficulty.

For GE items considered in this paper, solution density is defined as the proportion of correct solutions to the set of all "reasonable"[1] responses. The maximum values of these ranges of reasonable responses were defined objectively by the axis intercepts of the most restrictive constraint in an item (e.g., 0.25S + 0.40C = 10). The minimum values were typically stated in the item stem.

Figure 6 depicts the underlying solution space for the stamps problem shown earlier (Figure 1). The range of reasonable values for the number of commemorative stamps (2-25) is shown across the top of the graph, and the range of reasonable values for standard-issue stamps (2-40) is shown down the side. The gray wedge and asterisks show the correct solutions to the problem (e.g., four commemorative stamps and 22 standard-issue stamps). The solution density of this problem is 0.17 (159 correct solutions divided by 936 [39 times 24] total responses).

Jamal wants to buy some standard-issue stamps at $0.25 each and some commemorative stamps at $0.40 each. He wants more than 25 stamps, including at least two of each, but can spend no more than $10.00. What is one possibility for the number of standard-issue stamps and commemorative stamps Jamal could buy?

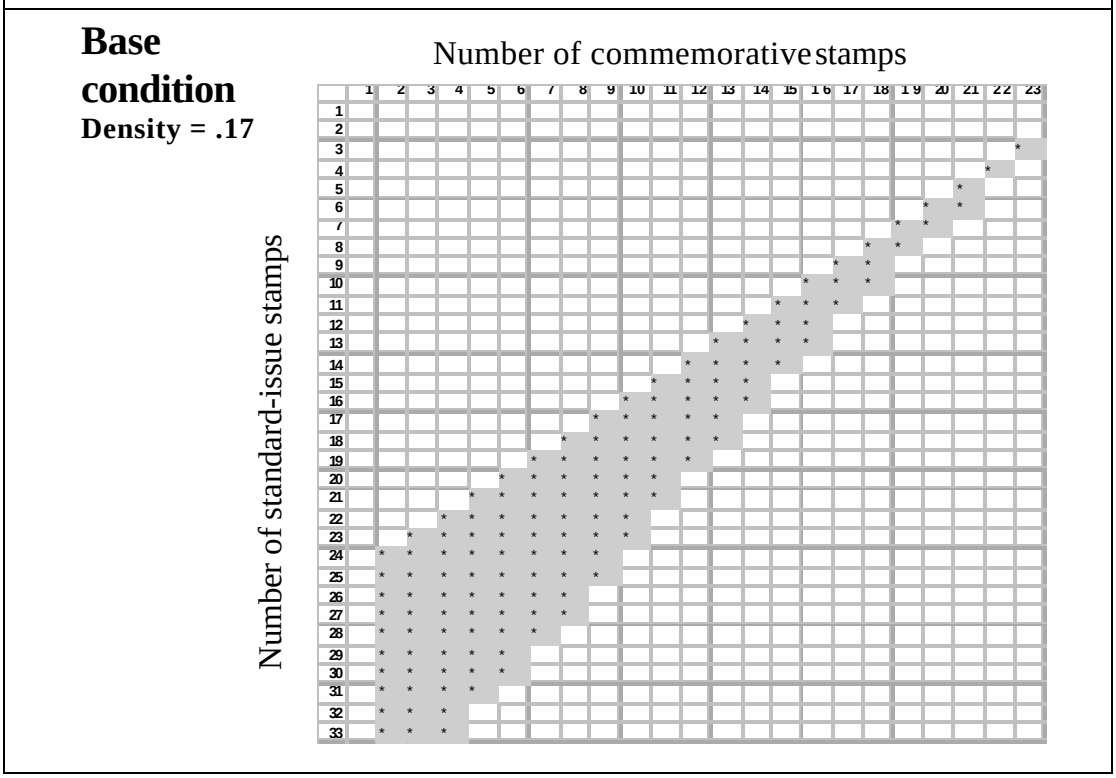
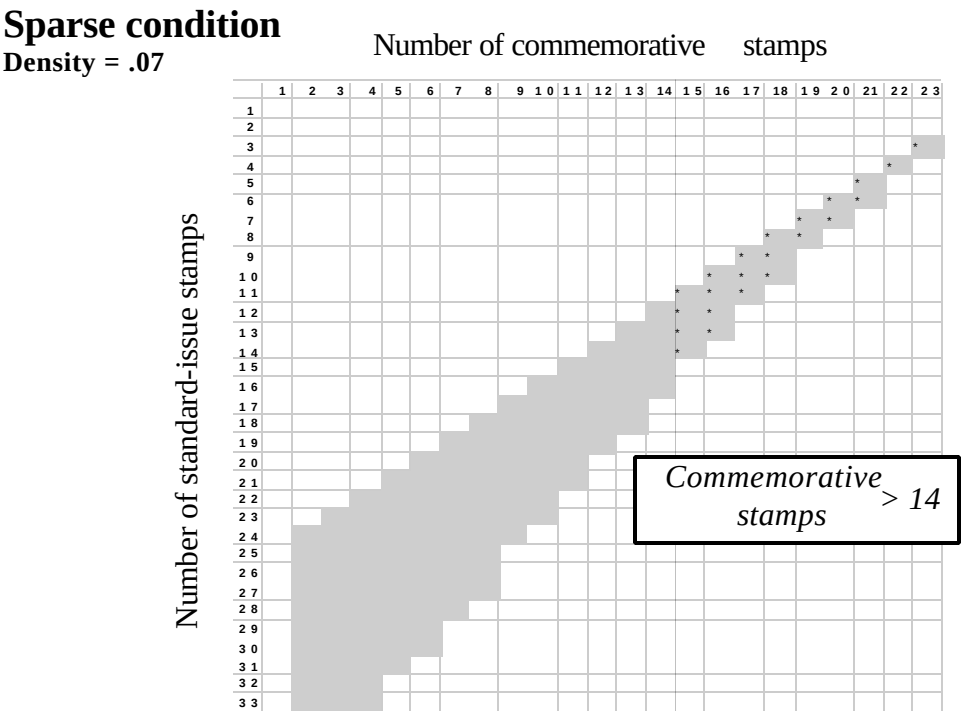


Figure 6. **Solution space for the GE "stamps" item.**

Changing the constraints, or adding another constraint, may change solution density. For example, as Figure 7 shows, adding the constraint that commemorative stamps must be greater than 14 reduces the number of correct solutions to 45 and the number of possible responses (both correct and incorrect) to 650 (25 standard-issue times 26 commemorative), thereby reducing the overall solution density to 0.07. On the other hand, *restricting* commemorative stamps to be *fewer* than 14 increases density to 0.26. Furthermore, by adding a different constraint (e.g., standard-issue stamps must be fewer than 20), solution density remains the same as the base problem.

Jamal wants to buy some standard-issue stamps at $0.25 each and some commemorative stamps at $0.40 each. He wants more than 25 stamps, including at least two of each, but can spend no more than $10.00. **If he wants more than 14 of the commemorative stamps,** what is one possibility for the number of standard-issue stamps and commemorative stamps Jamal could buy?



Note. The gray wedge represents the solution space of the original item; the asterisks represent the new solution space.

Figure 7. **Additional constraint that reduces solution density of the GE "stamps" item.**

In the three experiments described below, we investigated number of verifier constraints and solution density as predictors of GE item difficulty (in terms of both accuracy and speed). In Experiment 1, we systematically manipulated these factors to examine their predictive power. In Experiment 2, we investigated solution density using GE items that are more heterogeneous than those of the more controlled experiment. In Experiment 3, we examined the cognitive strategies underlying the effect of density on performance through an investigation of how the information in an item stem influences examinees' solution estimates. By drawing from both psychometric and cognitive approaches, these experiments provide a rich account of the GE item type and factors affecting performance.

Experiment 1: Modeling GE Item Difficulty

As noted above, Experiment 1 examined the predictive validity of the two theoretically motivated factors: number of verifier constraints and solution density. This experiment also investigated the contribution of item presentation, or "cover" – that is, whether item cover stories are presented in terms of words or equations. This factor is of interest from a practical standpoint because the quantitative portion of the GRE General Test consists of both word items and pure items. In addition, from a theoretical standpoint the presentation form of an item affects neither number of constraints nor solution density, yet there is considerable evidence of differential performance on equivalent items presented in these two forms (e.g., Kintsch, 1998; Nathan & Koedinger, 2000). These three factors were systematically manipulated to create a pool of GE items, which were then administered to a sample of past GRE test takers. In this way, we also examined the relationships between performance on GE items and on the GRE General Test.

Method

Participants

A total of 169 university students – 57 seniors, 101 first-year graduate students, 10 other graduate students, plus one student who did not specify his/her academic level – participated in the study. The sample included 116 female and 53 male students, all of whom had taken the GRE General Test between October 1996 and January 1999. Students were recruited from eight universities in different regions of the U.S.[2] Each student participated at his/her local institution and was paid $30.

The sample of students was reasonably representative of the GRE test-taking population,[3] with a few exceptions. The sample was 69% female, compared with the 58% female GRE test-taking population, and 96% of participants said that they understood English as well as or better than any other language, whereas 86% of the GRE test-taking population made this claim. Table 1 and Table 2 provide further demographic and academic information about the sample; they also show that participants were more academically adept than the GRE population both in terms of self-reported undergraduate GPA and GRE scores.[4]

Table 1

Demographic and Academic Information

		Sample (N = 169)	GRE population (N ~ 1,070,764)
Citizenship	U.S. citizen or resident alien	88%	78%
Self-description	American Indian or Native American	1%	1%
	Black or African American	2%	9%
	Mexican, Mexican American, or Chicano	8%	2%
	Asian, Asian American, or Pacific Islander	11%	7%
	Puerto Rican	1%	1%
	Other Hispanic or Latin American	3%	2%
	White (non-Hispanic)	73%	75%
	Other	2%	3%
Undergraduate GPA	3.5 - 4.0	49%	16%
	3.0 - 3.49	40%	28%
	2.5 - 2.99	10%	35%
	2.0 - 2.49	1%	13%
	1.5 - 1.99		7%
	Below 1.5		< 1%
GRE means (SDs)	GRE quantitative	589 (120)	569 (142)
	GRE verbal	508 (112)	470 (114)
	GRE analytical	593 (125)	545 (131)

Table 2

Other Academic Information

Undergraduate		Graduate (Actual or intended)	
Major		*Major*	
Social sciences	30%	Social sciences	25%
Humanities/arts	19%	Humanities/arts	10%
Life sciences	9%	Life sciences	8%
Education	10%	Education	22%
Physical sciences	4%	Physical sciences	1%
Engineering	6%	Engineering	5%
Business	1%	Business	
Other	21%	Other	29%
Degree sought			
Masters degree	65%		
Doctoral degree	35%		

Note. N for each question ranges from 163-169.

Materials

A total of 12 GE items were created, each with a unique cover story. All of these items involved two variables and two inequalities; because the two mathematical relationships were inequalities, the items were underdetermined. Of these 12 items, eight consisted of two verifier constraints (e.g., Figure 5) and four consisted of one verifier and one generator constraint (e.g., Figure 1). Solution density varied considerably among the base items, ranging from .027 to .302 (M = .166; SD = .079).

From each of these 12 base items, we created five variants to manipulate solution density and number of verifier constraints. Each variant consisted of the base item plus an additional constraint. For each type of added constraint (generator or verifier), density was either the same as the base item, lower than the base item, or (for additional generator constraints only[5]), higher than the base item. Table 3 describes the structure of each variant.

Table 3

Structure of Item Variants

	Type of added constraint	Density relative to base
1.	Generator (e.g., X > 20)	Same
2.	Generator	Lower
3.	Generator	Higher
4.	Verifier (e.g., X > 5Y)	Same
5.	Verifier	Lower

The pool of GE word problems consisted of 72 items (12 base items plus five variants for each base item). The number of verifier constraints per item was 1, 2, or 3 (for a total of 16 items, 40 items, and 16 items, respectively), and solution density of the items ranged from .001 to .506 (M = .145; SD = .110). A pure (equations only) version of each word problem was also created, as outlined below, for a total of 144 items.

To implement the manipulation of item cover (word form versus pure), we formulated a pure mathematics version of each of the 72 GE word problems. These presented only the underlying inequalities and restrictions on the values that could be taken by the variables (usually that the variables

were positive integers). The figures below are screen snapshots taken from the computerized test delivery system. Figure 8 presents an example GE word problem, and Figure 9 shows a pure problem. Both items were derived from the same base item, but contain different additional constraints.

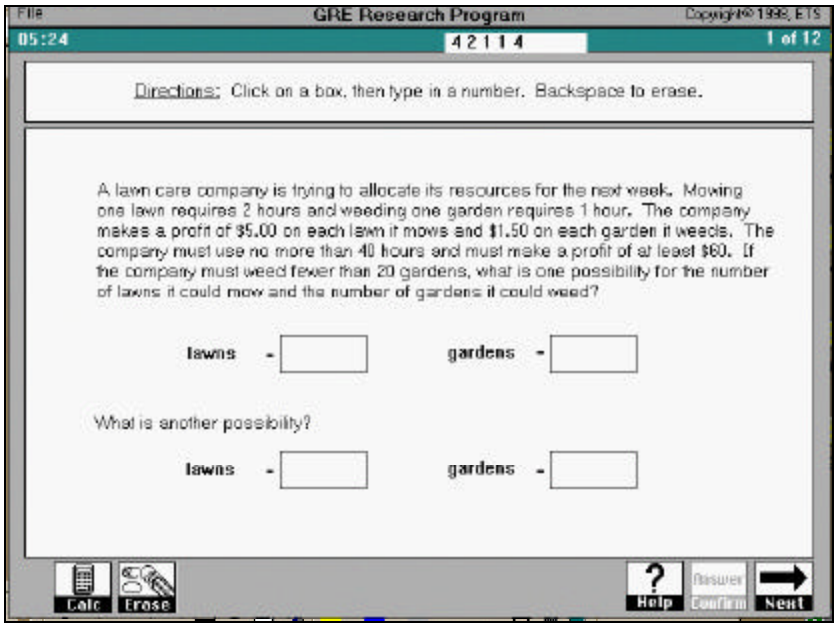


Figure 8. **Sample GE word problem (contains an added generator constraint).**

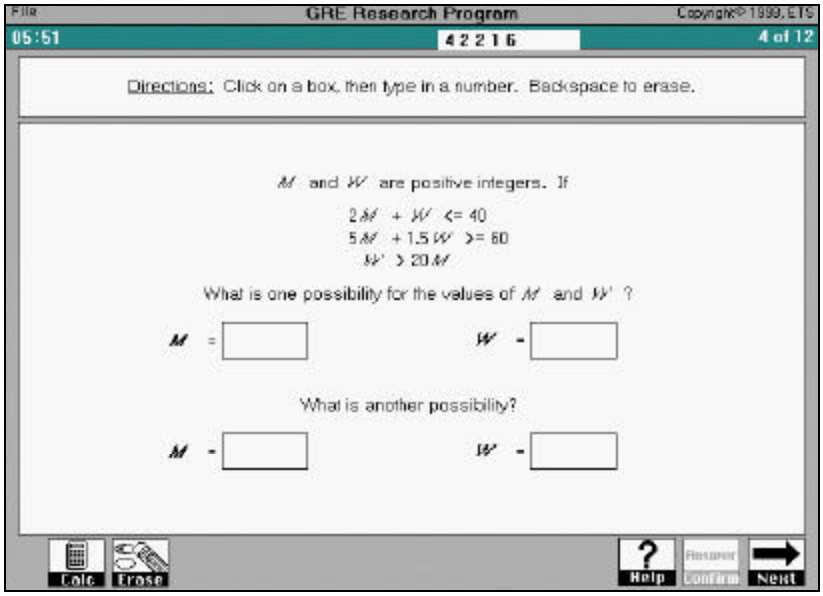


Figure 9. Sample GE pure problem (contains an added verifier constraint).

Design

Each student received all 12 items, in a random order, half presented as word problems and half as pure problems. The six word problems consisted of one each of the six item versions (base plus the five variants), as did the six pure problems. The assignment of which version of each item a student received was random, although variants derived from base items that contained one and two verifiers were represented equally between word and pure items. Test forms were created such that all 144 items were distributed approximately equally across the entire set of students. Thus, each student received a different set of items from any other student, subject to the constraints mentioned above.

Procedure

Students participated individually in sessions that lasted approximately one hour. Each student first viewed an on-line tutorial that explained how to use the test delivery software. Following the tutorial, students solved the 12 experimental items. After completing the test, students responded to a brief demographic and academic questionnaire.

Items were presented one-at-a-time, and students were not allowed to revisit items. They were given six minutes to complete each item, and were allowed to continue to the next item if they finished within the time limit, but when the item had been displayed for six minutes, the software displayed an alert and automatically proceeded to the next item. Students were asked to provide two solutions to each item, and were invited to solve the items however they wished. In addition to an on-screen calculator, they were provided with scratch paper and a pencil. The test delivery software recorded and time-stamped many of the students' interactions with the software, such as clicking in a text box and entering numbers, clicking on calculator keys, and moving to the next item (which involved clicking "next" then "confirm").

Response Scoring

Student responses were scored automatically based on scoring rubrics developed from the pure versions of each item. A sample of the responses was also independently scored by hand to verify the accuracy of the automatic scoring. Because students were asked to generate two answers to each item, scores on each item ranged from 0-2. The measure of seconds-to-solution was defined as the time that

passed between the moment an item appeared on screen and the moment the student clicked on the "Next Item" button after providing both responses to the item. If the student ran out of time, s/he was assigned the maximum seconds-to-solution time (360 seconds) for that item. Of the pool of 144 items (12 items times 6 versions times 2 cover options[word/pure]), four items inadvertently had fewer than two correct solutions. Because this restriction would impact the measure of accuracy, response data from these items were excluded from all analyses.

Results

Our discussion of the results of Experiment 1 is organized into two sections. First, we investigate the relationship between performance on the experimental GE items and performance on the GRE General Test. Thus, the units of this analysis are students. In the second section, we examine the potential for the three item-difficulty factors (and their interactions) to predict item scores and mean seconds-to-solution. Thus, the units of this analysis are items.

Student Performance

Students' mean score was 14.4 (out of 24); on average, students generated 1.2 correct responses to each item. Students completed all 12 items on average in 45 minutes, which represents a mean completion time of 3.7 minutes per item. Bridgeman and Cline (1999) report mean solution times of two minutes for algebra problem-solving items (both pure and word) on the computerized GRE exam. The longer time for the experimental GE items was expected because each item required the construction of two responses, whereas the typical GRE quantitative item requires selection of one response from a set of options.

Table 4 displays intercorrelations among GRE scores, total GE test score, and total seconds-to-solution, as well as a reliability measure for the GE test (coefficient alpha). As expected, GRE quantitative score correlates significantly with performance on the GE items used in this experiment. The overlap is not complete, however, suggesting that GE items might tap skills not assessed by the GRE quantitative test. By itself, GRE quantitative score accounts for approximately 38% of the variance in GE test score, but the reliability of the GE test suggests that about 36% of the variance remains to be explained.

Regarding solution time, no relation exists between the time it took to complete the test and any of the other measures, which is surprising. Intuitively, one expects that students with greater math proficiency would complete the test more quickly than less math-proficient students. This lack of correlation might be explained by the fact that, when a student exceeded the time limit on an item, s/he was recorded as spending 360 seconds on that item. Indeed, there is the expected negative correlation between the number of times students exceeded the time limit and their GRE quantitative scores (r = -.25, p < .01). On average, each student exceeded the time limit on .85 of the 12 items (SD = 1.4); approximately 43% of the students exceeded the time limit on at least one item.

Table 4

Intercorrelations of GE Measures and GRE Scores

	α	GE score	Seconds-to-solution	GRE quantitative score	GRE verbal score	GRE analytical score
GE score	.74	—	.05	.62**	.34**	.56**
Seconds-to-solution	.86		—	-.07	-.08	.00
GRE quantitative score				—	.46**	.70**
GRE verbal score					—	.53**
GRE analytical score						—

** p < .01

Difficulty Modeling

As noted earlier, in this section we investigate the predictive validity of the three difficulty factors (number of verifier constraints, solution density, and item cover) with respect to mean score and mean seconds-to-solution (over all students) for each of the 140 items. A hierarchical multiple regression analysis was used to ascertain the relative contributions to prediction of each factor, as well their interactions. Based on a random half of the data, we then constructed multiple regression models – one for each dependent variable (DV) – consisting of just the significant contributors. These models were then used to investigate the effects of each contributor and to test the generality of the models by applying them to the other half of the data.

Table 5 displays the intercorrelations between the DVs and the three difficulty factors, plus a reliability measure (coefficient alpha)[6] for each of the DVs. Number of verifier constraints and difficulty are moderately correlated, suggesting shared variance accounted for by each factor. However, preliminary analyses suggested that changing the order of the two factors in the hierarchical regression analysis yields no substantive difference in the results.

Table 5

Intercorrelations of GE Measures and Three Difficulty Factors

	α	GE score	Seconds-to-solution	Number of verifiers	Solution density	Item cover
GE score	.88	—	-.684**	-.584**	.569**	.140
Seconds-to-solution	.90		—	.603**	-.418**	-.403**
Number of verifiers				—	-.452**	.000
Solution density					—	.000
Item cover						—

** $p < .01$

Table 6 and Table 7 present hierarchical multiple regression analyses for score and seconds-to-solution, respectively. At each step in the analysis, an additional predictor was added: the main factors, the two-way interactions, and the three-way interaction. The results are similar for the two DVs: The complete models account for over 55% of the variance, and the same predictors contribute significantly to prediction by each model. Considering the reliability measures for each DV, there is still about 35% of the variance to be explained by heretofore unidentified factors.

Table 6

Hierarchical Multiple Regression Analysis on Score

Step	Predictor added	Cumulative R^2	Change R^2	*F*	*df*
1	Number of verifiers	.341	.341	71.5**	1, 138
2	Solution density	.458	.117	29.5**	1, 137
3	Item cover	.478	.020	5.1*	1, 136
4	Verifier-by-density	.543	.065	19.3**	1, 135
5	Verifier-by-cover	.552	.009	2.7	1, 134
6	Density-by-cover	.553	.001	.34	1, 133
7	Verifier-by-density-by-cover	.555	.001	.44	1, 132

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

Table 7

Hierarchical Multiple Regression Analysis on Seconds-to-Solution

Step	Predictor added	Cumulative R^2	Change R^2	*F*	*df*
1	Number of verifiers	.363	.363	78.7**	1, 138
2	Solution density	.390	.027	6.0*	1, 137
3	Item cover	.552	.162	49.2**	1, 136
4	Verifier-by-density	.568	.017	5.2*	1, 135
5	Verifier-by-cover	.569	.000	.05	1, 134
6	Density-by-cover	.580	.011	3.4	1, 133
7	Verifier-by-density-by-cover	.581	.002	.56	1, 132

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

Table 8 shows the regression coefficients for models fitted to a randomly selected half of the data (a randomly selected half of the original items, plus their variants, for a total of 68 items), as well as; the means and SDs for the predictors. For the DVs in this subsample, the mean score was 1.2 (*SD* = .43) and the mean seconds-to-solution was 223 (*SD* = 48).

These models are reasonably stable. Using the coefficients from Table 8 to predict GE scores and seconds-to-solution for the withheld item data (the remaining 72 items) yields correlations of .64

and .67, respectively. The fact that these multiple correlations are somewhat lower than those obtained for the other sample (.74 and .76, respectively) should not be surprising. The higher correlations occurred when fitting the multiple regression model to the data; these lower correlations reflect using the previously derived model to predict new data.

Table 8

Means (SDs) and Regression Coefficients for Models Based on Random Half of Items

		GE score[a]			Seconds-to-solution[b]		
Predictor	Mean (SD)	B	SE B	β	B	SE B	β
Number of verifiers	2.0 (.69)	-.57	.08	-.92**	58.9	8.6	.85**
Solution density	.16 (.12)	-1.66	.70	-.48*	117.6	78.5	.31
Item cover	.50 (.54)	.13	.06	.16*	-34.4	6.7	-.36**
Verifier-by-density	.28 (.21)	1.66	.37	.80**	-108.6	40.8	-.47*

[a] Constant = 2.12; R^2 = .68; [b] Constant = 133.7; R^2 = .68; * $p < .05$; ** $p < .01$

The regression equations can be used to describe the effect of each factor in terms of expected performance by a sample of examinees. As predicted, more verifier constraints led to more difficult and time-consuming items. Substituting the number of verifier constraints and the mean of the other predictors into the regression equations yields predicted scores of 1.6, 1.3, and 1.0 and predicted seconds-to-solution of 178, 218, and 260 for items with one, two, and three verifier constraints, respectively.

Also as predicted, solution density was negatively related to item difficulty: Items with higher densities were easier than those with lower densities. However, this effect was moderated by the number of verifier constraints. Figure 10 and Figure 11 show the effect of density on score and seconds-to-solution, respectively, with separate regression lines indicating the number of verifier constraints. The figures show that greater number of verifier constraints led to a greater influence of density on difficulty and solution time. A possible explanation for this effect is considered next in the discussion section for this experiment. Finally, items presented as words (score: 1.2; seconds-to-solution: 240) were more difficult and took more time to complete than items presented as equations (score: 1.3; seconds-to-solution: 205).

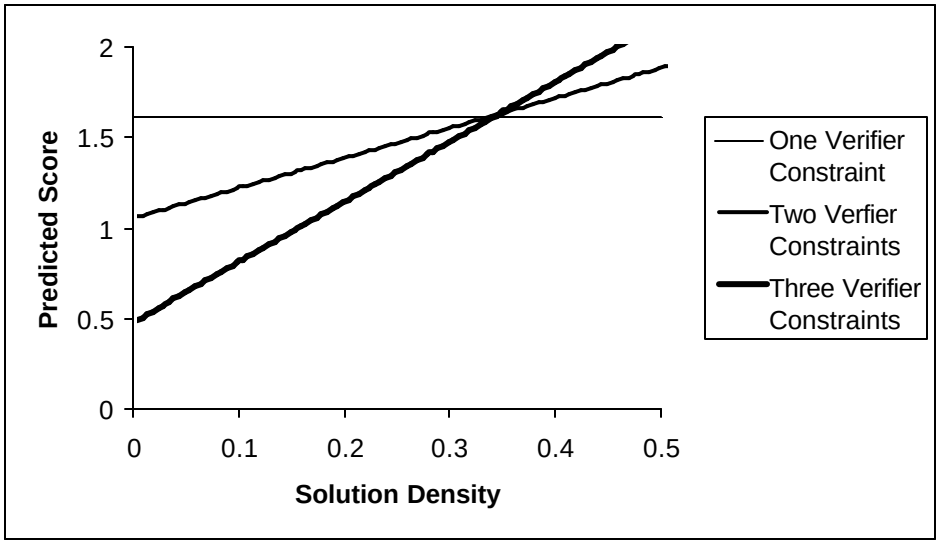


Figure 10. **Interaction of density and number of verifier constraints in predicting scores.**

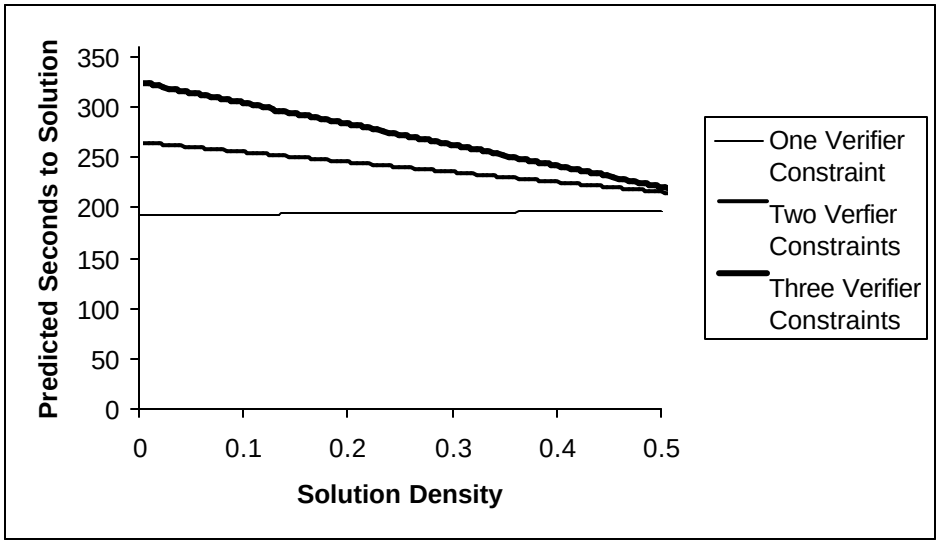


Figure 11. **Interaction of density and number of verifiers in predicting seconds-to-solution.**

Discussion

Experiment 1 demonstrated the success of number of verifier constraints and solution density in predicting the difficulty of GE items. As predicted by the cognitive model, more verifier constraints lead to greater difficulty, as do lower solution densities.

Whether the item was presented as words or as equations affected both speed and accuracy, with pure problems being easier and less time consuming than word problems. One might interpret this result in light of a two-step solution approach for word problems: (a) translate the item stem into equations, and then (b) use the resulting equations in a generate-and-test strategy. However, it was unclear whether the added burden of word problems was limited to the time needed to extract equations, or whether examinees continued to rely on the verbally presented constraints when solving the items. Future work might investigate the solution strategies adopted by examinees when solving GE items presented in words or equations. For example, direct presentation of equations might encourage a strategy involving more algebra (to reduce the equations and uncover implicit constraints); examinees solving a word problem might instead deal with the explicitly provided (albeit verbal) constraints.

What is the mechanism whereby density affects difficulty? One possible explanation is that density affects the likelihood of estimating a solution correctly. In the items used in the present experiment, more verifier constraints meant fewer generator constraints, and less information to use when making an estimate. Without any guidance from generator constraints, the probability of correctly guessing a solution should equal the density for an item. More generator constraints would limit the space of considered solutions and increase the likelihood that any estimates would be correct, thus making the absolute density of the item less of a factor in performance.

The predictive success of solution density is especially noteworthy, because a previous analysis of GE items (Bennett et al., 1999a) failed to find a relationship between density and performance. However, there were several critical differences between that work and the present experiment.

First, Bennett et al. (1999a) conducted their analysis of density post hoc: The items were not specifically crafted to manipulate density, and in fact, density and constraint type (generator or verifier) were not independently manipulated. Thus, the relatively large effect of verifier constraints might have overwhelmed the more subtle effect of density. The results of Experiment 1 support this interpretation

because solution density was a more powerful predictor when one included its interaction with number of verifier constraints.

Second, Bennett et al. (1999a) utilized a relatively heterogeneous pool of items – the items differed from each other in form, content, type of responses, and so on – in contrast to the homogeneous items used in the present experiment for purposes of experimental control. The effect of density on item difficulty might be small relative to the effect of other item factors; item density might prove to be useful only in the unlikely case when the item pool consists of similar GE items. This issue is important because it is likely that any use of GE items on operational tests will involve a wide variety of GE items. To further test this possibility, we conducted a pilot investigation of solution density utilizing a more heterogeneous set of items.

Experiment 2: Further Investigation of Solution Density

Method

Participants

Thirty-nine undergraduates (26 female; 13 male) from the experiment participation pool at George Mason University in Fairfax, Virginia, participated in Experiment 2 for course credit.

Materials and Design

Table 9 displays the four underdetermined algebra items that were created for Experiment 2, each of which has a unique cover story. Two variants of each of these four base items were created by adding a generator constraint to the original problems. For two base items (experimental items), the variants were crafted such that the additional constraint produced a higher density in one case and a lower density in another; the resulting item sets had average densities of .21, .08, and .34 for the base items, lower density variants, and higher density variants, respectively. Variants for the remaining two base items (control items) were crafted such that the additional constraint did not affect density; the resulting item sets for these items had average densities of .15, .13, and .16 for the corresponding base items and variants.

Table 9

GE Items Created for Experiment 2

Experimental items	Control items
Jamal bought *A* stamps at $0.25 each and *B* stamps at $0.40 each. Altogether, he bought more than 25 stamps, including at least two of each, and spent no more than $10.00. List two possible values for *A* and the corresponding values for *B*.	A and B are positive integers. Find two sets of values for A and B such that the following inequalities are true: $.75A + .67B > 5$ $.75A + .67B < 7$
Company *W* produces two skirts, one long and one short. The short skirt requires $2 worth of material and 4 hours of labor to manufacture. The long skirt requires $4 worth of material and 4 hours of labor to manufacture. The company wishes to produce more than 150 skirts, using less than $500 worth of material and 750 hours of labor. List two possible numbers of short skirts the company could produce and the corresponding numbers of long skirts.	A company sells both hardback and paperback books every month. It makes a profit of $3.30 on every hardback book it sells and a profit of $1.20 on every paperback book it sells. If it made a profit of $3,960 on hardback and paperback books last month, what are two possible combinations of hardback and paperback books it could have sold last month?

These items are more heterogeneous than those used in Experiment 1. For example, one item is similar to a pure problem, whereas the others are word problems. One item (skirts) involves three inequality constraints. Another item (books) involves a single linear constraint, whereas the other two items pose two inequality constraints. One of the items involves only whole numbers while the others involve decimals.

Each participant solved one item from all four item sets, plus two additional GE items that were not part of the current design (there were two additional variants for each of these as well), for a total of six items. Of these six items, each participant solved two each of base items, lower density variants, and higher density variants. As a result of this design, each variant developed from the four critical items was solved by 12 to 14 students.

Procedure

Participants were tested in groups of two to 15. Testing sessions lasted approximately one hour. Participants were supplied with a pen, calculator, and test booklet, the latter consisting of one item per

page. Below each item was an area for each set of answers requested (either one or two boxes) as well as any associated work.

Participants were given four minutes to solve each item. Participants who finished before this time limit were asked to wait until the four minutes were up; participants who were not finished when time was called were asked to proceed to the next item. After completing all items in the test booklets, participants filled out a background questionnaire and a questionnaire asking their opinion of the items.

Response Scoring and Analyses

Responses were scored by hand. Because of the extreme difficulty of these items for this sample of students (mean proportion correct for all items was 0.33), analyses used the proportion of correct first responses (recall that each item requested two responses) as a difficulty metric. A failure to respond at all was considered incorrect. Analyses utilizing other measures of difficulty (e.g., correctness of both first and second responses) yielded parallel results.

Results

For items that did not differ in density, the presence of an additional constraint slightly increased difficulty. The base item was answered correctly (nonsignificantly) more often than either of the variants. However, as predicted, the specific constraint had no effect on difficulty (see Figure 12). Table 10 provides a comparison of the lower and higher density variants of each item.

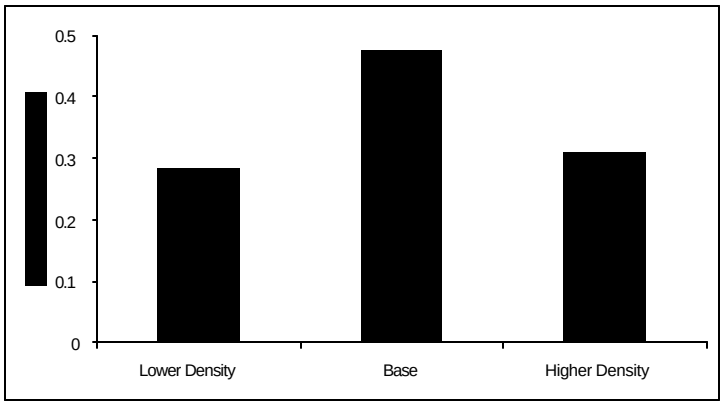


Figure 12. **Control items with minimal differences in density.**

Table 10

Comparison of Lower Density and Higher Density Variants

	Lower density		Higher density			
Item	No. correct	n	No. correct	n	*z*	*p*
	Control items					
A & B	5	12	6	13	.23	.41
Books	2	13	2	13	0	.50
	Experimental items					
Stamps	1	13	6	14	2.08	.02
Skirts	2	13	6	13	1.70	.05

Note. The *p*-values are reported for one-tailed tests, which are justified because of the expectation that lower density variants are more difficult than higher density variants.

For items that differed in density, we found the expected relation between density and difficulty. As Figure 13 shows, higher density variants tended to be easiest for participants, the base items were slightly more difficult, and lower density variants were most difficult. Post hoc comparisons of the proportion of students who correctly responded to lower and higher density variants (which have the same number of constraints) revealed significant differences for each item, as shown in Table 10.

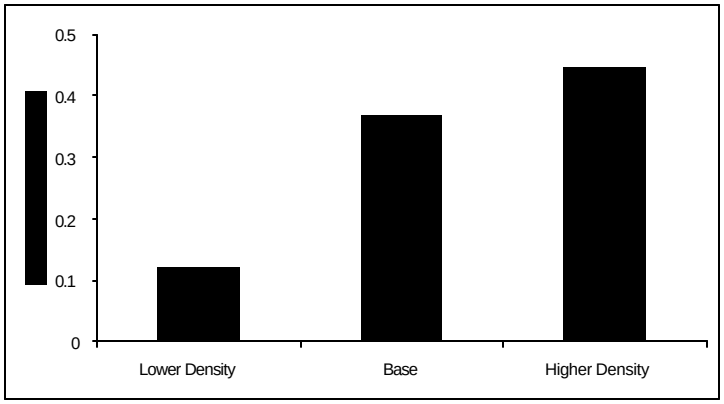


Figure 13. **Density manipulation affected item difficulty.**

Discussion

The results of this pilot experiment provide further support for the predicted positive relationship between the density of an item and its difficulty. The results also provide evidence for the generality of the relationship over different items: Compared with Experiment 1, the items in this experiment were more heterogeneous. Of course, GE items can be crafted that differ substantially from one another (see, e.g., Bennett et al., 1999a). Future experiments might investigate the relationship between density and difficulty through similarly systematic manipulations of item density, but utilizing items that differ even more widely in form and content than do the items used in Experiment 2.

Experiment 3: Investigations of Examinee Estimation Skill

Experiment 3 was intended to inform a model of student performance on GE items. Whereas Experiments 1 and 2 suggested several item factors that influence the difficulty of an item, Experiment 3 looked at possible mechanisms whereby item characteristics affect difficulty. Thus, this experiment represents a shift in emphasis from difficulty modeling to cognitive process modeling.

Because Experiment 3 investigated problem solving processes, we focused on a particular solution strategy: generate-and-test. In this experiment, students were instructed in the generate-and-test strategy, encouraged to use that strategy when solving items, and given an answer sheet that was structured around the generate-and-test strategy – that is, it had spaces for indicating estimates (solutions) and for checking those estimates against the inequalities in the items.

The dependent measure used in this experiment was students' initial estimate of the values of one (or both) of the variables in an item. This estimate represents the first step in a chain of events under the generate-and-test strategy (see Figure 2), which eventually leads to the student reporting a solution. In the current experiment, we manipulated item characteristics in two ways to observe their effect on students' first choice of solution estimates.

First, we manipulated the underlying structure of one set of items, systematically varying parts of the stem to observe effects on examinees' estimates. To distinguish the inequalities in the item, each item was crafted to include an easily interpreted "simple" inequality (a type of generator constraint, which will be called a "simple" equation: e.g., $x + y = 80$) and a more "complex" inequality (with non-unit

coefficients: e.g., 8x + 13y = 108). Analyses focused on the extent to which participants used either of the inequalities (or both) in generating their initial estimates.

Second, using a second set of items, we investigated how participants generated initial estimates in the absence of an easily interpreted inequality (i.e., these items contained two complex inequalities). One possibility we predicted was that examinees' estimates might become "anchored" to particular numbers given in an item. Through this second set of items, we introduced item variants that differed only in the numbers used; variants have the same underlying solution space.

Method

Participants

Sixty-four undergraduates from the experiment participation pool at George Mason University participated for course credit. Their self-reported SAT mathematics scores ranged from 320-670, with a mean of 512.

Materials and Design

Each participant received 14 GE word items in random order. All items consisted of two inequalities each involving two variables. Eight items contained one simple inequality (unitary coefficients) and one complex inequality (nonunitary coefficients); six items contained two complex equalities. For explanatory purposes, we refer to the eight items as the "information study" (which investigated students' use of simple vs. complex inequalities when generating an estimate) and to the six items as the "anchoring study" (which investigated whether mathematically irrelevant changes to items affect students' estimates). Note that students were asked to generate only a single set of estimates for each item, rather than the two sets of estimates requested for GE items in Experiment 1 and Experiment 2.

Stimuli and design: Information study. The information study followed a two-by-two within-subject factorial design. The two manipulated factors were (a) whether or not the constant in the simple equation was changed and (b) whether or not the constant in the complex equation was changed. Whenever the constant in an item was altered to create a variant, it was either increased or decreased

by approximately 30% such that the number of solutions in the variant was greater than the number of solutions in the base item. Table 11 shows how the manipulation was implemented for one GE item.

Table 11

Example GE Item From Experiment 3: Information Study

Jamal wants to go on some rides at the fair. He would like to go on at least 27 rides, but has only 80 tokens to spend. Regular rides cost 2 tokens and thrill rides cost 5 tokens. What is one possibility for the numbers of regular and thrill rides Jamal could go on?
Base item R + T = 27 2R + 5T = 80
Simple constant changed R + T = **19** 2R + 5T = 80
Complex constant changed R + T = 27 2R + 5T = **104**
Both constants changed R + T = **19** 2R + 5T = **104**

Note. The item stem shown is the base version. Numbers in bold represent the changes made to the items for each version. These numbers were changed in the *text* of the item; the equations were not shown to participants.

Each participant received two items in each of the four conditions. Two other factors were counterbalanced across items: (a) the order in which simple and complex inequalities were presented in prompts and (b) the direction of the inequalities (i.e., for half of the items, "=" was used for the simple inequality and "=" was used for the complex inequality; the assignment was reversed for the remaining items).

Stimuli and design: Anchoring study. For the anchoring study, stimuli consisted of six pairs of GE word problems, each consisting of two nonunitary inequalities. Each pair of items differed only in the particular numbers used in the prompt: One item in each pair used numbers that were approximately 150% larger than the other item in the pair. Beyond the change in numbers, all other characteristics (cover story, underlying inequalities, space of possible solutions, and so on) remained the same within each item pair. Table 12 shows an example item pair. Each participant received one item from each pair, for a total of six items. Three of the items consisted of smaller numbers; three consisted of larger numbers.

Table 12

Sample GE Items From Experiment 3: Anchoring Study

Smaller numbers	Larger numbers
An organization is arranging transportation for 960 people, by bus or by van. Each bus can transport 32 people and costs $160; each van can transport 8 people and costs $48. The organization can pay no more than $5,280 for transportation. What is one possibility for the number of vans and buses the organization could use?	An organization is arranging transportation for 1,440 people, by bus or by van. Each bus can transport 48 people and costs $240; each van can transport 12 people and costs $72. The organization can pay no more than $7,920 for transportation. What is one possibility for the number of vans and buses the organization could use?

Procedure

All participants were tested in groups of three to 20. At the start of each session, the experimenter gave a brief description of underdetermined (GE) problems, then demonstrated the generate-and-test strategy, using an example to solve a standard algebra word problem. Participants were shown how to estimate the value of one or both of the variables and then test those estimates against the equations (or inequalities) represented in the problem text.

Participants were given one answer sheet for each item. They were instructed to write down all of the work they used to solve each problem, in order, from the top of the page to the bottom of the page, including calculations performed on a provided calculator. All estimates generated were to be labeled to identify the variable being estimated. All participants then solved a practice item using the generate-and-test strategy. The experimenter verified that the answer sheets were being completed correctly and that participants understood how to implement the generate-and-test strategy.

Participants were allowed three minutes to work on each item. Participants who finished before the three minutes elapsed were asked to wait until the experimenter asked all participants to proceed to the next item. Participants who were not finished with an item when time was called were asked to proceed to the next item.

Response Scoring and Analyses

For each participant's response to an item, we recorded the first value generated for each variable (e.g., number of regular rides and number of thrill rides), as indicated on the answer sheet. This first response is referred to as the participant's "estimate" for each variable.

Results

Information Study

Overall, of the 512 answer sheets completed (eight items times 64 participants), participants' first estimates were correct 62% of the time. These first estimates are clearly not random guesses, but guesses that reflect some initial processing on the part of the student. However, this processing must have occurred mentally, rather than on paper: because these estimates are the first numbers written down by participants – that is, the first numbers that they chose to test against the constraints of an item.

Eighty-five percent of participants' responses contained first estimates for both variables. That is, in many cases, participants provided values for each of the two variables in the problem *before* making any calculations on the answer sheet. The remainder of responses either contained no estimates or contained an estimate of only one variable. In these latter cases, evidence suggests that the single estimate was tested before the second variable was estimated.

There was also evidence that participants primarily used the simple inequality to generate their initial estimates of values for the two variables in each item. Of the two-estimate responses, 55% summed exactly to the constant contained in the simple inequality. For example, for the item in which the simple inequality was R + T = 27, the sum of the estimates provided by participants was 27.

To analyze this behavior, we calculated the mean sum of initial estimates for each version of each item (eight items times four versions yielded 32 numbers in this analysis), which are shown in Table 13. We ran a repeated-measures MANOVA, using items as the unit of analysis, with two within-item factors: (a) whether the simple constant was increased and (b) whether the complex constant was increased. (This analysis required recategorization of the items, because for half the items, the "change of constant" manipulation involved decreasing the constant, whether of the simple or complex equation). As Table 14 indicates, there was a significant main effect of increasing the simple constant, as expected.

Table 13

Mean Sum of Initial Estimates

		Simple constant increased?	
		No	Yes
Complex constant increased?	No	39.3	46.2
	Yes	36.9	44.5

Table 14

Effect of Changing the Constant on First Estimates

Source	*df*	*F*	η^2	*p*
Simple constant	1	22.8**	.77	.002
Complex constant	1	.98	.12	.36
Simple-by-complex	1	.16	.02	.70
Simple-by-complex-by-items (within-group error)	7	(6.2)		

Note. Values enclosed in parentheses represent mean square errors.
** $p < .01$

How did students apportion the simple constant across the two variables? Table 15 shows the types of strategies participants employed when a response included estimates of both variables. In almost half of the cases, students' initial estimates contained at least one value that was a multiple of 10. Thus, a possible model of performance is that students choose a multiple of 10 that is close to the simple constant for one value, then mentally subtract that amount from the simple constant to arrive at an initial estimate for the second variable.

Table 15

How Constant Is Divided

Response types	Frequency[a]	Percent
All & nothing	5	2%
Multiple of 5	27	11%
Halves	42	17%
Idiosyncratic	64	26%
Multiple of 10	104	43%

[a] The total frequency is fewer than the total number of responses because only a subset of the responses included estimates of both variables.

Anchoring Study

Table 16 lists the average initial estimate of one of the variables[7] for each pair of items used in the anchoring study. Students were not influenced by the mathematically irrelevant change – multiplication by a factor – that was introduced to items: Initial estimates for each item in a pair did not differ significantly. In fact, for the item pair with the largest difference in estimates (item pair 3), mean estimates are opposite what was expected if students had been basing their estimates purely on the magnitude of the numbers in a problem.

Table 16

Mean (SD) Initial Estimates for Each Item Pair

Item pair	Smaller numbers	Larger numbers	*t*	*df*	d
1	61 (41)	72 (58)	0.68	39	.22
2	21 (14)	16 (12)	-1.4	52	-.38
3	107 (153)	46 (31)	-1.7	40	-.55
4	29 (21)	31 (20)	0.34	47	.10
5	64 (27)	82 (91)	0.89	45	.27
6	16 (7)	13 (8)	-1.4	48	-.40

Note. N per cell ranges from 18 to 28.

All Items

Items in the information and anchoring studies differed in terms of the complexity of the inequalities represented in the item prompts. Eight items containing both a simple and complex inequality and six items containing two complex inequalities were administered to each participant. The analyses presented above suggest the strong influence of a simple inequality on students' initial estimates, and one might expect that without an easily interpreted inequality, making an accurate estimate would be more difficult.

As predicted, items containing two complex inequalities were more difficult than items with one complex and one simple inequality. With two complex inequalities, initial estimates were accurate only 26% of the time, compared with the previously reported accuracy of 62% for items containing one

simple and one complex inequality. Although a similar percentage of responses indicated that participants generated two initial estimates (60% in the anchoring study vs. 55% in the information study) to find a solution, in the anchoring study only a small percentage of the first estimates produced a weighted sum equal to the constant in either equation (14% and 21%). Thus, as was found earlier, students rarely produced estimates that were consistent with the complex inequalities. Furthermore, these estimates were not even consistent with a simplified version of the inequalities (e.g., changing "=" into "="), as was the case for the simple inequalities.

How did students arrive at their initial estimates? Without the simple equation, students may have relied on other heuristics, such as using a multiple of 10. To investigate this possibility, we calculated two scores for each student: the proportion of first estimates that were multiples of 10 on (a) items having a simple equation and (b) items having no simple equation. As predicted, students' initial estimates were multiples of 10 significantly more often on items with two complex equations (M = .77, SD = .26) as compared with items having one simple and one complex equation [M = .56, SD = .21, paired t(63) = 6.0, p < .00001, Cohen's d = .89]. Without an easily interpreted inequality, students are more likely to choose an initial estimate that leads to simpler calculations.

Discussion

GE items are cognitively complex. They initially present as many as several hundred possible combinations of values to be checked against two inequalities. With all of these potential calculations, and the concomitant potential for slips, a brute-force generate-and-test strategy is daunting. As with most complex problems, students instead employ heuristics to simplify the problem, thus making them more cognitively manageable.

In Experiment 3, we observed several of the heuristics students use to simplify GE items, thereby reducing the space of possible solutions that must be searched and easing the verification of solutions. Students tended to make initial estimates using multiples of 10 that led to relatively simple calculations. Use of this heuristic was especially prominent when the item offered two complex inequalities not easily verified without using a calculator. Even when the item contained an easily verified inequality consisting of a sum of two variables, students further simplified the inequality by converting it

into a simple equation (e.g., by changing “R + T = 27” to “R + T = 27”). These heuristics simplify GE items and thus make both “generating” and “testing” easier.

General Discussion and Conclusions

The three experiments reported in this paper significantly extend our understanding about the factors that affect examinee performance on GE items, and about the heuristics examinees use in applying the generate-and-test strategy to solve GE items. A model comprising just two theoretically motivated predictors – number of verifier constraints and solution density – accounted for approximately 55% of the variance in scores for a random subset of items. These two factors are extensible to a wide range of GE items. Indeed, in the current work, we used different types of verifier constraints, which despite their surface dissimilarity, affected difficulty in ways predicted by the cognitive model. Bennett et al. (1999a) reported similar results using verifier constraints and GE items that differed even more widely than did the items analyzed in this report. We have confidence that solution density will similarly generalize to a wide range of GE items.

Density has two components: (a) the number of correct solutions and (b) the number of “reasonable” responses. For the purposes of this study, we defined the reasonable range of each variable in an item, which determines the number of reasonable responses, mathematically: We set the other variable to zero and considered the resulting range based on the most restrictive constraint of the item. Although this method results in an objective definition of “reasonable response,” participants’ judgment of range might be affected factors other than the strict mathematics of the item. Participants might be choosing estimates from a wider or narrower range, depending on what information in the item they use to construct their estimates. Future work might investigate alternative definitions of the space of possible responses to observe which resulting definitions of density best predict performance.

The concepts of solution density, verifier constraint, and generator constraint generalize across many different types of GE items. Indeed, the specific form of generator and verifier constraints varied widely in previously published work (Bennett et al., 1999a). For density to have a formal definition, two item characteristics are necessary: (a) a finite number of correct solutions and (b) a finite range of potential, reasonable responses. As noted earlier, some judgment is required to define the set of reasonable responses. For many GE algebra items, this set of reasonable responses consists of all the

values that each variable in the problem can take, where the range for each variable is determined by setting all of variables to a least-restrictive value (in the current experiments, this meant setting the other variable to zero). Other GE items define a space of possible responses. For example, given a matrix of distances between potential stops on a tour, and the restriction of making no more than four stops, the set of possible responses consists of all possible four-stop, three-stop, two-stop, and one-stop tours. Thus, depending on the type of GE item, it might be feasible to make density calculations automatically on different variants of a problem to achieve the desired levels of difficulty.

Theoretically motivated difficulty factors such as solution density and number of verifier constraints have the potential to provide much tighter control over the item development process, whether that process be automatic or manual. The availability of information on the likely difficulty level of an item can potentially reduce the sample size needed for pretesting (e.g., Enright, Morley, & Sheehan, 1999) or allow the dynamic creation of item variants at particular difficulty levels to meet the demands of computer-adaptive testing. Density is particularly useful for the latter purpose, because once a test developer specifies how to define the overall solution space for a particular item, it should be possible to automatically calculate the density of (and therefore predict difficulty for) any item variants produced as an examinee takes a test.

References

Bennett, R. E., Morley, M., Quardt, D., Rock, D. A., & Katz, I. R. (1999a). *Evaluating an underdetermined response type for the computerized SAT* (ETS Research Report 99-22). Princeton, NJ: Educational Testing Service.

Bennett, R. E., Morley, M., Quardt, D., Singley, M. K., Katz, I. R., & Nhouyvanisvong, A. (1999b). Generating examples: A new response type for measuring quantitative reasoning. *Journal of Educational Measurement*, *36*(3), 233-252.

Bridgeman, B., & Cline, F. (1999). *Variations in mean response times for questions on the computer-adaptive GRE General Test: Implications for fair assessment* (ETS Reseach Report 00-07). Princeton, NJ: Educational Testing Service.

Enright, M. K., Morley, M., Sheehan, K. M. (1999). *Items by design: The impact of systematic feature variation on item statistical characteristics* (GRE Research Report No. 95-15R). Princeton, NJ: Educational Testing Service.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist*, 39, 193-202.

Hall, R., Kibler, D., Wenger, E., & Truxaw, C. (1989). Exploring the episodic structure of algebra story problem solving. *Cognition and Instruction*, 6(3), 223-283.

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

Katz, I. R., Bennett, R. E., & Berger, A. (2000). Effects of response format on difficulty of SAT mathematics items: It's not the strategy. *Journal of Educational Measurement*, *37*(1), 39-57.

Katz, I. R., & Berger, A. E. (1995, April). *Strategies underlying score differences on SAT mathematical items.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition.* Cambridge, England: Cambridge University Press.

Kintsch, W., & Greeno, J. G. (1985). Understanding and solving word arithmetic problems. *Psychological Review*, *92*, 109-129.

Koedinger, K. R., & Tabachneck, H. J. M. (1994). *Two strategies are better than one: Multiple strategy use in word-problem solving.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Mayer, R. E., Larkin, J., & Kadane, J. (1984). A cognitive analysis of mathematical problem-solving ability. In R. Sternberg (Ed.), *Advances in the psychology of human intelligence*. Mahwah, NJ: Lawrence Erlbaum Associates.

Nathan, M. J., & Koedinger, K. R. (2000). Teachers' and researchers' beliefs about the development of algebraic reasoning. *Journal for Research in Mathematics Education, 31*(2), 168-190.

Mislevy, R. J., Sheehan, K. M., & Wingersky, M. S. (1993). How to equate tests with little or no data. *Journal of Educational Measurement, 30,* 55-78.

Nhouyvanisvong, A., & Katz, I. R. (1998). The structure of generate-and-test in algebra problem solving. In *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society* (pp. 758-763). Mahwah, NJ: Lawrence Erlbaum Associates.

Nhouyvanisvong, A., Katz, I. R., & Singley, M. K. (1997, March). *Toward a unified model of problem solving in well-determined and underdetermined algebra word problems.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Paige, J. M., & Simon, H. A. (1966). Cognitive processes in solving algebra word problems. In B. Kleinmuntz (Ed.), *Problem solving: Research, method, and theory* (pp. 51-119). New York: Wiley.

Simon, H. (1973). The structure of ill-structured problems. *Artificial Intelligence, 4,* 181-201.

Tabachneck, H. J. M., Koedinger, K. R., & Nathan, M. J. (1995). A cognitive analysis of the task demands of early algebra. In *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society.* Mahwah, NJ: Lawrence Erlbaum Associates.

Notes

[1] We use the term "reasonable" because, literally speaking, the possible range of any response is infinite; an examinee can consider and respond with whatever value s/he wishes.

[2] Participants included students enrolled at Auburn University, The Ohio State University, Penn State, University of Massachusetts—Amherst, University of Pittsburgh, University of South Florida, University of Texas—PanAmerican, and University of Washington.

[3] "GRE test-taking population" refers to all examinees who took the GRE General Test during the two years prior to this experiment – that is, between October 1996 and January 1999. The experimental participants were sampled from this population.

[4] All analyses used scores from the GRE database when possible; however, self-reported scores were used for 29 participants who could not be located in the GRE database (e.g., because the student did not provide a valid SSN).

[5] Because of the type of verifier constraints added (e.g., X > 40Y), it was impossible to add such a constraint and increase item density.

[6] We calculated coefficient alpha by using the 12 item variants for each item as the different "items" in the test; the different items represented the 12 "examinees" in the calculation. This approach is justified given the focus on understanding the within-item variability due to the factors that define the item variants.

[7] The average estimates reported are for the variable receiving the most number of estimates for an item pair. In all cases, the variable mentioned first in the problem received 18-28 estimates while the other variable received eight or fewer estimates. The low number of estimates for one variable suggests that, typically, participants estimated the first variable's value, then derived a value for the second variable through calculation.

54020-009051 • Y42P.4 • Printed in U.S.A.
I.N. 994267