ACT Research Report Series 89-6

Gender-Based Differential Item Performance in English Usage Items

Catherine J. Welch
Allen E. Doolittle

August 1989

ACT

GENDER-BASED DIFFERENTIAL ITEM PERFORMANCE IN ENGLISH USAGE ITEMS

Catherine J. Welch
Allen E. Doolittle
The American College Testing Program

ABSTRACT

Gender differences in performance on the ACT English Usage test were investigated. Random samples of male and female examinees were drawn from the October 1985 administration. Total test statistics and differential item performance statistics were used to detect gender-based performance differences. Results showed an overall tendency for female examinees to out-perform male examinees on the ACT English Usage test. The results did not suggest the existence of gender-based differential item performance in English Usage achievement items.

Gender-Based Differential Item Performance in English Usage Items

This study is designed to investigate relationships between characteristics of English usage achievement items and gender differences in performance. The performance of females is traditionally expected to be superior to that of males on verbal-based items (Huntley & Plake, 1980). Female high school students as a group perform slightly better than male high school students on English usage achievement tests (Green, 1987). Consistent with this observation, female high school students score about one-fifth of a standard deviation higher than male high school students on the ACT Assessment English Usage Test. NAEP investigations of the reading and writing skills of 17-year-old men and women in high school indicate that women have significantly out-performed men on these tests during the past 15 years. A possible explanation for these differences is that quantity or type of instructional background might affect performance. If this is true, instances of differential item performance (DIP) in the form of an instructional effect might exist in tests of English usage achievement.

Given examinees of equal abilities on the characteristic being measured by a set of items, the probability of answering an item correctly should not be related to group membership. Differential item performance is observed if group membership is related to performance (Petersen, 1980; Shepard, Camilli, & Averill, 1981). DIP could occur if the content of an item is less appropriate for one group of examinees than for another. A common way to examine the differences between two groups is to compare group performance on individual items among examinees obtaining comparable scores. (Holland, 1986; Green, 1987).

Recent research with high school examinees has indicated that "algorithmic" mathematics items tend to favor female examinees and reasoning-based items tend to favor males (Doolittle & Cleary, 1987) . Consistent with this research, it was expected that English usage items that seem algorithmic in nature, such as grammar, punctuation and sentence structure, would favor females; and that logic and organization, and diction and style items--usage items that seem to involve more reasoning skills--would relatively favor males.

Methodology

The Instrument

The ACT Assessment program contains four educational achievement tests, one of which is English Usage (ACTE). The ACTE is a 75-item, 40-minute test that measures the students' understanding of conventions of standard written English in punctuation, grammar, sentence structure, diction and style, and logic and organization. The test does not measure the rote recall of rules of grammar, but stresses the analysis of the kind of effective expository writing that will be encountered in many postsecondary curricula. The test consists of several prose passages with certain portions underlined and numbered. For each underlined portion, four alternative responses, including "NO CHANGE," are given. The student must decide which alternative is most appropriate in the context of the passage. Five types of ACTE items are included in the test and are described in Table 1.

Insert Table 1

Data Source

The data for this study were drawn from a sample of college-bound high school seniors from the October 1985 administration of ACTE. Seven forms of

the ACTE were administered to students in a spiraled fashion, thus creating seven randomly equivalent samples of students with each sample taking a different form of ACTE. Only students with a background in certain English courses were considered (see below). The final data sets were seven randomly equivalent samples of 2100-2250 students each (see Table 2). Approximately 60% of the students were female.

Insert Table 2

Measures of Instructional Background

As part of the registration process for the ACT Assessment, examinees were asked to indicate whether or not they had taken specific English courses. Examinees were included in the study if they reported having completed a course in literature or composition in grades 9, 10, and 11, and if they were currently enrolled in such a course. Approximately 85% of the college-bound seniors met this requirement. Raw score means and standard deviations, by form, for the selected students and the total group of examinees are shown in Table 2.

Index of Differential Item Performance

A contingency table procedure was used to measure DIP (Mantel & Haenszel, 1959). The Mantel-Haenszel statistic (MH-CHISQR, see Holland and Thayer, 1986) is based upon 2 x 2 contingency tables for each total score category. The MH-CHISQR statistic is distributed as a chi-square with one degree of freedom and is therefore a powerful unbiased test (Cox, 1970). Two statistics related to the MH-CHISQR, $\hat{\alpha}_{MH}$ and $\hat{z}_{MH}$, were also examined. The common odds ratio, $\hat{\alpha}_{MH}$, across the 2 x 2 tables, is given by

$$\hat{\alpha}_{MH} = \frac{\sum A_j D_j / T_j}{\sum B_j C_j / T_j} .$$

Where T_j is the total number of examinees in the jth matched set. A_j and C_j represent the number of examinees in the reference and focal groups who answered an item correctly. B_j and D_j are the number of examinees who responded incorrectly from the reference and focal groups. The reference group establishes a standard against which the performance of the focal group is compared. This ratio is on a scale of 0 to ∞ with $\alpha = 1$ representing a null value or no differential item performance.

The value of $\hat{\alpha}_{MH}$, for a studied item, is the "average factor by which the odds that a member of the reference group is correct on the studied item exceeds the corresponding odds for a comparable member of the focal group" (Holland and Thayer, 1986). Holland and Thayer suggest taking the log of $\hat{\alpha}_{MH}$ to put it into a symmetric scale with zero as the null value. Thus, we propose

$$\hat{z}_{MH} = -\frac{1}{1.7} \ln (\hat{\alpha}_{MH})$$

as a measure of the amount of differential item performance.

The value of $\hat{z}_{MH}$ is the average amount more difficult that a female examinee found the studied item than did a comparably-scoring male examinee. Positive values imply that the males found the item relatively easier than the females; negative values indicate that males found the item relatively harder.

In this study, the typical Mantel-Haenszel calculations were supplemented by a log-linear test of the three-way interaction between reference group membership, item response, and score category. The significance of this likelihood ratio chi-square value was calculated for each item. If the interaction was significant, then the assumption of the Mantel-Haenszel statistic--that no three-way interaction exists--was violated. This violation was used to qualify the interpretation of DIP in items with significant MH-CHISQR values.

Design and Analysis

A single factor design with replicated experiments (Winer, 1962, p. 213) was used to investigate the effect of English Usage item category on gender-based DIP. Item category was considered a fixed effect and test form was considered a random effect. All five ACTE item categories were crossed with the seven test forms, used essentially as replications, creating 35 distinct cells. Individual items were nested within form and item category.

The Mantel-Haenszel procedure was used separately for each form to estimate DIP indices for each of the 75 items. Negative values of the index represented items that were relatively easier for females, positive values represented items that were relatively easier for males, when matched on general level of achievement on the ACTE. The analysis was unweighted (Winer, 1962, p. 241) with the observed score in each cell as the signed, mean DIP index for the items in the cell. Analysis of variance was used to determine whether or not there was a significant item category effect on gender-based DIP.

Results

Table 2 shows the raw score means and standard deviations on ACTE performance by form. Although examinees were selected with similar course backgrounds, ACTE was easier for the female population across all forms of the test. They out-performed the male population by 2-3 raw score points. The population of students selected according to similar course backgrounds slightly out-performed the entire population of examinees on six of the seven test forms.

Insert Table 2

Table 3 shows the means and standard deviations of the DIP indices for each item category and each form. Means and standard deviations of the index values for the five item categories, averaged across all forms, are also presented. Diction and Style items have positive means for six of the seven forms. On the other hand, Sentence Structure items have negative means for six of the seven forms. The Grammar, Punctuation, and Logic and Organization items have means split between males and females. The overall means for Grammar and Punctuation were negative while the mean for Logic and Organization was positive.

Insert Table 3

The results of the calculation of the DIP indices did not show any significant three-way interactions for any of the items. Therefore, the assumption of no three-way interaction was not violated, indicating that the relationships between group membership and item response were consistent across score category.

The results of the analysis of variance are summarized in Table 4. No significant results were found. Since the test forms were constructed to be as equivalent as possible based upon detailed specifications, it was no surprise that the form main effect and the category by form interaction were not significant. However, the expectation of a significant category effect was not supported.

Insert Table 4

Discussion

Given comparable coursework, the results of this study show an overall tendency for female examinees to out-perform male examinees on the ACTE.

However, the results do not suggest the existence of gender-based differential item performance in English usage achievement items. When the ACT item classifications were used, no evidence of systematic DIP was found.

The results of this study are not surprising given the procedures followed in the construction of these tests. Items were selected for the test to match the content and statistical specifications of the test. Each version of the items was subjected to several reviews to help ensure the accuracy of the items. The completed test forms were also carefully reviewed by content experts, measurement experts and reviewers sensitive to issues of test and item bias to eliminate any items which did not seem to be appropriately focusing on achievement in English usage.

Despite the lack of support found by this study for its primary hypothesis--that systematic DIP would be found for algorithmic English usage items favoring females and for reasoning-oriented English usage items favoring males--a conclusion that these relationships do not exist may be premature. Close examination of the items in the ACTE indicates that most if not all of the items, regardless of category, may be considered similar in that they are more algorithmic than reasoning-oriented. That is, the assumption held in this study, that logic and organization and diction and style items focus on reasoning skills, may have been false. Comparable research done with different sets of English items, presumably some that require more reasoning, might yield support for the hypothesis. Since the specifications for the ACTE have been recently revised to include more items involving "rhetorical strategies" (and possibly reasoning), new forms of the ACTE may contain the necessary blend of test items that could facilitate further investigation. However, this is only speculation. The results of the present research found

no evidence of DIP based on gender in the ACTE and no support for the hypothesis that different categories of English usage test items perform differently for males and females.

REFERENCES

Cox, D.R. (1970) Analysis of Binary Data. London: Methuen and Co., Ltd.
Holland, P.W. (1985). On the study of Differential Item Performance without IRT. Proceeding of the Military Testing Association, October 1985, in press.

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

Green, D.R. (August, 1987). Sex differences in item performance on a standardized achievement battery. Paper presented at the annual meeting of the American Psychological Association, New York City, New York.

Holland, P.W. and Thayer, D.T. (1986, April). Differential Item Performance and the Mantel-Haenszel Procedure. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, California.

Huntley, R.M. and Plake, B.S. (April, 1980). Effect of selected item-writing practices on test performance: can relevant grammatical clues result in flawed items? Paper presented at the annual meeting of the American Educational Research Association, Boston, Massachusetts.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analyses of data from retrospective studies of disease. Journal of the National Cancer Institute, 22, 719-748.

National Assessment of Educational Progress. (1986). Writing trends across the decade, 1974-1984. Princeton, New Jersey: Educational Testing Service.

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. Chichester, England: John Wiley & Sons, 103-122.

Shepard, L., Camilli, G., and 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.

Winer, B.J. (1962) Statistical principles in experimental design. New York: McGraw-Hill.

TABLE 1

Description of ACTE Item Categories

1. **Punctuation.** The items in this category test such conventions as the use and placement of commas, colons, semicolons, dashes, parentheses, apostrophes, and quotations, questions, and exclamation marks.

2. **Grammar.** The items in this category test adjectives and adverbs, conjunctions, and agreement between subject and verb, and between pronouns and their antecedents.

3. **Sentence Structure.** The items in this category test relationships between/among clauses, placement of modifiers, parallelisms, and shifts in construction.

4. **Diction and Style.** The items in this category test precision in word choice, appropriateness in figurative language, and economy in writing.

5. **Logic and Organization.** The items in this category test the logical organization of ideas: paragraphing, transitions, unity and coherence.

TABLE 2

Sample Sizes, Raw Score Means and Standard Deviations
of ACTE Performance by Form

		Form						
		A	B	C	D	E	F	G
Sampled Students								
Males	N	850	940	907	925	948	970	905
	Mean	50.73	47.15	45.96	45.79	45.80	46.83	53.25
	SD	13.04	12.31	12.31	12.75	12.05	12.16	12.02
Females	N	1300	1300	1230	1302	1265	1280	1170
	Mean	54.00	50.03	49.43	49.17	48.33	49.44	55.44
	SD	12.34	11.85	12.32	11.50	11.33	11.94	10.46
Total	N	2150	2240	2137	2227	2213	2250	2075
	Mean	52.70	48.82	47.96	47.77	47.25	48.31	54.48
	SD	12.52	12.11	12.32	12.47	11.72	12.06	11.40
All Students								
	N	2532	2611	2479	2593	2554	2596	2532
	Mean	52.08	48.20	47.48	47.04	46.57	47.80	53.92
	SD	12.88	12.29	12.79	12.05	12.05	12.45	11.51

TABLE 3

Means and Standard Deviations of DIP Indices

Test Form		Item Categories				
		Grammar	Punctuation	Diction & Style	Logic & Organization	Sentence Structure
1	Mean	-.0100	-.0445	.0544	-.0123	-.0776
	SD	.1190	.1367	.1793	.2060	.1576
2	Mean	.0336	.0346	.0045	-.0127	-.0400
	SD	.0751	.0919	.1262	.1412	.1316
3	Mean	-.0053	.0217	-.0538	.0100	.0033
	SD	.0866	.0849	.1130	.1742	.0841
4	Mean	.0046	-.0370	.0640	.0106	-.0433
	SD	.0840	.0867	.0820	.1610	.1168
5	Mean	-.0511	.0183	.0408	.0418	-.0229
	SD	.1384	.0947	.1336	.0895	.1225
6	Mean	.0062	-.0021	.0100	-.0500	-.0050
	SD	.1447	.0684	.1399	.1034	.1397
7	Mean	-.0337	-.0585	.0259	.0350	-.0078
	SD	.0952	.1235	.1139	.1303	.1066
All Forms		-.0089	-.0136	.0234	.0045	-.0275
		.1113	.1016	.1312	.1451	.1249

TABLE 4

ANOVA Summary Table: Single Factor, Replicated Experiments Analysis (Unweighted)

Source	SS	df	MS	F	F prob.
Item Category	.00150	6	.00025	.197	--
Form (Replications)	.00946	4	.00237	1.866	--
Category x Form	.03043	24	.00127		
Total	.04139	34			