ACT Research Report Series **87-16**

Gender Differences in Performance on Mathematics Achievement Items

Allen E. Doolittle

September 1987

For additional copies write:
ACT Research Report Series
P.O. Box 168
Iowa City, Iowa 52243

GENDER DIFFERENCES IN PERFORMANCE ON MATHEMATICS ACHIEVEMENT ITEMS

Allen E. Doolittle

ABSTRACT

Gender differences in performance on three types of mathematics test items were investigated using data from students with three different course backgrounds. Eight randomly equivalent samples of high school seniors were each given a unique form of the ACT Assessment Mathematics Usage Test. Only students with three specific profiles of high school mathematics coursework were considered in the analysis. The three background conditions ranged from little mathematics (Algebra I only) to a modest background (two Algebra courses and Geometry) to a full mathematics program including Beginning Calculus. For each background condition, examinee performance was analyzed in a 2 x 3 x 8 (gender by item category by test form) split plot factorial design. The results indicated that, at each of the studied background levels, females performed less well than males on geometry and strategy/reasoning items. On the other hand, females performed as well as males on algorithmic, operations-oriented items.

Gender Differences in Performance on Mathematics Achievement Items

In recent years, many investigators in educational and psychological measurement have given attention to a topic frequently referred to as item bias, but perhaps more precisely termed differential item performance (DIP). Differential item performance is observed if, given examinees of equal abilities in the characteristic being measured by a set of test items, the probability of answering an item correctly is related to group membership (Shepard, Camilli, & Averill, 1981; Petersen, 1980). Much of the attention has been focused on developing and evaluating procedures for the detection of DIP. Comparatively little work has been done in investigating relationships between characteristics of items and differential performance. The research reported here is of the latter type and focuses on the characteristics of mathematics achievement items on which male and female high school students seem to perform differently.

In the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1985), the responsibility of test developers to understand the role that item format and content may have in causing group differences in test scores is emphasized. Standard 3.10 states that "operational use of a test will often afford opportunities to check for group differences in test performance and to investigate whether or not these differences indicate bias." Conceivably, if bias is evident, such investigations could lead the test developer to institute revisions in the test items or specifications. However, even if bias is not indicated and the test seems to be functioning appropriately, such investigations can be useful for better understanding the nature of existing group differences in performance.

It is well known that male high school students as a group tend to perform better than female high school students on mathematics achievement tests (Armstrong, 1981; Clark & Grandy, 1984; Fennema & Carpenter, 1981). Benbow and Stanley (1980) suggest that these differences may be due in part to gender differences in spatial abilities. Another possible explanation is that male students typically have different experiences that may be relevant to the development of mathematics skills than do females. Fennema and Sherman (1977) argue that these differences are primarily due to differences in instruction--that males typically receive more and higher levels of mathematics instruction than do females. Differences in instructional background might also contribute to differential performance on mathematics items. For example, differential performance might be shown to exist for a higher level mathematics item if one group of students has been appropriately instructed in the relevant concepts and another group of students has not.

In a series of three studies (Doolittle, 1984, 1985; Doolittle & Cleary, 1987), the plausibility of a differential instruction interpretation of gender-based DIP in mathematics was investigated. In all three studies, a procedure suggested by Linn and Harnisch (1981) was used to detect differentially performing items for subgroups of examinees defined by various combinations of gender and high school mathematics background. Two notable observations were supported by these studies:

1. Gender-based DIP that is not clearly attributable to differences in instruction may exist in mathematics achievement items;
2. Differential item performance can be predicted based upon characteristics of the items and the sex of the examinees.

The primary focus of the present investigation was to expand upon the previous research by specifically controlling for background in mathematics.

The results of the previous studies are suggestive but unclear because of difficulty in assessing academic background. In the present research, the problem is reduced since students were categorized according to specific profiles of self-reported high school coursework. In addition, several background levels were studied to determine whether the same patterns of differential performance occur for students with different mathematics backgrounds. One background group consisted of students reporting an Algebra I course as their only high school mathematics course. At the other extreme, a group was comprised of students with a full program of mathematics, including Beginning Calculus. Somewhere in the middle was a group consisting of students reporting the equivalent of three courses: Algebra I, Algebra II, and Geometry. This course profile was chosen because it is the most common of all profiles among college-bound high school seniors.

A second focus of the research was to investigate specific item content as it relates to instructional background and gender. Multiple forms of the ACT Assessment Mathematics Usage Test (ACTM) were used to gather information on the relative performances of males and females on a large group of items classified into three categories. The results of the previous studies suggest that these content categories might be relevant to an understanding of gender-based differences in mathematics test performance. When mathematics background was controlled, an item category by gender interaction was expected. Geometry items and items such as word problems that emphasize reasoning skills were predicted to favor male examinees. On the other hand, algorithmic, calculation-oriented items were predicted to relatively favor females. Examination of these hypotheses was intended to contribute, in the spirit of Standard 3.10, to a greater understanding of the nature of differential performance in mathematics items as it relates to gender.

Methodology

The Instrument

The ACT Assessment Program contains educational achievement tests in four content areas, one of which is Mathematics Usage (ACTM). The ACTM is a 40-item, 50-minute measure of mathematics achievement. It emphasizes the solution of practical, quantitative problems that are encountered in many post-secondary programs and includes a sampling of mathematical techniques covered in high school courses. The test stresses quantitative reasoning rather than the memorization of formulas, knowledge of techniques, or computational skill. In general, the mathematical skills required for the test involve proficiencies emphasized in high school plane geometry and first- or second-year algebra. Each item in the test is a question followed by five alternative answers. Six categories of items, described in Table 1, are included in the test.

Item Classification

For the purposes of this study, the ACTM items were reclassified based on a theoretical framework developed by Mayer (1977, 1982) for describing the domain of mathematics problem solving. Mayer's formulation is of particular value for this research because it provides a useful structure for classifying mathematics problems. In particular, algorithmic knowledge was considered to relate to the solution of problems that emphasize computations and other well-defined operations; and strategic knowledge was considered to be required primarily in the solution of reasoning-focused items. Word problems are most likely to be placed in this category because they are widely considered to best represent thinking and understanding in mathematics learning (Nesher, 1986).

Although Mayer's theory does not clearly specify where geometry items should be included, most might plausibly be considered as items primarily measuring strategic knowledge. That is, the solution of geometry problems would seem to be more "strategic" than "algorithmic." However, since the solution of geometry problems is sometimes considered to draw upon spatial skills, and since differences in spatial skills are commonly discussed in the research literature on gender differences (Maccoby & Jacklin, 1974; Petersen, 1979), geometry items were classified in a category separate from other "strategic" items. In sum, ACTM items were classified into three categories:

1. Algorithmic;
2. Strategic, Non-Geometric; and
3. Strategic, Geometric.

A set of guidelines was prepared to assist in classifying the items. Each of the 40 items on each of the eight forms was independently classified by two raters. Whenever the raters could not agree on a classification, the item was withdrawn from consideration; only those items for which the raters were in complete agreement were included. Each form of the ACTM contained approximately 40% Algorithmic items, 35% Strategic, Non-Geometric items, and 20% Strategic, Geometric items. About 1-2 items per form (5%) were not included because of difficulty in classification.

Many of the Strategic, Non-Geometric items were previously classified by ACT as Arithmetic and Algebraic Reasoning items (Table 1, Category 2); most of the Strategic, Geometric items were classified by ACT as Geometry items (Table 1, Category 3); and the Algorithmic items came primarily from ACT's remaining categories (Table 1, Categories 1, 4, 5, & 6). Table 2 presents the precise number of items (out of 40) for each category and form that were retained for analysis. Because each form of the ACTM is constructed to precisely match a set of test specifications, the variability in the numbers of items in each category, shown in Table 2, simply reflects the differences between the operational classification scheme and the classification scheme used here.

Instructional Background

Since Fall 1985, as part of the registration process for the ACT Assessment, examinees have been asked to indicate whether or not they have taken courses in six areas of mathematics:

1. Algebra I (also Beginning Algebra, but not pre-Algebra or general mathematics);
2. Algebra II (also Advanced Algebra, but not a second year of Algebra I);
3. Geometry (includes Plane Geometry or Solid Geometry, but not Analytic Geometry);
4. Trigonometry;
5. Advanced Mathematics (includes Pre-Calculus, Analytic Geometry, Analysis, or Statistics, but not Trigonometry, Algebra, or computer mathematics);
6. Beginning Calculus.

Students are able to indicate background in any number of these courses or content areas. Since this data is student-reported and does not come from high school transcripts, it is not expected to be perfectly reliable. However, research at ACT has demonstrated that similar data is approximately 90%

accurate. In the present research, specific combinations of courses were used to match students on high school mathematics background.

Data

The data for this research were drawn from a sample of college-bound, high school seniors on a recent administration of the ACTM. Eight forms of the ACTM were administered to approximately 20,000 students in a spiraled fashion, thus creating eight samples, presumed to be randomly equivalent, of about 2,500 students apiece. Approximately 55% of the sampled students were female.

Each of the samples was further divided into subgroups based upon reported mathematics coursework in high school. Subgroups for three mathematics course-taking profiles were selected for further study in this research. Groups 1 (Algebra I only) and 3 (full math program) were selected to represent extremes in background. Group 2 was selected as the most typical profile reported by college-bound, high school students. The three profiles, with approximate percentages of students from the whole sample, are shown below.

1. Algebra I only (5.0%)
2. Algebra I, Algebra II, Geometry (24.6%)
3. Algebra I, Algebra II, Geometry, Trigonometry, Advanced Mathematics, Beginning Calculus (4.4%)

The numbers of male and female examinees given each form of the test are shown in Table 3. So that the analysis of the data could be readily interpreted, individual cell sample sizes were balanced by limiting all cells to the number in the smallest cell. Because the smallest cell was the number of males given Form D with an Algebra I-only background, all cell sizes were set to 35. Thus, 35 male and 35 female examinees were selected for each test form and each background condition. A random number generator was used to approximate a random sampling of the students. All together, data from 1,680 examinees were retained for analysis.

Design and Analysis

A split-plot factorial design, similar to that used by Schmeiser (1983), was used to investigate the effects of item category on gender differences in performance. The observed score for each examinee was the proportion correct of the items in each specific item category. Performance for a group was measured by mean proportion correct.

In this design, gender, and test form were considered between-group "treatments" and item category was a within-group "treatment." Three analyses, one for each background profile, were carried out following the same design.

For each background category, the three item categories were crossed with gender and the eight unique forms used as replications (Figure 1). The design includes 3 x 2 x 8 = 48 cells, for each background condition. Since a sampled examinee is either male or female and was given only one of the eight forms, examinees were nested within gender and form. Examinees and item category, on the other hand, were crossed. To illustrate, the responses of female examinees with an Algebra I only background, who also were given Form A, are shaded in Figure 1.

The model for the design is:

$$Y_{pgfc} = \mu + \alpha_g + \gamma_f + \alpha\gamma_{gf} + \pi_{p(gf)} + \psi_c + \alpha\psi_{gc}$$

$$+ \gamma\psi_{fc} + \alpha\gamma\psi_{gfc} + \psi\pi_{cp(gf)} + \varepsilon_{pgfc}$$

(Equation 1)

where:

Y_{pgfc} = proportion of items correct for person p of gender g on item category c for form f,

μ = overall population mean,

α_g = gender effect,

γ_f = form effect,

$\alpha\gamma_{gf}$ = interaction of gender and form,

$\pi_{p(gf)}$ = effect of persons, nested within gender, and form,

ψ_c = item category effect,

$\alpha\psi_{gc}$ = interaction of gender and item category,

$\gamma\psi_{fc}$ = interaction of form and item category,

$\alpha\gamma\psi_{gfc}$ = interaction of gender, form, and item category,

$\psi\pi_{cp(gf)}$ = interaction of item category and persons, nested within gender and form,

ε_{pgfc} = residual error.

Results

The results of the analysis of variance for each of the three background categories are presented in Tables 5, 6, and 7. The null hypothesis of principle interest in this study--that there is no interaction between gender and item classification--should be rejected for the two lower background groups. However, the results of the ANOVA presented in Table 7 (full math background students) are not sufficient to reject the null hypothesis for the gender by item category effect.

Mean performances of male and female examinees at each background level and for each item category, summarized across forms, are graphically presented in Figure 2. The nature of the gender by item category interaction is visually clear in this figure. Consistent with expectations, males and females performed similarly on the Algorithmic items, but females performed less well relative to males on the Strategic, Non-Geometric and the Strategic, Geometric items. Although the gender by category effect was not found to be statistically significant for the full background group (Table 7), mean performances for this group, shown in Figure 2, are consistent with those for the other background groups. Relative to males, females performed less well on Strategic, Geometric and Non-Geometric items than they did on Algorithmic items. Ceiling effects may have been partially responsible for mitigating the gender by item category interaction and the item category main effects for Background 3.

Also shown in Figure 2 are substantial performance differences between the students at each background category. Because there is an obvious confounding of the effects of instruction and student ability, little can be concluded about the sensitivity of the test to curriculum. However, the difference in student performance on geometry items between Background 1 (no Geometry) and Background 2 (includes Geometry) is noteworthy, as is the difference in performance on Algorithmic items from Background 2 to Background 3. This latter result might be attributed to improved performance on some of the more challenging, "algorithmic" algebra items following coursework in Advanced Mathematics and Introductory Calculus.

All three ANOVA summaries (Tables 5-7) were similar in showing a significant test form effect and a significant form by category interaction. The size and direction of these effects can be seen in part in Figure 3. For background categories 1 and 2, only the mean proportion correct for the total set of items is presented. For Background 3, however, means for each item category are presented for each form. The variation in the item category means, pictured for Background 3, is illustrative of the patterns that also occurred for background categories 1 and 2. These flip-flopping means are the source of the significant form by category interactions. The differences in the means for all studied items in each form are the cause of the significant form effect.

Both the significant form by category and the overall form effects were somewhat of a surprise, though perhaps they should not have been. The detailed test specifications used to construct the tests were based on a different classification scheme than that used for this analysis. In addition, the test items are all unique so the resulting forms can never be precisely parallel. It is to adjust for such differences in the test forms that the ACT Assessment and other standardized tests are statistically equated. Because the data analyzed here are based on unequated raw scores, these differences appear in the results.

Finally, it is noteworthy that a gender by form interaction was not found at any of the background levels. These results suggest that it is immaterial for female examinees which form of the test they take.

Discussion

Despite differences in methodology, the results of this study are consistent with previous research reported by the author (Doolittle, 1984, 1985; Doolittle & Cleary, 1987) and others (Becker, 1983; Donlon, 1973; Donlon, Hicks, and Wallmark, 1980; Marshall, 1984). There seem to be systematic differences between male and female examinees in their performance on mathematics achievement items. Relative to males, females perform less well on Strategic (both Geometric and Non-geometric) items than they do on Algorithmic items. A major outcome of this study is that the observed differences in performance for each item type were stable across ACTM forms, when examinees were matched by high school course background.

Although the differential performance between males and females is statistically significant and seems to be real, the practical significance of the differences needs to be evaluated. From Figure 2, it appears that mean differences of about .05 occur between instructionally matched males and females on the Strategic items (both Geometric and Non-geometric). Because approximately 22-23 Strategic items appear on a test form (see Table 2), the impact of these mean performance differences is about one raw score point, which converts to an approximate one point difference on ACT's standard score scale as well. Depending upon a student's overall performance relative to the standards used for making admissions or scholarship decisions, a one-point difference on the ACTM may or may not be considered significant.

However, mean performance differences of this magnitude should be of significance to test developers and educators. Test developers, for example, might choose to revise their specifications in light of known group differences in performance. This is not always an appropriate solution, though, because many standardized testing programs like the ACT Assessment have specifications that are closely tied to curriculum. As long as test items are reflective of the curriculum, they should not be removed simply because of observed group differences.

Figure 4 presents four items that were among those relatively more difficult for females than for males. In reviewing these items, it is not readily apparent why such group differences exist--but they do. The problem might very well have its source in student backgrounds. For example, there may be differences in student experiences, unaccounted for in this study, that partially explain differential performances on mathematics items. Or there may be gender differences, either learned or biological, in approaches to mathematics problem-solving. These thoughts are only speculation. The results of this study merely suggest that when students are matched on high school coursework, small but possibly consequential differences in the performances of male and female examinees do exist on the ACT Assessment Mathematics Usage Test.

REFERENCES

AERA, APA, & NCME (1985). Standards for educational and psychological testing. Washington, DC: American Psychological Association, Inc.

Armstrong, J.M. (1981). Achievement and participation of women in mathematics: results of two national surveys. Journal for Research in Mathematics Education, 12, 356-372.

Becker, B.J. (1983, April). Item characteristics and sex differences on the SAT-M for mathematically able youths. Paper presented at the annual meeting of American Educational Research Association, Montreal.

Benbow, C.P., & Stanley, J.C. (1980). Sex differences in mathematical ability: fact or artifact? Science, 210, 1262-1264.

Clark, M.J. & Grandy, J. (1984). Sex differences in the academic performance of Scholastic Aptitude Test takers (College Entrance Examination Board Report 84-8; ETS Research Bulletin 84-43). New York: College Entrance Examination Board.

Donlon, T.F. (1973). Content factors in sex differences on test questions. (ETS RM 73-28). Princeton, NJ: Educational Testing Service.

Donlon, T.F., Hicks, M.M., & Wallmark, M.M. (1980). Sex differences in item responses on the Graduate Record Examination. Applied Psychological Measurement, 4(1), 9-20.

Doolittle, A.E. (1984, April). Interpretation of differential item performance accompanied by gender differences in academic background. Paper presented at the annual meeting of the American Educational Research Association, New Orleans. (ERIC Document Reproduction Service No. ED 247 237.)

Doolittle, A.E. (1985, April). Understanding differential item performance as a consequence of gender differences in academic background. Paper presented at the annual meeting of the American Educational Research Association, Chicago. (ERIC Document Reproduction Service No. ED 263 218.)

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

Fennema, E., & Sherman, J. (1977). Sex-related differences in mathematics achievement, spatial visualization and affective factors. American Educational Research Journal, 14, 51-71.

Linn, R.L., & Harnisch, D.L. (1981). Interactions between item content and group membership on achievement test items. Journal of Educational Measurement, 13, 109-118.

Maccoby, E., & Jacklin, C. (1974). Psychology of sex differences. Palo Alto, CA: Stanford University Press.

Marshall, S.P. (1984). Sex differences in children's mathematics achievement: Solving computations and story problems. Journal of Educational Psychology, 76(2), 194-204.

Mayer, R.E. (1977). Thinking and problem solving. Glenview, IL: Scott, Foresman & Co.

Mayer, R.E. (1982). The psychology of mathematical problem solving. In F.K. Lester & J. Garofalo (Eds.), Mathematical problem solving: Issues in research. Philadelphia: The Franklin Institute Press.

Nesher, P. (1986). Learning mathematics: A cognitive perspective. The American Psychologist, 41(10), 1114-1122.

Petersen, A.C. (1979). Hormones and cognitive functioning in normal development. In M.A. Wittig & A.C. Petersen (Eds.), Sex-related differences in cognitive functioning: Developmental Issues. New York: Academic Press.

Petersen, N.S. (1980). Bias in the selection rule -- bias in the test. In L.J. Th. van der Kamp, W.F. Langerak, & D.N.M. de Gruijter (Eds.), Psychometrics for educational debates. John Wiley & Sons, Ltd.

Schmeiser, C.B. (1983). Doctoral dissertation, The University of Iowa.

Shepard, L.A., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. Journal of Educational Statistics, 6, 317-375.

TABLE 1

ACTM Item Categories

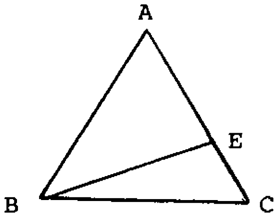
Description	Example
1. Arithmetic and Algebraic Operations (AAO). The four items in this category explicitly describe operations to be performed by the student: manipulating and simplifying expressions containing arithmetic or algebraic fractions, performing basic operations in polynomials, solving linear equations in one unknown, and performing operations on signed numbers.	• $(2^4)^3 = ?$ A. $2^{\frac{4}{3}}$ B. 2^7 * C. 2^{12} D. 2^{43} E. 2^{64}
2. Arithmetic and Algebraic Reasoning (AAR). The fourteen word problems in this category present practical situations in which algebraic and/or arithmetic reasoning is required. The problems require the student to interpret the question and to either solve the problem or find an approach to its solution.	• If 8 French francs were worth 1 U.S. dollar, and 2 U.S. dollars were worth 1 British pound, then 16 British pounds would be worth how many French francs? * A. 256 B. 128 C. 64 D. 32 E. 4
3. Geometry (G). The items in this category cover such topics as measurement of lines and plane surfaces, properties of polygons, the Pythagorean theorem, and relationships involving circles. Both formal and applied problems are included. Each form of the ACTM includes eight G items.	• In the figure below, $\overline{AB}$ and $\overline{AC}$ have the same length, and E lies on $\overline{AC}$. If the measure of ∠ABC is 54° and the measure of ∠BEC is 103°, what is the measure of ∠EBC?  A. 18° * B. 23° C. 27° D. 36° E. 49°

TABLE 1--continued

ACTM Item Categories

Description	Example
4. Intermediate Algebra (IA). The eight items in this category include such topics as dependence and variation of quantities related by specific formulas, arithmetic and geometric series, simultaneous equations, inequalities, exponents, radicals, graphs of equations, and quadratic equations.	• What value of $\mathbf{y}$ satisfies the system of equations below? $2x + 3y = 5$ $x - 2y = 6$ A. -11 * B. - 1 C. 1 D. 2 E. 7
5. Number and Numeration Concepts (NNS). The four items in this category cover such topics as rational and irrational numbers, set priorities and operations, scientific notation, prime and composite numbers, numeration systems with bases other than 10, and absolute value.	• For all positive real numbers $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ with $\mathbf{a} = \mathbf{b} + \mathbf{c}$, which of the following inequalities is ALWAYS true? A. $\mathbf{a} < \mathbf{b}$ B. $\mathbf{b} < \mathbf{c}$ * C. $\mathbf{c} < \mathbf{a}$ D. $\mathbf{ab} < \mathbf{ac}$ E. $\mathbf{a} + \mathbf{b} < \mathbf{a} + \mathbf{c}$
6. Advanced Topics (AT). The items in this category cover such topics as trigonometric functions, permutations and combinations, probability, statistics, and logic. Only simple applications of the skills implied by these topics are tested. Each form of the ACTM includes two AT items.	• A 6-sided die with sides numbered 1 to 6 is tossed at the same time that a fair coin is flipped. A typical outcome is (5,H)--a 5 on the die and a head on the coin. How many different outcomes are possible? A. 8 * B. 12 C. 32 D. 36 E. 64

TABLE 2

Number of Items in Each Category
for Each Form

Item Category	Test Form A	B	C	D	E	F	G	H
Algorithmic	15	15	20	18	17	18	15	12
Strategic, Non-geometric	18	14	10	13	14	12	16	16
Strategic, Geometric	6	9	9	8	9	10	8	9
not classified	1	2	1	1	0	0	1	3
Total items	40	40	40	40	40	40	40	40

TABLE 3

Number of Examinees by Course Background

	Test Form							
Course Background	A	B	C	D	E	F	G	H
A1								
Males	48	42	53	35	41	45	41	43
Females	82	99	83	77	87	82	80	84
A1, A2, G								
Males	223	233	250	237	236	231	247	215
Females	387	419	394	371	389	378	396	416
A1, A2, G, T, AM, BC								
Males	67	63	51	61	54	49	78	58
Females	54	50	58	46	51	58	53	57

A1: Algebra I

A2: Algebra II

G: Geometry

T: Trigonometry

AM: Advanced Mathematics

BC: Beginning Calculus

TABLE 4

Mean ACTM (Scaled Score) Performance by Course Background

	Test Form							
Course Background	A	B	C	D	E	F	G	H
A1								
Males	10.7	7.9	9.9	9.4	9.2	7.6	8.5	10.7
Females	8.7	7.9	7.0	7.8	7.0	6.6	6.6	8.9
A1, A2, G								
Males	15.6	16.5	16.3	16.0	15.9	15.5	16.2	16.4
Females	14.2	15.0	14.3	14.8	15.2	14.8	14.5	15.8
A1, A2, G, T, AM, BC								
Males	25.7	27.0	27.2	26.8	26.5	27.9	27.0	25.2
Females	24.2	24.3	26.0	24.9	25.2	25.5	24.7	26.0

TABLE 5

Analysis of Variance Summary Table:
Background Category 1 (Algebra 1 Only)

Source	df	MS	F	F prob.
Gender	1	0.3518	14.36	0.007
Form	7	0.0872	2.23	0.030
Gender x Form	7	0.0245	0.63	0.734
Persons Within Form x Gender	544	0.0391	---	---
Item Category	2	2.0882	17.04	0.000
Gender x Category	2	0.0890	4.72	0.027
Form x Category	14	0.1225	6.01	0.000
Gender x Form x Category	14	0.0189	0.92	0.532
Persons x Category Within Form x Gender	1088	0.0204	---	---

TABLE 6

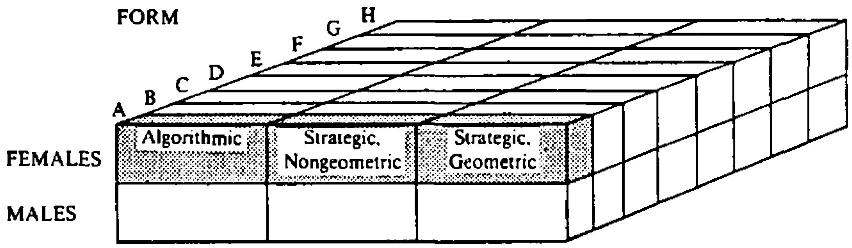
Analysis of Variance Summary Table:
Background Category 2 (A1, A2, Geometry)

Source	df	MS	F	F prob.
Gender	1	0.2841	9.12	0.019
Form	7	0.1523	2.47	0.017
Gender x Form	7	0.0312	0.51	0.830
Persons Within Form x Gender	544	0.0615	---	---
Item Category	2	0.5836	3.72	0.051
Gender x Category	2	0.3225	13.69	0.001
Form x Category	14	0.1569	7.07	0.000
Gender x Form x Category	14	0.0236	1.06	0.389
Persons x Category Within Form x Gender	1088	0.0222	---	---

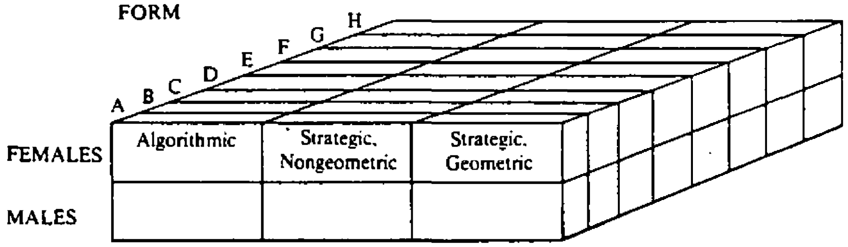
TABLE 7

Analysis of Variance Summary Table:
Background Category 3 (Full Math Program)

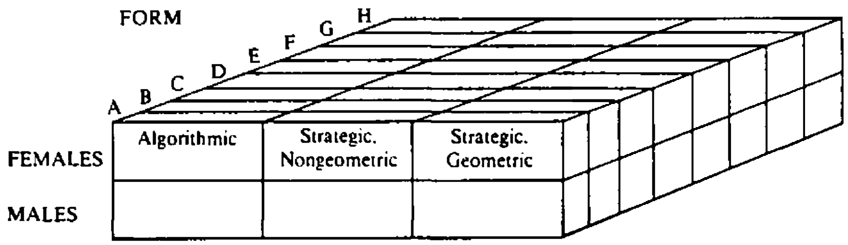
Source	df	MS	F	F prob.
Gender	1	1.2460	10.33	0.015
Form	7	0.1577	2.15	0.037
Gender x Form	7	0.1207	1.64	0.120
Persons Within Form x Gender	544	0.0734	---	---
Item Category	2	0.1053	1.76	0.208
Gender x Category	2	0.0143	1.07	0.370
Form x Category	14	0.0597	4.27	0.000
Gender x Form x Category	14	0.0134	0.96	0.493
Persons x Category Within Form x Gender	1088	0.0140	---	---



Background 1 (Algebra I only)



Background 2 (Algebra I, Algebra II, Geometry)



Background 3 (Algebra I, Algebra II, Geometry, Trigonometry, Advanced Mathematics, Intro. Calculus)

Figure 1. Pictorial Representation of the Design

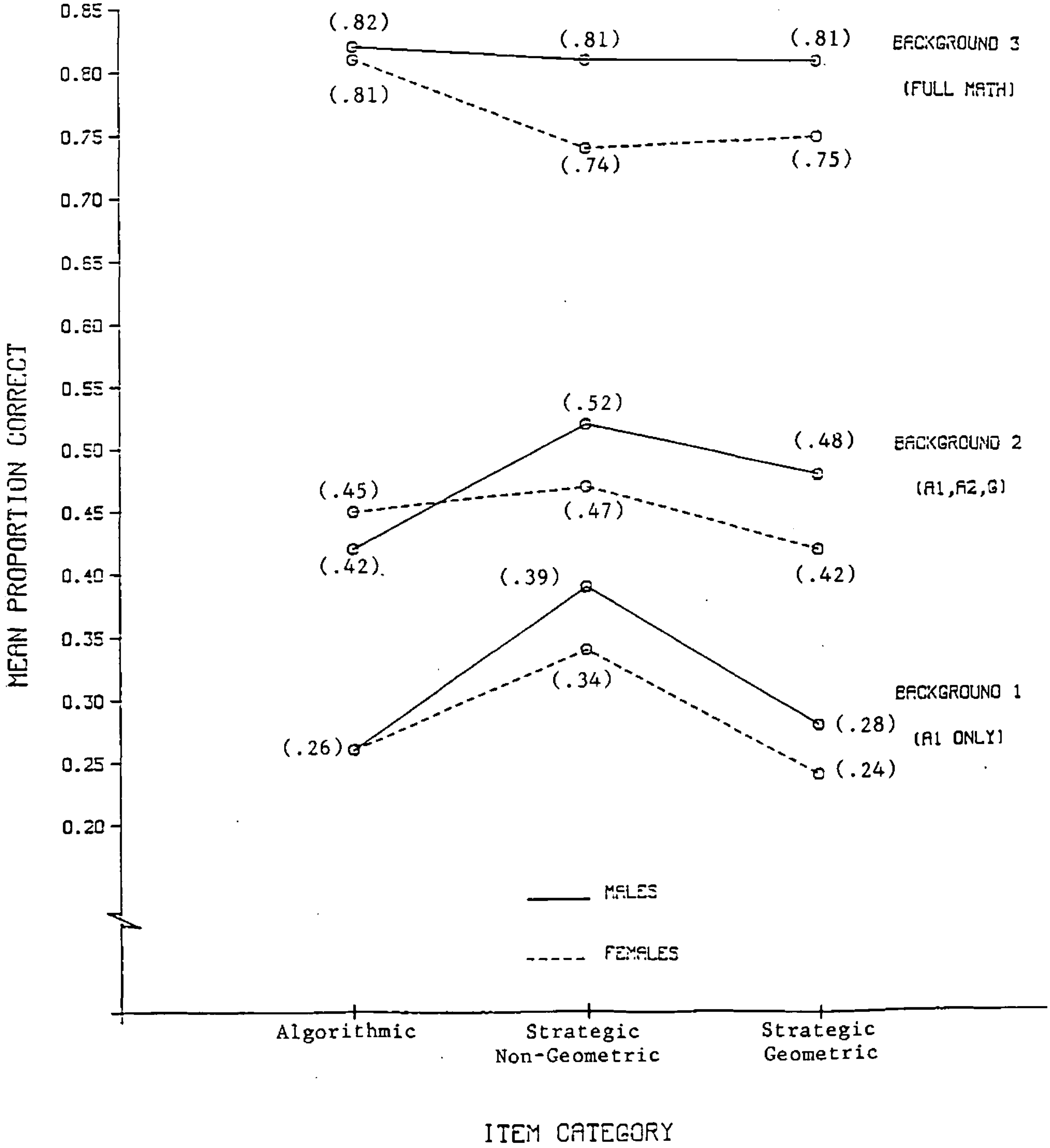


Figure 2. Gender x Item Category Effects for Each Background Level

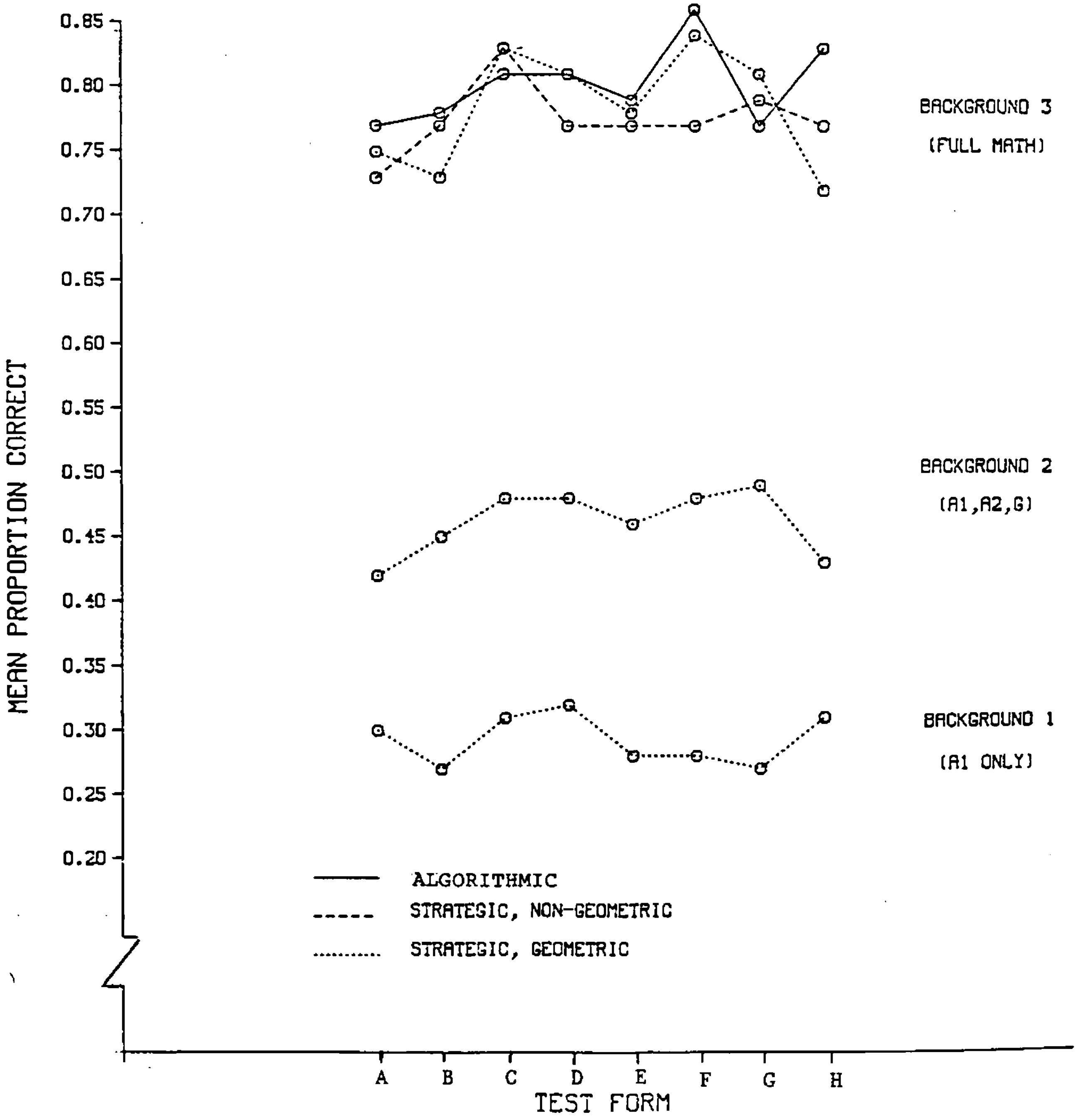


Figure 3. Mean proportion correct by form for each background level. Means shown by item category for Background 3

Strategic, Non-Geometric

1. An omelet made with 2 eggs and 30 grams of cheese contains 280 calories. An omelet made with 3 eggs and 10 grams of cheese contains the same number of calories. How many calories are in an egg?

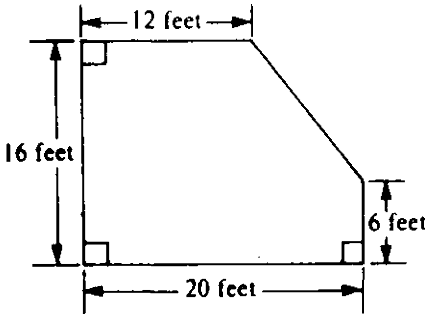
   A. 27
   B. 50
   * C. 80
   D. 102
   E. 160

2. A pair of slacks has a regular price of $32. If the slacks are on sale at 15% off the regular price and a sales tax of 5% of the sale price is added, what is total cost (tax included) of the slacks?

   A. $28.80
   * B. $28.56
   C. $25.84
   D. $25.70
   E. $25.60

Strategic, Geometric

3. What would be the area, in square feet, of a room with the measurements indicated in the figure below?



   A. 392
   B. 336
   C. 312
   * D. 280
   E. 240

Figure 4. Examples of Items That are Relatively More Difficult for Female Than for Male Examinees With Comparable Mathematics Backgrounds