ACT Research Report Series 89-9

Gender Differences in Performance on a College-level Achievement Test

Allen Doolittle
Catherine Welch

November 1989

ACT

Gender Differences in Performance on a College-Level Achievement Test

Allen Doolittle
Catherine Welch

ABSTRACT

Gender differences in performance on five tests of the Collegiate Assessment of Academic Proficiency (CAAP) were investigated. Random samples of male and female examinees were drawn from the first pilot administration of CAAP to form the basis for this study. Total test summary statistics and differential item performance methodology were used to detect gender-based performance differences. Differences were found on the Mathematics Test (favoring males), and on the multiple-choice Writing Skills Test (favoring females) as well as the essay-based Writing Test (favoring females). Although no overall performance differences were found between males and females on the Reading and Critical Thinking tests, there were notable differences associated with specific types of content within those tests. These results were viewed as consistent with similar research on different testing programs.

Gender Differences in Performance on a College-Level Achievement Test

Concern about equity with respect to men and women has generated considerable interest in educational achievement. Differences in the educational backgrounds and achievement of the two groups are likely to contribute to disparities in the allocation of cognitively demanding roles in our society. Consequently, group differences in relevant test scores are cause for concern. The focus of this study is on the measurement of gender differences in achievement test performance at the college level.

Differences in performance patterns on standardized test batteries have frequently been found for males and females. Stanley and his colleagues (Brody, 1987; Dauber, 1987; Lupkowski, 1987; Stanley, 1987) investigated gender differences on some 82 nationally standardized tests. To measure the size of differences in mean scores, they used Cohen's (1977) concept of effect size (mean score differences in standard units). Fairly large effect sizes (.50 to .90) were found for aptitude tests and for advanced achievement tests such as the advanced tests of the Graduate Record Examinations. Effect sizes were smaller for other standardized achievement tests, including college admissions tests. Recent ACT assessment data (ACT, 1988) yielded an effect size of .23 in English Usage favoring females and effect sizes of .22 (Social Studies Reading), .33 (Mathematics), and .38 (Natural Sciences Reading) favoring males.

Gender differences found on the ACT are generally consistent with those found for the SAT. A possible inconsistency is that females do better than males on the ACT English Usage Test, but that males do better than females on the SAT-Verbal (Clark & Grandy, 1984). However, the SAT-Verbal includes some

scientific and technical reading items on which females do substantially less well than males (Wendler & Carlton, 1987). This effect is consistent with performance differences favoring males on the ACT Natural Sciences Reading Test.

ACT has recently developed the Collegiate Assessment of Academic Proficiency (CAAP) as a new achievement test battery for use in higher education. In Fall 1988, CAAP was pilot-tested on a national sample of college students. This research was done as part of the initial analysis of that CAAP data and had as its focus the investigation of gender differences in test performance.

Methodology

The Instrument

CAAP has been developed as a test battery with components directed toward the measurement of academic skills typically attained in the first two years of college. The various tests in the CAAP battery are each 40 minutes in length and can be used independently or in any configuration. No overall composite score is offered.

In Fall 1988, CAAP included four objective tests--Reading, Writing Skills, Mathematics, and Critical Thinking--and a direct measure of writing proficiency.

The ***Reading Test*** measures student achievement in reading comprehension using questions based on reading selections in prose fiction, the humanities, the social sciences, and the natural sciences. Each form of the 36-item test contains four reading passages that are representative of the kinds of texts commonly encountered in college and university curricula.

Each passage is accompanied by a set of multiple-choice questions that require students to derive meaning, manipulate information, cite comparisons, make generalizations, and draw conclusions. The test focuses on a complex set of skills that students must use in comprehending written materials from a range of subject areas and purposes.

The 72-item *Writing Skills Test* is an indirect measure of writing skill. The test requires examinees to analyze prose similar to that found in a typical course of college study. Several prose passages are included, each of which is accompanied by a sequence of multiple-choice test items measuring understanding of the conventions of standard written English and rhetorical skills such as strategy, organization, and style. To provide a variety of rhetorical situations, a range of discourse is employed.

The 35-item *Mathematics Test* measures the achievement of mathematical skills generally taught in first- or second-year college mathematics courses. It emphasizes the solution of quantitative problems that are encountered in many postsecondary algebra courses and also includes some trigonometry and introductory calculus. The test emphasizes quantitative reasoning rather than memorization of formulas, knowledge of techniques, or computational skills.

The *Critical Thinking Test* measures the ability to clarify, analyze, evaluate, and extend arguments. The test consists of 32 items related to three passages that are representative of the kinds of issues commonly encountered in a postsecondary curriculum. Each passage presents one or more arguments and may use one of a variety of formats, including case studies, debates, dialogues, overlapping positions, statistical arguments, experimental results, and editorials.

The *Writing (Essay) Test* constitutes a direct approach to the measurement of writing. Each form of the test consists of two independent writing prompts. The two prompts involve different issues and audiences, but each requires the examinee to formulate a clear thesis; support the thesis with an argument or reasons relevant to the issue, position taken, and audience; and present the argument in a well-organized, logical manner.

As initially administered, each examinee received two scores per prompt. A "purpose" score reflected how well the examinees responded to the task required by the situations described in the prompts; and a "language usage" score reflected the raters' impressions of the relative presence of usage or mechanical errors and the degree to which such errors impeded the flow of thought in the essays. Each paper was scored separately on a 4-point scale for purpose and language usage by each of two raters working independently. The evaluations of both raters were averaged to obtain the purpose and language usage scores for each prompt. Additionally, the scores for the two prompts were averaged to yield a composite purpose score and a composite language usage score on a scale of 1.0 to 4.0.

Data Source

CAAP was pilot-tested on a national sample of students from about 100 postsecondary institutions. The sample included a variety of institutions, two- and four-year, public and private. The sample was not, however, designed to be nationally representative. The involved institutions were simply a sample of those interested in the CAAP program and able to begin a new testing program in Fall 1988. The students tested at these institutions were primarily incoming college freshmen, and consequently the test was fairly difficult for many of them.

Random samples of 1,000 males and 1,000 females for each objective test were drawn for analysis. Because the total number of examinees given the Writing (Essay) Test was not large, all of these students were included in the essay analyses.

Analyses

Mean performance for males and females was compared on all CAAP tests and for each available essay score. To investigate possible passage effects, mean performances were also compared for each passage-related set of items in the Reading and Critical Thinking tests. T-tests were run and effect sizes were calculated to assist in evaluating group differences in performance.

Finally, the Mantel-Haenszel (M-H) procedure (Holland & Thayer, 1986) was used at the individual item level of the objective tests to measure gender-based differential item performance. The intent of these analyses was to identify categories of items that seemed to be operating differently for the two groups.

Results

Table 1 presents the means, standard deviations, t-statistics, and effect sizes found for each test. These results indicate that females tended to perform better than males on the multiple choice Writing Skills Test and on the essay test; males tended to outperform females on the Mathematics Test.

Insert Table 1 About Here

Although no overall performance differences were found between males and females on the Reading and Critical Thinking tests, there were notable differences associated with individual passages. Females performed relatively

well on the items associated with Reading Passage 2 (art topic); and males performed relatively well on the items associated with Reading Passage 1 (scientific context) and Critical Thinking Passage 2 (scientific context).

In terms of the magnitude of performance differences on the tests, effect sizes were generally small. The Mathematics (favoring males) and the Writing Skills (favoring females) effect sizes were .20 and -.28, respectively. The effect sizes for the Writing (Essay) composite scores were larger: -.41 for the purpose score and -.32 for the language usage score, both scores favoring females.

Mantel-Haenszel procedures were used, but not in the typical sense of identifying individual items for differential performance. This use of differential item performance methodology was exploratory in nature, intended to look for categories of items that favored either males or females. A summary of these exploratory analyses is presented in Table 2. In these analyses, a very relaxed criterion (± 0.2 on the M-H delta) was used to identify items that seemed to perform differently for the two groups. The number of items in each category that seemed to favor males or females are shown.

Insert Table 2 About Here

Generally the results in Table 2 portray seemingly random distributions of items favoring males or females in the various subcategories. However, the disproportionate numbers of items favoring males in Passage 1 of the Reading Test and Passage 2 of the Critical Thinking Test are consistent with the subscore results presented in Table 1, showing a relative advantage for males

on science-oriented items. Also, the pattern of items favoring females in Passage 2 of the Reading Test, an arts-oriented passage, is consistent with Table 1. Results for several other item categories were less strong, but still suggestive:

- Writing Skills "grammar" items -- 4 items favoring females to 1 item favoring males;
- Writing Skills "sentence structure" items -- 8 to 3 favoring females;
- Writing Skills "organization" items -- 6 to 1 favoring males;
- Mathematics "applications" items -- 6 to 3 favoring males.

Discussion

The outcomes of this study with the CAAP tests are generally consistent with results found for other tests and examinee populations. The effect size of .20 found for the CAAP Mathematics Test was smaller than that found in other research with different programs, but still consistent with them in showing higher scores for males. Although the patterns of differential performance for several of the mathematics categories in Table 2 seem consistent with previous research (Doolittle & Cleary, 1987; Doolittle, in press; Marshall, 1984), additional research would be necessary to substantiate these relationships.

Effect sizes favoring females for the writing instruments (multiple choice, -.28; essay, -.41 and -.32) were also generally consistent with previous findings. However the differences for the essay scores were somewhat larger than expected, based on the results with the multiple-choice Writing Skills Test.

Finally the results for the Reading and Critical Thinking tests are interesting in that notable gender differences are found for items associated with specific passages, but not for the overall tests. Clearly, because of the limited number of passages examined here, further research with additional test forms would seem to be necessary. However, it appears that, consistent with Wendler and Carlton (1987), females may do better than males with items based on humanities-oriented reading passages, but poorer than males on items associated with science-oriented passages. (It is important to note that the test items do not directly measure knowledge of the content of associated passages, but rather reading, understanding, or reasoning within context.)

It appears that the performance differences between males and females found with CAAP are similar to those found with other achievement tests and populations. Clearly, when mean differences are usually less than half a standard deviation apart, there is considerable overlap in score distributions. However, these seem to be stable, group-level differences that are observed in many testing situations.

Differential background, interests, and even demographic factors related to male and female examinee groups, may be relevant for an accurate interpretation of group differences in test performance. But to the extent that the differences are real -- on content that is a significant part of the domain of interest -- they must be viewed as reflections of the differential achievement of students.

This research has focused on identifying item types and categories that perform differently for males and females. The challenge of future research will be in attempting to offer explanations for differential performance. For example, is there a theoretical framework that can be used in explaining gender differences in item performance? Research of this type will be important before appropriate interventions can be identified and implemented.

References

American College Testing Program (1988). ACT Assessment Program Technical Manual. Iowa City, IA: Author

Brody, L. E. (1987, April). Sex differences on the GPE, MCAT, LSAT, and GMAT. Paper presented at the AERA Annual Meeting, Washington, D.C.

Clark, M. J., & Grandy, J. (1984). Sex differences in the academic performance of Scholastic Aptitude Test takers. Report No. 84. New York: College Entrance Examination Board.

Cohen, J. (1977). Statistical power analysis for the behavioral sciences.. New York: Academic Press.

Dauber, S. L. (1987, April). Sex differences on the SAT-M, SAT-V, TSWE, and ACT among college-bound high school students. Paper presented at the AERA Annual Meeting, Washington, D.C.

Doolittle, A. E. (in press). Gender differences in performance on mathematics achievement items. Applied Measurement in Education.

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

Holland, P. W., & Thayer, D. T. (1986, April). Differential item performance and the Mantel-Haenszel procedure. Paper presented at the AERA annual meeting, San Francisco.

Lupkowski, A. E. (1987, April). Sex differences on the DAT. Paper presented at the AERA annual meeting, Washington, D.C.

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

Stanley, J. C. (1987, April). Sex differences on the College Board Achievement Tests and Advanced Placement Examinations. Paper presented at the AERA annual meeting, Washington, D.C.

Wendler, C. L. W., & Carlton, S. T. (1987, April). An examination of SAT-verbal items for differential performance by women and men: An exploratory study. Paper presented at the AERA annual meeting, Washington, D.C.

TABLE 1
Mean Comparisons of Male and Female Examinees

	Males			Females					
Test/Score	N	Mean	S.D.	N	Mean	S.D.	*t*	Prob.	Effect Size
Reading	1000	20.42	6.52	1000	20.46	6.12	- .14	.889	-.01
Passage 1		5.62	1.92		5.24	1.83	4.44	.000	.20
Passage 2		5.50	2.08		5.76	1.93	-2.92	.004	-.13
Passage 3		4.71	2.21		4.77	2.15	- .58	.559	-.03
Passage 4		4.59	2.74		4.68	2.63	- .77	.444	-.03
Writing Skills	1000	43.39	13.65	1000	47.09	12.43	-6.31	.000	-.28
Mathematics	1000	16.18	4.41	1000	15.34	3.82	4.58	.000	.20
Critical Thinking	1000	19.29	5.15	1000	18.92	5.07	1.60	.110	.07
Passage 1		7.32	2.00		7.15	1.94	1.94	.050	.09
Passage 2		6.41	2.32		6.05	2.33	3.45	.001	.15
Passage 3		5.57	2.26		5.71	2.15	-1.68	.094	-.06
Writing (Essay)	1490			2282					
Prompt 1 Purpose		2.50	.79		2.79	.79	-11.11	.000	-.36
Prompt 1 Lang. Usage		2.62	.67		2.82	.66	-8.98	.000	-.30
Prompt 2 Purpose		2.16	.81		2.42	.85	-9.27	.000	-.31
Prompt 2 Lang. Usage		2.60	.68		2.81	.66	-9.10	.000	-.31
Purpose (Composite)		2.33	.66		2.61	.68	-12.33	.000	-.41
Lang. Usage (Composite)		2.61	.62		2.81	.61	-9.84	.000	-.32

Table 2
Differential Item Performance (by Favored Group) for Item/Passage Categories[1]

Test	Subcategory	Total Items	Number Favoring Males	Number Favoring Females
Reading	Referring	8	2	4
	Reasoning	28	10	9
	Passage 1	9	7	0
	Passage 2	9	1	5
	Passage 3	9	3	3
	Passage 4	9	1	5
Writing Skills	Grammar	8	1	4
	Sentence Structure	18	3	8
	Organization	10	6	1
	Style	14	6	3
	Strategy	16	5	5
	Punctuation	6	2	1
	Passage 1	12	4	3
	Passage 2	12	5	4
	Passage 3	12	2	4
	Passage 4	12	4	2
	Passage 5	12	3	4
	Passage 6	12	5	5
Mathematics	Pre-Algebra	7	4	2
	Algebra	20	5	7
	Trig./Calculus	8	2	1
	Basic Skills	24	5	7
	Applications	11	6	3
Critical Thinking	Analysis	16	4	5
	Evaluation	17	3	1
	Extension	9	3	3
	Passage 1	11	3	2
	Passage 2	11	6	2
	Passage 3	10	1	5

[1]An extremely loose criterion on the Mantel-Haenszel delta statistic (Holland & Thayer, 1986) was used to identify items performing differently for males and females. Although this criterion was far too loose for evaluating individual items, it was used for exploratory purposes in evaluating trends in the various subcategories of items. This criterion flagged about 65% of the test items as favoring one group or the other.