GRE®

RESEARCH

Cognitive Patterns of Gender Differences on Mathematics Admissions Tests

Ann Gallagher
Jutta Levin
Cara Cahalan

September 2002

GRE Board Professional Report No. 96-17P

ETS Research Report 02-19

ETS® Educational Testing Service
Princeton, NJ 08541

Cognitive Patterns of Gender Differences on Mathematics Admissions Tests

Ann Gallagher, Jutta Levin, and Cara Cahalan

GRE Board Report No. 96-17P

September 2002

This 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 not necessarily represent official Graduate Record Examinations Board position or policy.

****

Educational Testing Service
Princeton, NJ 08541

Abstract

A two-part study was conducted to determine whether theoretical work examining gender differences in cognitive processing can be applied to quantitative items on the Graduate Record Examination (GRE®) to minimize gender differences in performance. In Part I, the magnitude of gender differences in performance on specific test items was predicted using a coding scheme. In Part II, a new test was created by using the coding scheme developed in Part I to clone items that elicited few gender-based performance differences. Results indicate that gender differences in performance on some GRE quantitative items may be influenced by cognitive factors such as item context, whether multiple solution paths lead to a correct answer, and whether spatially-based shortcuts can be used.

Key words: Gender, assessment, problem solving, mathematics, graduate admissions

Table of Contents

List of Tables

Page

List of Figures

Page

Introduction

Most standardized tests of mathematical or quantitative reasoning show gender differences in performance, with males outperforming females. Although some differences in strategy use and other mathematical skills (such as fact retrieval) are found in preadolescent samples, differences in mathematical reasoning skills usually appear around adolescence and increase in size as samples become more selective (De Lisi & McGillicuddy-De Lisi, 2002; Willingham & Cole, 1997). In many cases these differences persist despite attempts to control for background variables, such as coursework (Bridgeman & Wendler, 1991). Although gender differences in performance on the mathematics portion of standardized achievement tests appear to be decreasing over time (Hyde, Fennema, & Lamon, 1990), gender differences in performance have persisted on high-stakes tests that are intended to assess mathematical reasoning and problem solving, such as the SAT I: Reasoning Test (SAT I®) and the Graduate Record Examinations (GRE®) General Test (De Lisi & McGillicuddy-De Lisi, 2002; Willingham & Cole, 1997).

Analyses of item content have yielded inconsistent results as to which specific content consistently favors males over females (Doolittle & Cleary, 1987; McPeek & Wild, 1987; O'Neill, Wild, & McPeek, 1989), but general patterns of performance are evident. Most studies find that women generally perform better on algebra problems than on geometry problems, and that men outperform women most consistently on problems requiring unconventional use of mathematical knowledge or those classified as reasoning problems (Armstrong, 1985; Dossey, Mullis, Lindquist, & Chambers, 1988; Gallagher & DeLisi, 1994; Willingham & Cole, 1997). In the two studies presented here, we sought to determine whether theoretical work examining gender differences in cognitive processing skills could be applied to quantitative items on the GRE General Test to help minimize gender differences in performance.

The evidence of gender differences in performance on standardized tests of mathematics — combined with the fact that, even at the most advanced levels, women do as well or better than men in mathematics courses (Kimball, 1989) — suggests that these high-stakes tests may be assessing a different or more narrowly defined construct than that which is reflected by course grades. This is not surprising, given the differences in time allotment and, as a consequence, the types of tasks students and examinees are required to perform in each setting. Standardized tests generally rely on a large number of rapidly generated responses, whereas coursework at advanced levels usually requires sustained work on a relatively smaller set of complex problems.

Work by Gallagher and DeLisi (1994) suggests that, among high-performing students, attributes of test questions may influence the types of strategies students use for solving problems. In their study, gender differences in performance were related to differential strategy use on specific kinds of questions. This interaction between question attributes and solution strategies may explain some of the gender differences found among highly trained students applying to graduate programs in technical fields. More recent work by Fennema, Carpenter, Jacobs, Franke, and Levi (1998) supports work by Gallagher and DeLisi and suggests that the same type of differential strategy use may be evident even in grade school.

Halpern suggests that there may be common cognitive processes underlying tasks that favor males or females across a variety of content domains (see Halpern, 1997, p. 1102 for a complete list of tests and tasks that show gender effects). According to Halpern's hypothesis, women appear to excel at tasks that require rapid access and retrieval of information from long-term memory. These tasks include associational fluency tasks (such as generating synonyms), language production, anagrams, and computational tasks. Men appear on average to excel at tasks that require the retention and manipulation of visual representations in working memory. These tasks include mental rotation and spatial perception tasks, verbal analogies, and some types of mathematical problem solving. This perspective has recently received independent confirmation in research conducted by Casey, Nuttall, Pezaris, and Benbow (1995) and by Casey, Nuttall, and Pezaris (1997). Both studies show that mental rotation is a critical mediator of gender differences on the mathematics portion of the SAT I test, especially for students with high ability.

In the two-part study described here, we sought to combine theoretical perspectives from this prior work. In Part I, cognitive correlates of gender differences in task performance were added to the item and solution attributes identified by Gallagher and DeLisi (1994; see Appendix A for a description of the coding categories used by Gallagher & DeLisi). The resulting expanded coding scheme includes factors identified by Halpern (1997) and by Casey and her colleagues (1997) in accounting for gender differences in mathematical problem solving. The new coding scheme was applied to quantitative questions on the GRE General Test in an effort to predict group differences in test performance.

In Part II, information about factors affecting gender differences in performance was used to create a prototype test. For this prototype, items were generated by cloning existing items with known statistics, but keeping in mind information about solution strategy requirements. Items to be cloned were selected from a pool of 70 items that had been created in 1995 for the prototype GRE Mathematical Reasoning Test.[1] Clones of these items were then administered to students in the target population for the trial test — that is, those planning graduate study in mathematics or technical sciences.

Part I

As noted above, the purpose of Part I was to determine whether specific cognitive attributes of test items and their solutions were related to the magnitude and direction of gender differences in performance on those items. The new coding scheme used in Part I was based on prior research and hypotheses generated from an examination of common features of groups of test items with either very large or very small gender differences in performance. Specifically, solution attributes identified in Gallagher and DeLisi's (1994) work and tasks identified by Halpern (1997) as favoring either males or females (see Appendix B) were used as the starting point for examining "high impact" and "low impact" items from an experimental GRE Mathematical Reasoning Test data collection.

Method

In developing the new taxonomy, one of our guiding principles was the idea that women tend to solve mathematics problems the way they have been taught to solve such problems. Questions that are similar to problems in current textbooks were predicted to show less impact than questions that *look* like standard textbook questions but actually require unusual strategies. An additional, related principle was that males and females tend to differ in their use of spatial representations to solve problems. When the use of a spatial representation is optional (e.g., the problem can be solved with either a spatial method or an analytical method), males are more likely than females to use it as a preferred strategy. In contrast, when the solution to a problem clearly requires the use of a spatial representation, both male and female test takers will incorporate the representation into their strategy. Use of spatial strategies, therefore, will only affect performance on items where other kinds of strategies can lead to a correct answer, but where a spatial solution provides some advantage (e.g., speed or accuracy).

Data collected during prototype testing of the GRE Mathematical Reasoning Test in the fall of 1995 and the spring of 1996 were examined for evidence of performance patterns by gender. The 10 highest impact items and the 10 lowest impact items were examined to determine whether sources of variation suggested by Gallagher and DeLisi (1994), Halpern (1997), and Casey et al. (1997) were evident in these items. Selected elements suggested by all three studies were clearly evident in both sets of items.

Figure 1 presents the new coding scheme. It includes item context (that is, the problem setting or story surrounding the problem) and use of specific language and spatial skills, but also contains features that were used to classify items as conventional and unconventional in Gallagher and DeLisi's (1994) work. Individual codes with similar characteristics were grouped into categories. The six resulting categories were verbal, spatial, conventional, unconventional, multistep, and multiple-solution. Finally, overall low-impact and high-impact categories were constructed by combining categories described as likely to favor females or males, respectively. To test whether this coding scheme was capable of predicting gender differences in performance on specific problems administered under actual test-taking conditions, items from three forms of the GRE quantitative section were coded using the new taxonomy, and gender differences among two groups of test takers were examined.

Subjects. Given differences reported in earlier work (Kimball, 1989) and differences reported by Casey et al. (1995; 1997) for high- versus average-ability groups on the mathematics portion of the SAT I, it was important that analyses be conducted with relatively homogeneous groups of students (in terms of mathematics training) and that items were pitched at an appropriate level of difficulty. For this reason, analyses were conducted on two groups of students who employ mathematics to varying degrees in their studies: a) arts, humanities, and social science students, and b) technical science students (who use calculus on a regular basis in their coursework and are likely to be homogenous in mathematics training).

We hypothesized that the strongest relationship between taxonomy categories and gender differences in performance would be found among students in the arts, humanities, and social sciences group, and that the weakest relationship would be found for students in the technical sciences group. A key consideration in making this prediction was the appropriate match between item difficulty and mathematics training within each group. If difficulty was either too high or too low, there would be little variation in performance.

Spatial

- Requires the conversion of a word problem to a spatial representation (i.e., generation of spatial format). Spatial representation is an important part of the problem.
- Requires using a given spatial representation (e.g., converting it to a mathematical expression or extracting information to be used in solving a problem). Spatial representation is an important part of the problem.
- Spatial information must be maintained in working memory while other spatial information is being transformed (e.g., maintaining a particular shape in working memory so that it can be compared with a transformed shape). Formerly called “visual spatial memory span.” For example, given the graph of the derivative of f, find the graph of f.

Verbal

- Requires the conversion only. Typically mathematical expression items or mathematical calculation with expressions as multiple-choice answers.
- Verbal information must be maintained in working memory while additional information is being processed; primarily used for items with heavy verbal load.
- Reading math (e.g., using a newly defined function or understanding the properties of an algebraic expression).

Multiple-solution

- More than one solution path leads to a correct answer, and the quick solution is imaginative or insightful, while the slower solution is more systematic and planful.
- More than one solution path leads to a correct answer, and one solution, usually the faster one, involves drawing a picture.
- More than one solution path leads to a correct answer. Test-taking skills can contribute to a faster solution.

Multistep

- The problem is multistep and requires accuracy and a systematic approach. For example, two successive calculations must be done and the second calculation uses information from the first calculation.

Unconventional

- More than one solution path leads to a correct answer, and the quick solution is imaginative or insightful, while the slower solution is more systematic and planful.
- More than one solution path leads to a correct answer, and one solution, usually the faster one, involves drawing a picture.
- More than one solution path leads to a correct answer. Test-taking skills can contribute to a faster solution.
- The context looks like a familiar one, but the solution is NOT one that is generally associated with the context; (e.g., on the first glance the problem appears to deal with averages, but to solve it one needs to use a rate of growth).
- Quantitative comparison items in which the relationship cannot be determined from the information given.

Conventional

- Requires labeling the problem as a specific type of problem and/or retrieving a formula or routine that should be known from memory, but is not immediately apparent. (We used this only for nonobvious cases; very obvious, standard retrieval problems were coded using the definition that follows.)
- Typical text-book problems — the context is a familiar one, frequently seen in mathematics coursework; the solution path is one that is generally associated with the context (not used for quantitative comparison items).
- The problem is multistep and requires accuracy and a systematic approach. For example, two successive calculations must be done and the second calculation uses information from the first calculation.
- Pure algebraic manipulation
- Pure calculation
- Reading math (e.g., using a newly defined function or understanding the properties of an algebraic expression).
- Quantitative comparison items in which the two quantities are equal.

Figure 1. **New coding categories for items.**

The difficulty level of the GRE quantitative test is most appropriate for students in the arts, humanities, and social sciences group and least appropriate for students in the technical sciences group (who generally score very high on tests of mathematical reasoning, creating a ceiling effect). Because the cognitive attributes of any problem are a function of the problem-solver's knowledge base, a problem that would be considered a reasoning problem for someone with little mathematical training might be considered a standard application of a memorized algorithm for someone with extensive training.

Data source. Item response data were drawn from three forms of the GRE quantitative test administered during the fall of 1995. Participants consisted of all examinees reporting their best language to be English. They were divided into two groups based on self-reported undergraduate major:

- arts, humanities, and social sciences (e.g., English and foreign languages, history, performance and studio arts, anthropology, economics, political science, psychology, and sociology); n = 51,333
- technical sciences (e.g., chemistry, computer science, engineering, mathematics, and physics); n = 24,962

Table 1 displays descriptive statistics for each group by test form, undergraduate major, and gender.

Table 1

Average Percent Correct by Test Form, Undergraduate Major, and Gender

		Percent correct						
		Male			Female			Impact:
Form	Major	M	N	SD	M	N	SD	$P^+_f - P^+_m$
A	Arts, humanities, and social sciences	.54	10131	.17	.49	20282	.16	-.059
	Technical sciences	.74	9282	.15	.69	3186	.16	-.058
B	Arts, humanities, and social sciences	.57	3769	.19	.50	8086	.17	-.072
	Technical sciences	.77	2676	.14	.72	1336	.16	-.053
C	Arts, humanities, and social sciences	.60	2882	.19	.53	6183	.18	-.071
	Technical sciences	.80	6423	.15	.77	2059	.15	-.030

Procedure. Items from the three GRE test forms were coded for the cognitive attributes of their questions and solutions according to the 22 characteristics listed in Figure 1. Coding was performed independently by two mathematicians experienced in writing test items for the GRE quantitative test. Initial interrater agreement for all item codings was 70 percent. Disagreements in coding were resolved through discussion, and a single final set of codes was agreed upon.

Items were coded only for the most salient characteristics of solutions, since many items are highly complex in the set of skills that are tapped. In some cases, more than one code applied to items. For example, 28 of the dual-coded items required recall of a formula or algebraic manipulation (predicted to produce low impact), but also required the translation of an expression given mathematically into a spatial array (predicted to show high impact). An additional 12 items required substantial amounts of calculation or verbal memory (likely to show low impact), but could be solved by way of two different strategies, one being substantially quicker and likely to involve drawing a picture (again, likely to show high impact).

Once items were coded, item codes were clustered into the six higher-level item-type categories described in Figure 1 (verbal, spatial, conventional, unconventional, multiple-solution, and multistep). Three of these categories were predicted to have high impact (spatial, unconventional, and multiple-solution) and three were predicted to have low impact (verbal, conventional, and multistep). Two items did not clearly fit any of the 22 coding categories. Final codings of the 178 items from the three forms yielded 82 items coded as likely to show low impact and 63 items coded as likely to show high impact. The remaining 33 items were double coded. For example, an item that requires multiple steps to interpret a graph would have been coded as multistep (low impact) and spatial (high impact). Also, any one item may have been coded as more than one type within an impact category — for example, a given high-impact item may have been coded as both spatial and unconventional. Table 2 shows the number of items for each test form falling into code categories.

Table 2

Number of Items by Code Category and Test Form

Code Category	Form A	Form B	Form C
Spatial	20	18	18
Unconventional	17	25	15
Multiple-solution	14	22	6
Total number of items predicted to have high impact*	32	33	31
Verbal	10	10	8
Conventional	38	30	35
Multistep	17	9	8
Total number of items predicted to have low impact*	41	33	41

* A large proportion of items were assigned to more than one code category. Therefore, the sum of the numbers in the table is larger than the total number of items.

Analyses

Items were grouped into coding categories based on their predicted levels of impact. Because many items were assigned codes that predicted both high- and low-impact, items were divided into three categories:

- combined — assigned both codes that predict high impact AND codes that predict low impact
- only high — assigned only codes that predict high impact
- only low — assigned only codes that predict low impact

Table 3 displays the distribution of items by code category and predicted impact. Here again, any given item may have been coded as more than one type within an impact category or as having both low impact and high impact.

Table 3

Frequency Distributions of Items by Code Category and Predicted Impact Category

	Predicted impact category						
	Combined		Only high		Only low		Total
Code category	N	%	N	%	N	%	N
Spatial	26	46.43	30	53.57	0	0.00	56
Unconventional	7	12.28	50	87.72	0	0.00	57
Multiple-solution	3	6.67	42	93.33	0	0.00	45
Total number of items predicted to have high impact*	33	34.38	63	65.63	0	0.00	96
Verbal	5	17.86	0	0.00	23	82.14	28
Conventional	26	25.24	0	0.00	77	74.76	103
Multistep	16	34.78	0	0.00	30	65.22	46
Total number of items predicted to have low impact*	33	28.70	0	0.00	82	71.30	115

* A large proportion of items were assigned to more than one code category. Therefore, the sum of the numbers in the table is larger than the total number of items.

For each item, effect size was calculated using the following equation:

$$d = \frac{\overline{X}_f - \overline{X}_m}{\sqrt{\frac{SD_m^2 + SD_f^2}{2}}}$$

where $\overline{X}_f$ is the mean percent correct for female examinees and $\overline{X}_m$ is the mean percent correct for male examinees, and SD$_f$ and SD$_m$ are the respective standard deviations for female and male examinees. Thus, positive values indicate that the item is more likely to favor female test takers and negative values indicate that the item is more likely to favor male test takers. This formula is less dependent on subgroup sample sizes than a formula using weighted standard deviations (Willingham & Cole, 1997, p. 21). Calculations were done separately for test takers in the two undergraduate major groups — a) arts, humanities, and social sciences, and b) technical sciences.

For all analyses, skipped items — defined as items that were not answered, but were followed by answered items — were coded as incorrect. Items that were not reached — defined as items that were left blank following the last answered item on the test — were not included.

All analyses were conducted for the two groups based on undergraduate major. To control for the effects of item difficulty, analyses of covariance (ANCOVAs) were used to examine data. We felt it important to control difficulty to insure that our findings were not confounded. Prior research has found larger gender differences in more select populations (e.g., Willingham & Cole, 1997), suggesting that greater differences favoring males may be found on more difficult items. We also felt it important to control for difficulty because mean difficulty varied across the groups of items that were examined.

ANCOVAs were conducted to determine whether the effect size was different between the two category groups (items likely to have high impact and items likely to have low impact) while controlling for difficulty (percent correct). In addition to the overall high-impact and low-impact categories, some of the codes were combined to create three contrasting subcategories: a) spatial versus verbal, b) multiple-solution versus multistep, and c) conventional versus unconventional, which are described below.

Spatial versus verbal. Problems coded as spatial were predicted to be high impact and problems with verbal codes were predicted to have low impact. Spatial problems do not simply contain a graph or figure; they rely heavily on spatial skills (e.g., converting a word problem to a spatial representation, using a spatial representation, maintaining spatial information in working memory). Similarly, verbal problems rely heavily on verbal skills (e.g., translating math to English or English to math, reading math, maintaining verbal information in working memory).

Multiple-solution versus multistep. Multiple-solution problems were predicted to have high impact, whereas multistep problems were predicted to have low impact. Multiple-solution problems, by definition, could be approached more than one way; in this case, all had at least one solution that was faster, and often required drawing a picture or using test-wiseness skills. Multistep problems require a systematic approach — for example, the solution to one problem could require two successive calculations, with the second calculation using information from the first calculation.

Unconventional versus conventional. Unconventional problems were predicted to have high impact, while conventional problems were predicted to have low impact. Unconventional problems include a) problems with multiple-solution paths, of which one is much quicker, b) problems with contexts that looks familiar, but solution methods that are unfamiliar, or c) quantitative-comparison items for which the answer cannot be determined. Conventional problems, on the other hand, include a) problems that require examinees to label a problem as a specific type, b) multistep problems, c) problems with familiar contexts and solutions, d) pure calculation problems, e) pure algebra problems, f) problems that require high-level reading of mathematics (e.g., using a newly defined function or understanding the properties of an algebraic expression) or g) quantitative-comparison items in which the two given quantities are equal.

Results

Items in the categories “only high” and “only low” were compared using ANCOVA to control for item difficulty. Items that fell into the “combined” category were not included in this analysis. Item difficulty was defined as the percent of all test takers within each undergraduate major who answered the item correctly. Using effect size as the dependent variable and the broad coding categories as the independent variable, a significant difference was found for technical science majors [F (1, 145) = 5.35, p <.05] and arts, humanities, and social science majors [F (1, 145) = 5.81, p <.01]. Table 4 displays mean effect sizes by category and major; Appendix C and Appendix D display the complete ANCOVA tables.

Table 4

Effect Sizes by Predicted Item Impact Category

	Effect size		
	N	M	SD
Arts, humanities, and social sciences			
Only high, no low codes	63	-.16	.08
Only low, no high codes	82	-.13	.08
Technical sciences			
Only high, no low codes	63	-.14	.09
Only low, no high codes	82	-.10	.08

Analyses were also conducted to investigate whether the number of characteristics related to impact affected the magnitude of the effect size. These analyses compared items with one, two, or three high-impact codes (within the category of “high impact”) as well as items with one, two, or three low-impact codes (within the category of “low impact”). Neither analysis revealed significant differences. It appears to make no difference whether an item has one high-impact code or three.

Finally, ANCOVAs were conducted to examine differences in the effect size of selected contrasting code categories — spatial versus verbal, multiple-solution versus multistep, and unconventional versus conventional — after controlling for item difficulty. As noted earlier, verbal, conventional, and multistep items were expected to have low impact, while spatial, unconventional, and multiple-solution items were expected to have high impact. ANCOVA was used for these contrasts, with item difficulty (mean percent correct) as the covariate.

Analysis of the spatial-versus-verbal contrast showed a significant difference for technical science majors [F (1, 74) = 4.74, p < .05], but not for arts, humanities, and social science majors. By contrast, the unconventional-versus-conventional contrast showed no significant difference for the technical science majors, but revealed a significant difference in effect sizes for arts, humanities, and social science majors [F (1, 160) = 6.49, p < .01]. Finally, the multiple-solution-versus-multistep contrast showed a significant difference in effect for both technical science majors [F (1, 85) = 10.17, p < .01] and arts, humanities, and social science majors [F (1, 85) = 7.27, p < .01]. Table 5 presents descriptive statistics for code categories used in contrasts for arts, humanities, and social science majors, while Table 6 provides the same information for technical science majors.

Table 5

Average Effect Sizes by Code Category for Arts, Humanities, and Social Science Majors

	Effect size		
Code category	N	M	SD
Verbal	23	-.14	.07
Spatial	51	-.16	.07
Conventional	103	-.13	.08
Unconventional	57	-.17	.07
Multistep	43	-.13	.07
Multiple-solution	42	-.17	.07

Note. Items coded as both verbal and spatial or both multistep and multiple-solution were not included in analyses.

Table 6

Average Effect Sizes by Code Category for Technical Science Majors

	Effect size		
Code category	N	M	SD
Verbal	23	-.09	.05
Spatial	51	-.12	.07
Conventional	103	-.11	.08
Unconventional	57	-.13	.09
Multistep	43	-.10	.06
Multiple-solution	42	-.14	.10

Note. Items coded as both verbal and spatial or both multistep and multiple-solution were not included in analyses.

Discussion

The coding scheme created here was based on the notion that specific cognitive processing requirements are likely to affect the size of gender differences in performance on mathematical reasoning items. This coding scheme — which was based on cognitive characteristics previously identified as likely to favor either male or female students — was successful in explaining a portion of gender differences in performance on GRE quantitative items. By dividing test takers into groups based on major field of study, we were able to control (in a rough manner) for mathematics coursework.

Items involving the use of language skills (i.e., the language of mathematics) and storage and retrieval of information from memory were predicted to show the smallest gender differences. Items requiring generation of a spatial representation and manipulation of a given representation, among other things, were predicted to show the largest differences. Results of data analyses support these predictions. Results also support Halpern's (1997) theory of the types of tasks at which male students and female students excel, and are consistent with recent work highlighting the importance of spatial skills for performance on standardized tests of mathematics (Casey et al., 1997). One interesting finding was that items with two or three high-impact characteristics did not have significantly greater impact than items that only had one high-impact characteristic.

Finally, we predicted that the largest gender differences would be found among students with arts, humanities and social science majors because the difficulty of the test was most appropriate for that group. Although this was true in most cases, in the spatial-versus-verbal contrast, differences in effect sizes were only significant for the technical sciences group. This suggests that gender differences in performance on these code categories may only be found on the most difficult items.

Part II

The purpose of Part II was to determine whether constructs that were identified in Part I could be used to control impact. To investigate this possibility, items showing low impact in the GRE Mathematical Reasoning Test prototype data were used to create a low-impact test, and this test was cloned. The cloned test was then administered to students in the GRE mathematical reasoning target population as part of a 27-section research package.

Method

Procedure. Draft test specifications were used to select 24 items from a pool of 70 items that had been developed for two GRE Mathematical Reasoning prototype tests, which were administered to students in the fall of 1995 and spring of 1996. Items were selected on the basis of three criteria:

1. their demonstrated impact in the prototype testing
2. cognitive characteristics of their solutions, as outlined in the current classification scheme
3. requirements delineating relative proportions of code categories within the test

The selected items were assembled as a two-section prototype test, with each section containing 12 items. After the prototype was assembled, all of the items in the new test were cloned. The goal of the cloning was to create new items that were different in appearance but that maintained relevant task requirements. Problem scenarios, values in computations, variable names, and other names were changed in cloned versions of problems. Elements identified in Part I as likely to affect performance were maintained.

Item clones were designed to require types of mathematical operations similar to the original problems and to retain the original level of complexity. If the original problem was an applied problem, the clone was set in a similar context or field. The length of the stem and the reading load for clones was also designed to be similar to the original problems. Appendix E presents examples of original low-impact items and their clones.

The cloned prototype test was then administered to 60 students as part of the GRE Mathematical Reasoning Test development project. All examinees completed both sections of the test, and were allowed one hour to complete each 12-item section.

Subjects. All of the 60 students who participated in the GRE Mathematical Reasoning Test development project were majoring in either engineering, mathematics, chemistry, physics, or computer science at the time they took the test. Thirty-five of the 60 were males, and 25 were females. All but six students reported an overall grade-point average (GPA) of B or better, with a greater proportion of males (60%) reporting an overall GPA of A or A- than was reported by females (52%). Of the 60 examinees, 32 reported their race as White and 27 reported their race as African American. One student did not report race/ethnicity. Fifty of the students reported that English was their best language. This sample of examinees is very similar to the sample that took the test on which the clone was based.

Analyses

Data were examined using analysis of covariance. Students' self-reported GRE quantitative scores was used as a covariate and mean percent correct was the dependent variable.

Results

Table 7 displays average GRE quantitative scores, as well as mean percent correct on the prototype test, for male and female participants. Results indicate that, like the original items, the impact of the cloned items was low. As Table 7 shows, male and female examinees were well matched on GRE quantitative score (the male mean was 674 while the female mean was 676). Even though mean percent correct on the prototype is higher for females than males, using GRE quantitative score as a covariate showed there was no significant difference in performance ($\underline{F}$ [1, 57] = 3.4, $\underline{p} < .07$). Impact for this prototype was 0.51 using the formula described earlier, with females outperforming males.

Table 7

Average GRE Quantitative Score and Mean Percent Correct for Prototype by Gender

	Males		Females	
Test	M	SD	M	SD
GRE quantitative score	674	95.0	676	95.0
Prototype: Mean percent correct	69.2	16.6	74.8	21.1

Discussion

The cloned prototype test was highly successful in maintaining the very low impact of the original test, even though cloned items were quite different from the original items. This strongly suggests that factors selected to be held constant in these items during cloning were relevant influences on gender impact, and that factors that were manipulated were not. These results should prove useful to test developers when cloning test questions in the future.

Conclusions

Performance on standardized tests — such as the quantitative portion of the GRE General Test — is one of several factors that influence students as they plan careers in higher education. For both theoretical and applied reasons, it is important to gain a better understanding of factors that affect performance on these kinds of tests (Casey et al., 1997). From a cognitive-strategy perspective, the work reported here has contributed to our understanding of the sources of score differences between male and female examinees with similar academic backgrounds. The magnitude of gender differences in performance on some GRE quantitative items may be affected by factors such as the context of the problem setting, whether multiple solution paths lead to a correct answer, and whether spatially-based shortcuts can be used.

The findings of this two-part study have important implications for the test development process — particularly test creation and item-writing. The increasing need to produce more test items at a faster rate has led to an increase in item cloning. Small changes in surface features of items frequently result in changes to item statistics. The research presented here provides important information regarding the kinds of changes that can be made to low-impact items without altering their impact statistics.

The study also has validity implications. Qualitatively different approaches to mathematics problems that may be used by male and female test takers can lead to performance differences that may or may not be relevant to the test construct. Factors affecting performance should be evaluated with regard to their importance to mathematical reasoning. If deemed important, they should form part of the test specifications. If not, efforts should be made to minimize such factors. These qualitative differences should be taken into account when test questions are written and when questions are assembled into tests. More detailed and explicit statements regarding which cognitive skills and abilities are important to the test construct and which skills are not relevant will help to insure that performance differences reflect what the test intends to predict.

References

Armstrong, J. M. (1985). A national assessment of participation and achievement of women in mathematics. In S.F. Chipman, L.R. Brush, & D. M. Wilson (Eds.), *Women and mathematics: Balancing the equation* (pp. 59-94). Hillsdale, NJ: Erlbaum.

Bridgeman, B., & Wendler, C. (1991). Gender differences in predictors of college mathematics performance and in college mathematics course grades. *Journal of Educational Psychology, 83*, 275-284.

Casey, M. B., Nuttall, R., & Pezaris, E. (1997). Mediators of gender differences in mathematics college entrance test scores: A comparison of spatial skills with internalized beliefs and anxieties. *Developmental Psychology, 33,* 669-680.

Casey, M. B., Nuttall, R., Pezaris, E., & Benbow, C. P. (1995). The influence of spatial ability on gender differences in mathematics college entrance test scores across diverse samples. *Developmental Psychology, 31,* 697-705.

De Lisi, R., & McGillicuddy-De Lisi, A. (2002). Sex differences in mathematical abilities and achievement. In A. McGillicuddy-De Lisi & R. De Lisi (Eds.), *Biology, society and behavior: The development of sex differences in cognition* (pp. 155-182). Westport, CT. Ablex.

Doolittle, A. E., & Cleary, T. A. (1987). Gender-based differential item performance in mathematics achievement items. *Journal of Educational Measurement*, *24*(2), 157-166.

Dossey, J. A., Mullis, I. V. S., Lindquist, M. M., & Chambers, D. L. (1988). *The mathematics report card: Are we measuring up? Trends and achievement based on the 1986 National Assessment*. Princeton, NJ: Educational Testing Service.

Fennema, E., Carpenter, T. P., Jacobs, V. R., Franke, M. L., & Levi, L. W. (1998). A longitudinal study of gender differences in young children's mathematical thinking. *Educational Researcher, 27*(5), 6-11.

Gallagher, A. M., & DeLisi, R. (1994). Gender differences in Scholastic Aptitude Test mathematics problem solving among high ability students. *Journal of Educational Psychology*, *86*(2), 204-211.

Halpern, D. F. (1997). Sex differences in intelligence. Implications for education. *American Psychologist, 52*, 1091-1102.

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin, 107*, 139-153.

Kimball, M. M. (1989). A new perspective on women's math achievement. *Psychological Bulletin, 105*, 198-214.

McPeek, W. M., & Wild, C. L. (1987). *Characteristics of quantitative problems that function differently for men and women.* Paper presented at the 95th annual convention of the American Psychological Association, New York.

O'Neill, K., Wild, C. L., & McPeek, W. M. (1989). *Gender-related differential item performance on graduate admissions tests.* Paper presented at the annual meeting of the American Psychological Association, San Francisco.

Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment*. Mahwah, NJ: Erlbaum.

Notes

[1] This trial test was administered in 1995 and 1996. It had been developed specifically for students in mathematics and the technical sciences, where the average scores are consistently close to the top of the scale. However, large subgroup differences proved difficult to eliminate, and pending further research, further development of the test has been postponed.

[2] Given that males outperform females on most items, items on which males tended to have an advantage were designated as “high impact” while items on which males tended to have little or no advantage were deemed “low impact.” (Impact was defined as the difference between the percent correct for male test takers and the percent correct for female test takers. Negative values indicate a male advantage, while positive values indicate a female advantage.)

Appendix A: Coding Categories (Gallagher & De Lisi, 1994)

Conventional

- Solution consists primarily of computational strategies generally taught in school. This includes computations and algebraic formulas using abstract terms or givens from the problem stem.
- Solutions where values are assigned to variables given in the problem stem. This includes trial-and-error solutions that use random numbers and solutions where the assigned values make the operations more concrete.
- Solutions where the student works backward from the options and systematically plugs in choices.

Unconventional

- Solutions that use a mathematical algorithm but are simplified or shortened by the student's insight, logical reasoning, or estimation. This includes solutions for which the student realizes that it is not necessary to complete an equation or algorithm to choose an option.
- Solutions based primarily on the application of mathematical principles or logic, either alone or in combination with estimation or insight. These solutions generally do not include computations or algorithms, but may include minor mental calculations.

Appendix B: Coding Categories (Halpern, 1997)

Tasks on which women
obtain higher scores

- tasks that require rapid access to and use of phonological, semantic, and other information in long-term memory
- production and comprehension of complex prose
- fine motor tasks
- perceptual speed
- decoding nonverbal communication
- perceptual thresholds (e.g., lower thresholds for touch, taste, and odor)
- speech articulation (e.g., tongue twisters)

Tasks on which men
obtain higher scores

- tasks that require visual transformations in short-term memory (e.g., mental rotation)
- tasks that involve moving objects
- motor tasks that involve aiming
- tests of fluid reasoning (e.g., proportional reasoning, mechanical reasoning, verbal analogies)

Appendix C: ANCOVA Tables for Arts, Humanities, and Social Science Majors

Analysis of Covariance of Effect Size by Predicted Impact

Source	df	*F*	*p*
Difficulty	1	5.24*	.02
Predicted impact (high vs. low)	1	5.81**	.01
Error	142	(0.01)	

Analysis of Covariance of Effect Size (1, 2, or 3 High Codes)

Source	df	*F*	*p*
Difficulty	1	7.36**	.01
Number of high codes	2	0.43	.66
Error	59	(0.01)	

Analysis of Covariance of Effect Size (1, 2, or 3 Low Codes)

Source	df	*F*	*p*
Difficulty	1	0.99	.32
Number of low codes	2	0.23	.79
Error	78	(0.01)	

Analysis of Covariance of Effect Size (Verbal vs. Spatial)

Source	df	*F*	*p*
Difficulty	1	2.67	.11
Verbal vs. spatial	1	1.18	.28
Error	71	(0.01)	

Analysis of Covariance of Effect Size (Conventional vs. Unconventional)

Source	df	*F*	*p*
Difficulty	1	8.65***	.00
Conventional vs. unconventional	1	6.49**	.01
Error	157	(0.01)	

Analysis of Covariance of Effect Size (Multistep vs. Multiple-Solution)

Source	df	*F*	*p*
Difficulty	1	4.22*	.04
Multistep vs. multiple-solution	1	7.27**	.01
Error	82	(0.00)	

*p < .05. **p < .01. ***p < .001.

Appendix D: ANCOVA Tables for Technical Science Majors

Analysis of Covariance of Effect Size by Predicted Impact

Source	df	*F*	*p*
Difficulty	1	20.89***	.00
Predicted impact (high vs. low)	1	5.35*	.02
Error	142	(0.01)	

Analysis of Covariance of Effect Size (1, 2, or 3 High Codes)

Source	df	*F*	*p*
Difficulty	1	7.85**	.01
Number of high codes	2	0.26	.77
Error	59	(0.01)	

Analysis of Covariance of Effect Size (1, 2, or 3 Low Codes)

Source	df	*F*	*p*
Difficulty	1	16.53***	.00
Number of low codes	2	2.35	.10
Error	78	(0.01)	

Analysis of Covariance of Effect Size (Verbal vs. Spatial)

Source	df	*F*	*p*
Difficulty	1	1.39	.24
Verbal vs. spatial	1	4.74*	.03
Error	71	(0.00)	

Analysis of Covariance of Effect Size (Conventional vs. Unconventional)

Source	df	*F*	*p*
Difficulty	1	16.83***	.00
Conventional vs. unconventional	1	3.66	.06
Error	157	(0.01)	

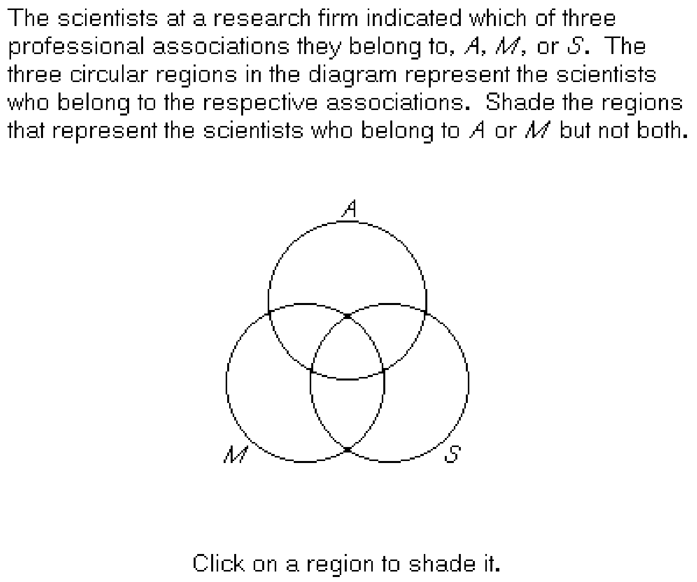
Analysis of Covariance of Effect Size (Multistep vs. Multiple-Solution)

Source	df	*F*	*p*
Difficulty	1	11.93***	.00
Multistep vs. multiple-solution	1	10.17***	.00
Error	82	(0.01)	

*p < .05. **p < .01. ***p < .001.

Appendix E: Example Low-Impact Items and Clones

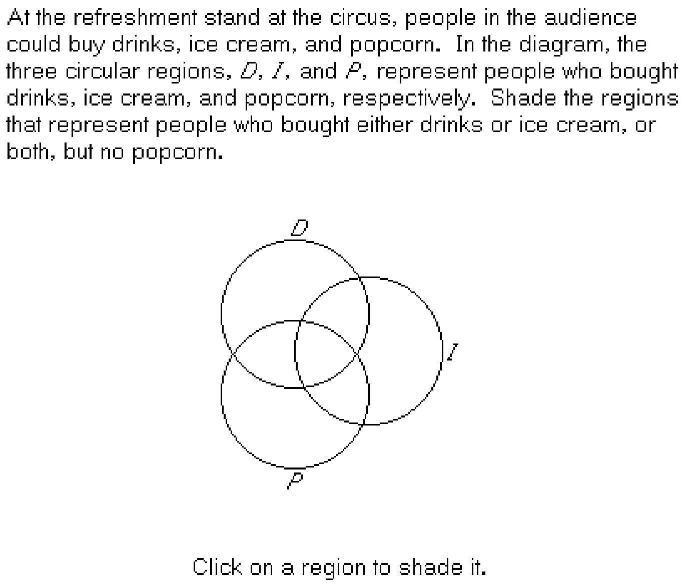
Original Item 1



The scientists at a research firm indicated which of three professional associations they belong to, *A*, *M*, or *S*. The three circular regions in the diagram represent the scientists who belong to the respective associations. Shade the regions that represent the scientists who belong to *A* or *M* but not both.

Click on a region to shade it.

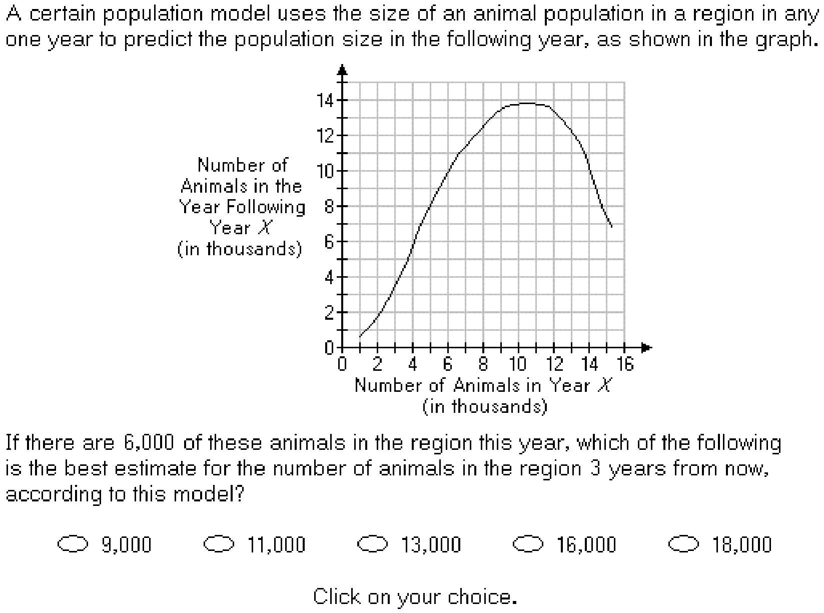
Item 1 Clone



At the refreshment stand at the circus, people in the audience could buy drinks, ice cream, and popcorn. In the diagram, the three circular regions, *D*, *I*, and *P*, represent people who bought drinks, ice cream, and popcorn, respectively. Shade the regions that represent people who bought either drinks or ice cream, or both, but no popcorn.

Click on a region to shade it.

Original Item 2



A certain population model uses the size of an animal population in a region in any one year to predict the population size in the following year, as shown in the graph.

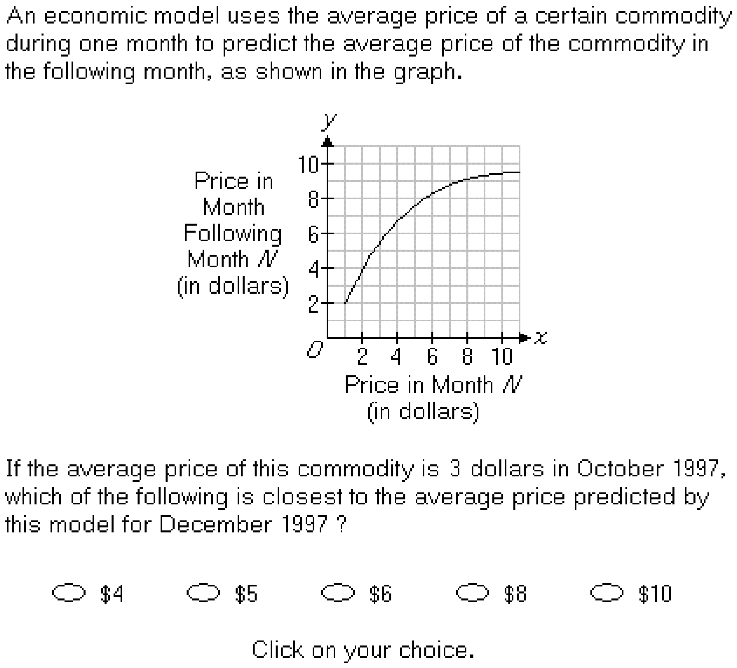
Number of Animals in the Year Following Year X (in thousands)
0
2
4
6
8
10
12
14
16
Number of Animals in Year X (in thousands)

If there are 6,000 of these animals in the region this year, which of the following is the best estimate for the number of animals in the region 3 years from now, according to this model?

○ 9,000 ○ 11,000 ○ 13,000 ○ 16,000 ○ 18,000

Click on your choice.

Item 2 Clone



An economic model uses the average price of a certain commodity during one month to predict the average price of the commodity in the following month, as shown in the graph.

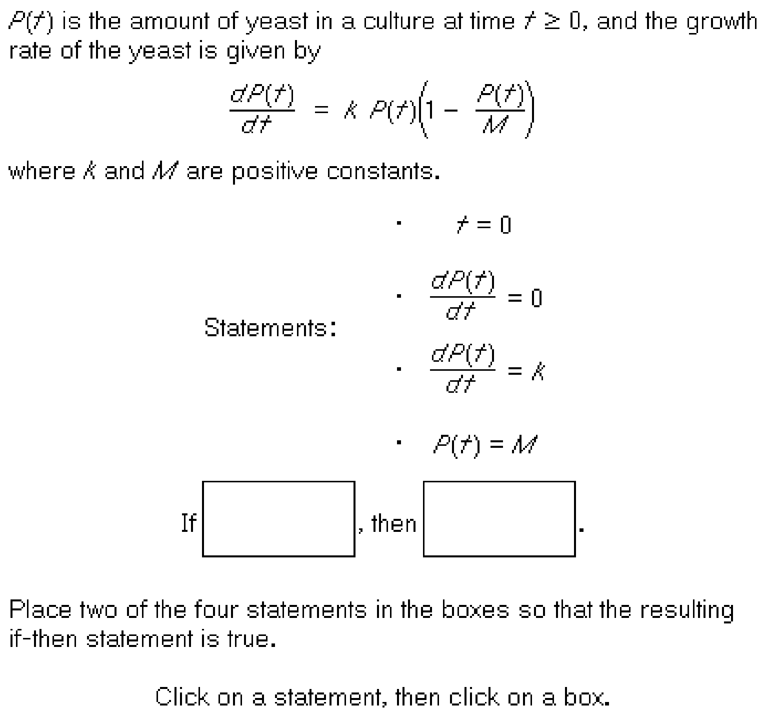
y
Price in Month Following Month N (in dollars)
10
8
6
4
2
O
2
4
6
8
10
x
Price in Month N (in dollars)

If the average price of this commodity is 3 dollars in October 1997, which of the following is closest to the average price predicted by this model for December 1997 ?

○ \$4 ○ \$5 ○ \$6 ○ \$8 ○ \$10

Click on your choice.

Original Item 3



$P(t)$ is the amount of yeast in a culture at time $t \geq 0$, and the growth rate of the yeast is given by

$$\frac{dP(t)}{dt} = k\, P(t)\left(1 - \frac{P(t)}{M}\right)$$

where k and M are positive constants.

Statements:

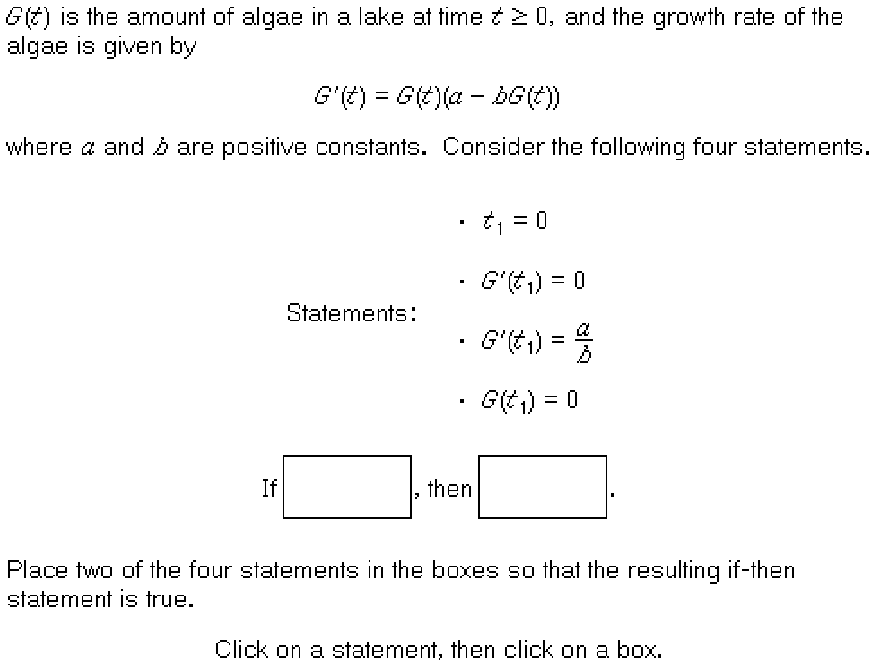
- $t = 0$
- $\frac{dP(t)}{dt} = 0$
- $\frac{dP(t)}{dt} = k$
- $P(t) = M$

If □, then □.

Place two of the four statements in the boxes so that the resulting if-then statement is true.

Click on a statement, then click on a box.

Item 3 Clone



$G(t)$ is the amount of algae in a lake at time $t \geq 0$, and the growth rate of the algae is given by

$$G'(t) = G(t)(a - bG(t))$$

where a and b are positive constants. Consider the following four statements.

Statements:

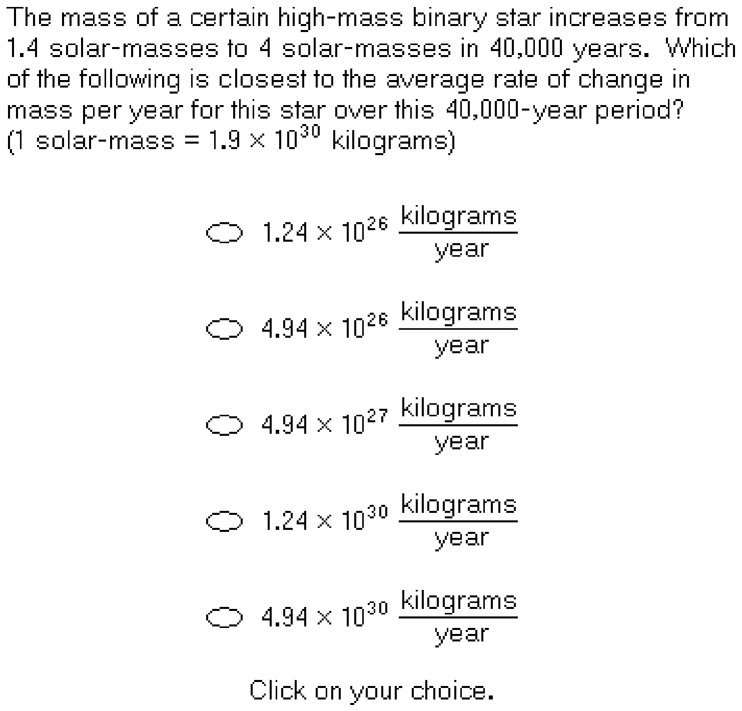
- $t_1 = 0$
- $G'(t_1) = 0$
- $G'(t_1) = \frac{a}{b}$
- $G(t_1) = 0$

If □, then □.

Place two of the four statements in the boxes so that the resulting if-then statement is true.

Click on a statement, then click on a box.

Original Item 4



The mass of a certain high-mass binary star increases from 1.4 solar-masses to 4 solar-masses in 40,000 years. Which of the following is closest to the average rate of change in mass per year for this star over this 40,000-year period? (1 solar-mass = 1.9×10^{30} kilograms)

○ $1.24 \times 10^{26} \dfrac{\text{kilograms}}{\text{year}}$

○ $4.94 \times 10^{26} \dfrac{\text{kilograms}}{\text{year}}$

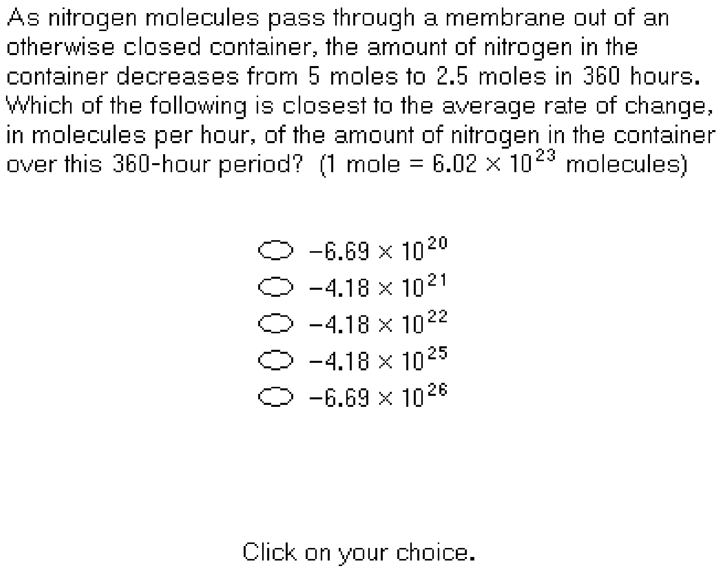
○ $4.94 \times 10^{27} \dfrac{\text{kilograms}}{\text{year}}$

○ $1.24 \times 10^{30} \dfrac{\text{kilograms}}{\text{year}}$

○ $4.94 \times 10^{30} \dfrac{\text{kilograms}}{\text{year}}$

Click on your choice.

Item 4 Clone

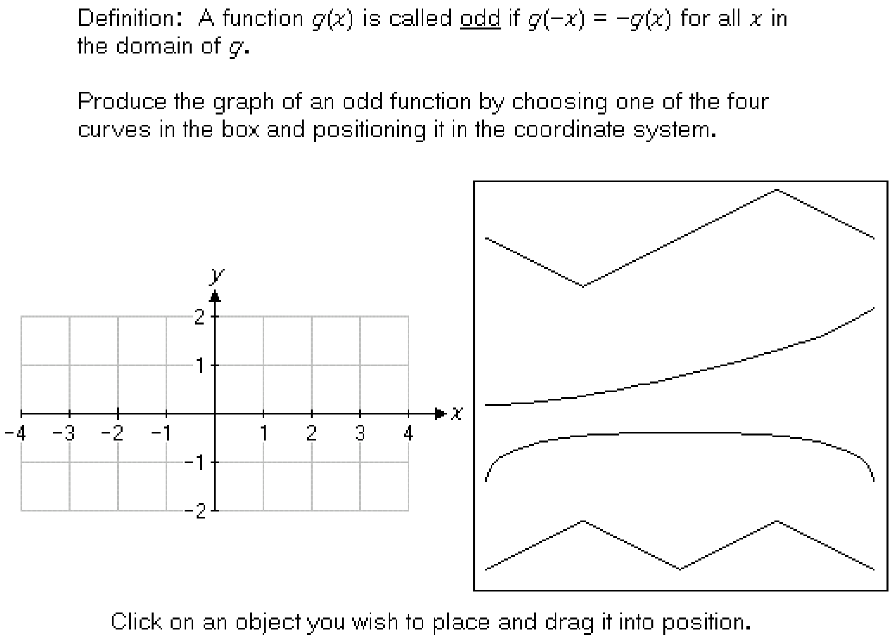


As nitrogen molecules pass through a membrane out of an otherwise closed container, the amount of nitrogen in the container decreases from 5 moles to 2.5 moles in 360 hours. Which of the following is closest to the average rate of change, in molecules per hour, of the amount of nitrogen in the container over this 360-hour period? (1 mole = 6.02×10^{23} molecules)

○ -6.69×10^{20}
○ -4.18×10^{21}
○ -4.18×10^{22}
○ -4.18×10^{25}
○ -6.69×10^{26}

Click on your choice.

Original Item 5

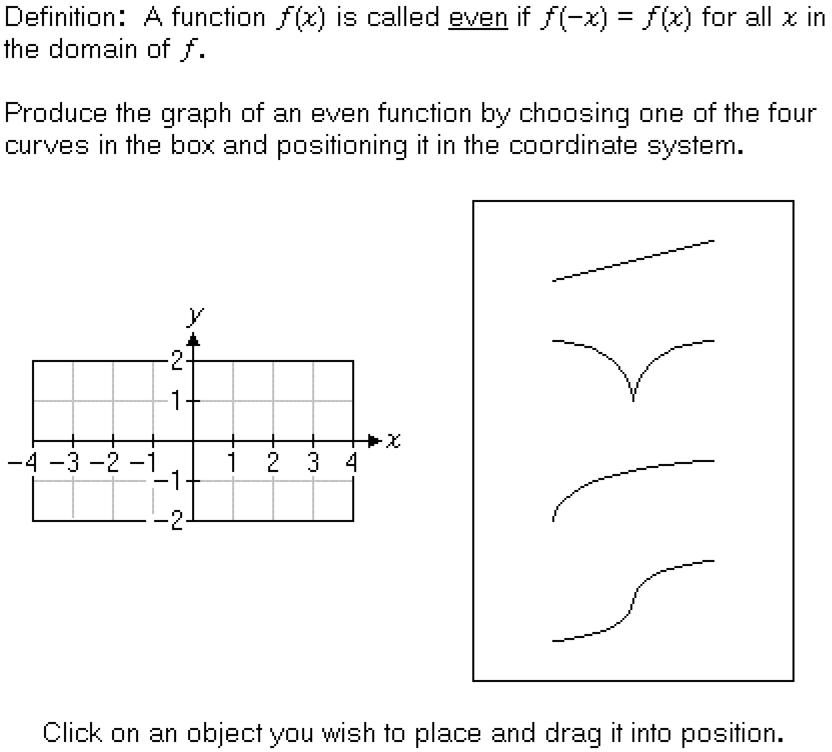


Definition: A function $g(x)$ is called odd if $g(-x) = -g(x)$ for all x in the domain of g.

Produce the graph of an odd function by choosing one of the four curves in the box and positioning it in the coordinate system.

Click on an object you wish to place and drag it into position.

Item 5 Clone

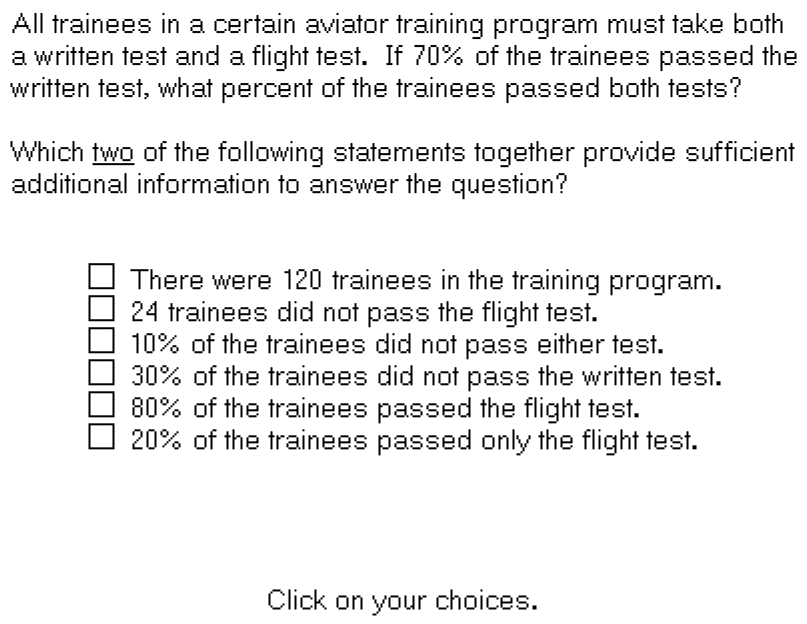


Definition: A function $f(x)$ is called even if $f(-x) = f(x)$ for all x in the domain of f.

Produce the graph of an even function by choosing one of the four curves in the box and positioning it in the coordinate system.

Click on an object you wish to place and drag it into position.

Original Item 6



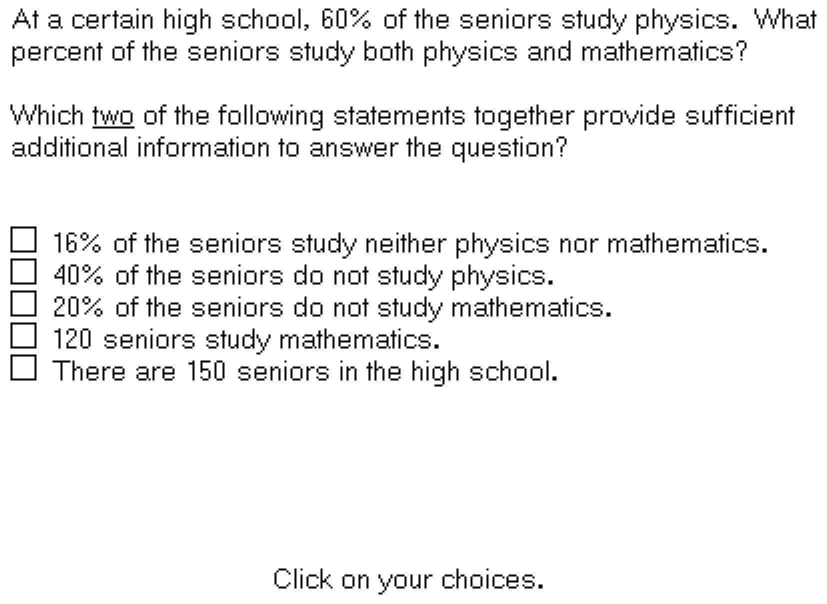
All trainees in a certain aviator training program must take both a written test and a flight test. If 70% of the trainees passed the written test, what percent of the trainees passed both tests?

Which two of the following statements together provide sufficient additional information to answer the question?

□ There were 120 trainees in the training program.
□ 24 trainees did not pass the flight test.
□ 10% of the trainees did not pass either test.
□ 30% of the trainees did not pass the written test.
□ 80% of the trainees passed the flight test.
□ 20% of the trainees passed only the flight test.

Click on your choices.

Item 6 Clone



At a certain high school, 60% of the seniors study physics. What percent of the seniors study both physics and mathematics?

Which two of the following statements together provide sufficient additional information to answer the question?

□ 16% of the seniors study neither physics nor mathematics.
□ 40% of the seniors do not study physics.
□ 20% of the seniors do not study mathematics.
□ 120 seniors study mathematics.
□ There are 150 seniors in the high school.

Click on your choices.

04041-013286 • Y102P.2 • Printed in U.S.A.
I.N. XXXXXX