ACT Research Report Series 89-8

Differential Performance on a Direct Measure of Writing Skills for Black and White College Freshmen

Catherine Welch
Allen Doolittle
Joyce McLarty

November 1989

ACT

Differential Performance on a Direct Measure of Writing Skills for Black and White College Freshmen

Catherine Welch
Allen Doolittle
Joyce McLarty

ABSTRACT

The purpose of this study was to examine the differential performance for black and white college freshman found on a direct measure of writing skills. The direct measure consisted of responses to two individual prompts each requiring twenty minutes of testing time. Each essay was scored by two independent raters. Analysis of the data indicated that black examinees did not perform as well as white examinees on the essay test. The differences between the populations on the essay were similar in magnitude to differences found on a multiple-choice test of writing skills.

DIFFERENTIAL PERFORMANCE ON A DIRECT MEASURE OF WRITING SKILLS FOR BLACK AND WHITE COLLEGE FRESHMEN

The direct assessment of writing skills continues to be a central issue in education. Meredith and Williams (1984) reported that the direct assessment of writing is a major aspect of many state policies on testing. A recent issue of Educational Measurement: Issues and Practices was devoted to the assessment of writing. Recent studies of instruction have shown that schools are giving more attention to writing instruction for high school juniors and seniors (NAEP, 1986). This additional emphasis on direct writing assessment raises the question of what effect direct writing assessment has for various population subgroups. "NAEP results suggest that across the 10-year period from 1974 to 1984, trends in student achievement were much the same for many population subgroups. At ages 13 and 17, Black, Hispanic and White students showed relatively parallel trends in performance, with inconsistent trends or declines between 1974 and 1979 and gains from 1979 to 1984 (p. 6)."

Breland and Griswold (1981) concluded that black students at a given score level on a traditional college entrance test tended to write less well than the average white student at the same score level. White and female students tended to write better overall. Breland and Jones (1982) confirmed these earlier findings using the Cleary (1968) definition of bias, which stated "a test is biased if the criterion score predicted from the common regression line is consistently too high or too low for members of the subgroup (p. 115)." They concluded: "multiple-choice scores predict essay scores similarly for all groups sampled, with the exception that lower-performing groups tend to be overestimated by multiple-choice scores. Analyses showed that women consistently wrote more superior essays than would be predicted by the multiple-choice measures and that men and minority groups wrote fewer superior essays than would be predicted (p. 21)."

White (1985) reports results which suggest that a direct measure of writing may be fairer than a multiple choice test for ethnic minorities. "The TSWE, a conventional multiple-choice usage test, does not distribute the scores of minority students the way trained and accurate evaluators do. A carefully devised and properly scored essay test seems not to contain the same problem of validity for minorities (p. 81)."

This study will explore the following questions: 1) What are the differences between black and white examinee performance on a direct measure of writing skills? 2) Are these differences comparable in magnitude to objective test relationships? 3) What characteristics of the test contribute the most (or least) to these differences?

Methodology

The Instrument

The American College Testing Program (ACT) has been working to develop a test which will directly measure upper-level (college) writing proficiency by means of student writing samples. Each form of the CAAP Writing (Essay) Test consists of two independent writing prompts, each administered within 20 minutes. 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 in writing appropriate to the conventions of standard English.

Each examinee received two scores per response to a prompt. A "purpose" score reflected how well the examinees responded to the task required by situations described in the writing prompt. 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 on a 4-point scale by each of two raters working independently. The two scores assigned for each type of score were averaged so the resulting scale ranged from 1.0 to 4.0 in increments of 0.5. The two raters' scores for each essay had to either agree or be adjacent to be averaged. If the raters' scores differed by two or more points, a third rater adjudicated and determined the reported score. Detailed descriptions of the score points for both types of scores are provided in Appendix A.

Data Source

The essay test is part of ACT's Collegiate Assessment of Academic Proficiency (CAAP) program that was administered in the fall of 1988 to approximately 4,200 students. These students were primarily college freshmen from 15 postsecondary institutions distributed nationally. The institutions represented a variety of two- and four-year, and public and private institutions. The CAAP program offers a modular format which consists of multiple-choice tests in writing skills, mathematics, reading, critical thinking and science reasoning; and the essay test. Institutions may opt to administer one or any combination of these tests. Due to this modular format, sample sizes vary across tests. A sample of 140 black examinees and 375 white examinees took the complete CAAP test battery (all five tests). A larger sample (858 black examinees and 3,352 white examinees) took at least the essay test.

Analyses

The overall performance for all CAAP tests was compared for black and white examinees. Due to interest in evaluating the writing skills test performance and the essay performance for the same set of examinees, data from the smaller sample of examinees who took the complete CAAP battery were used

for these analyses. Unfortunately, the black examinees in this sample were primarily from one institution. To compare the performances of the black and white examinees a Hotelling's T^2 and post hoc *t*-tests were run.

Correlational analyses were conducted to show the basic relationships among the various CAAP tests separately for black and white examinees. Plots of the essay scores verse the multiple-choice test scores were examined for the black and white examinees. This allowed the range of multiple-choice scores at each essay score point to be examined for both groups.

The frequency and cumulative frequency distributions were compared for the two groups on both essays and across both types of scores. A Kolmogorov-Smirnov two-sample test was computed to determine whether or not the two distribution functions associated with the black and white examinees were drawn from the same population.

Results

Table 1 provides the means and standard deviations for the CAAP examinees who took the complete test battery and for examinees who took each of the tests of the CAAP battery. The results indicate that the students who took the complete battery did not perform as well as those students who took separate tests. The results also indicate that the black examinees did not perform as well as the white examinees on the multiple-choice tests or the essay.

Insert Table 1 about here

The equality of means of all seven variables listed in Table 2 were tested simultaneously by the multivariate Hotelling's T^2 procedure. The T^2 results were transformed to an F statistic, and the results indicate that the centroids for the black and white examinees are significantly different beyond the .001 level.

Table 2 provides the results of t-tests between the black and white examinees who took the complete CAAP test battery.* A statistically significant difference was found for all four multiple-choice tests and the essay test. The results suggest that the difference between black and white examinees was consistent across two measures of writing, direct and indirect.

Insert Table 2 about here

These results indicate that black examinees perform about the same relative to the white examinees on both the purpose scores (effect size = 1.42) and the language usage scores (effect size = 1.41). These values are consistent with the results for the multiple-choice writing skills test (effect size = 1.41).

Tables 3 and 4 provide the frequency distributions for black and white examinees on both essays. On essay one, the largest discrepancy occurred between the two and three point mark on the score scale. This was true for both the Purpose and the Language Usage score. On essay two, as shown in Table 4, black examinees received much lower Purpose scores than they had on the first essay. The Language Usage score distribution for essay two was

*Significant differences between the variances of the two samples were found for several of the score distributions even though the use of t as a test statistic assumes that the populations are equally variable. We believe this to be an artifact of the atypical black examinee sample. Results of the nonparametric Mann-Whitney test which does not make this assumption supported the results of the t-test.

Insert Tables 3 and 4 about here

similar to that found with essay one. The results from the Kolmogorov-Smirnov (K-S) tests are also reported in Tables 3 and 4. The K-S Z is computed from the largest difference between the two distributions. The differences between the black and white distributions were significant for all scores.

Intercorrelations of the various CAAP tests for both groups are reported in Tables 5 and 6. These correlations suggest that the relationship between the writing skills and the essay test may be slightly stronger for the white examinees than for the black examinees. The correlations between the Purpose score and the Writing Skills Test scores are noticeably lower than the correlations between the Language Usage score and the Writing Skills Test scores. This result is found for both groups. Examination of the scatter plots suggest positive relationships between the essay scores and the Writing Skills scores. The multiple-choice Writing Skills Test was a better predictor of both the Language Usage score and the Purpose score for white examinees than it was for black examinees.

Insert Table 5 and 6 about here

The reliability of the writing skills test differed slightly for black and white examinees. The largest difference in reliability was found on the Purpose score of the essay test. For white examinees the reported reliability was .52, for black examinees the reported reliability was .37. The interrater reliability estimates for the essays were higher for the black examinees on both Language Usage scores. The largest difference between groups in

correlations was on the Purpose score for essay two. The estimates of reliability for the multiple-choice and essay test, and interrater reliabilities for the essays are reported in Table 7.

Insert Table 7 about here

Discussion

The results from the sample in this study indicate that black examinee performance on the CAAP essay test was consistent with black examinee performance on the CAAP multiple-choice test of writing skills.

The black examinee performance remained fairly consistent across essays on the Language Usage score. However, the second essay resulted in lower Purpose scores for both groups. This difference was greater for black examinees than for white examinees. The topics of the two essays were examined to provide a possible explanation for this difference. (The test is a secure instrument at this time and the specific prompts can not be released.) The first prompt asked examinees to allocate funds to various projects and the second asked students' opinions on grading policies. No apparent reason for the discrepancy between essays was found.

Based on the statistical results a small sample of papers was examined for the black and white examinees. This sample included twenty papers each for black and white examinees for both prompts. The most obvious difference between the two sets of papers was the length of the responses provided. For both prompts the black examinees' responses were shorter than the white examinees' responses. Length was measured by the average number of sentences written and the average number of words per sentence.

Both black and white examinees provided longer responses to the first prompt. This could be a result of both the topic and the position of the

prompt. Beyond the brevity of responses there were no common errors committed by examinees that stood out for either black or white examinees. The brevity of the black examinees' responses could possibly be an institutional effect since the majority of the black examinee sample came from one institution.

Even though strict guidelines are provided for test administration, ACT has no control over the actual administration of the CAAP tests. Institutions are responsible for their own test administrators and time schedules. This makes comparability of administrations between institutions somewhat difficult, especially for the essay tests where timing is a critical factor.

The results indicate that for the sample studied white examinees outperformed black examinees to a similar degree on the CAAP multiple-choice and essay tests. These results must be interpreted cautiously, as the samples, and particularly the black sample, were probably atypical. A larger, carefully controlled sample of black examinees is a necessity before these results can be generalized to the college freshman population.

References

Breland, H.M., & Griswold, P.A. (1981). Group comparison for basic skills measures. New York: College Entrance Examination Board.

Breland, H.M., & Jones, R.J. (1982). Perceptions of writing skills. New York: College Entrance Examination Board.

Cleary, T.A. (1968). Test bias: prediction of grades of Negro and white students in integrated colleges. Journal of Educational Measurement, 3, 115-124.

Cooper, P.L. (1984). The assessment of writing ability: a review of research. GRE Board Research Report GREB No. 82-15R.

Meredith, V., & Williams, P. (1984). Issues in direct writing assessment: Problem identification and control. Educational Measurement: Issues and Practice, 3,11-15.

National Assessment of Educational Progress. (1986). Writing trends across the decade, 1974-84. Princeton, NJ: Educational Testing Service.

White, E.M. (1985). Teaching and assessing writing. San Francisco: Jossey-Bass.

Table 1

Means and (Standard Deviations) of CAAP Examinees Who Took the Complete Test Battery

	Blacks	Whites	Total
Writing Skills	29.11 (9.32)	44.65 (11.06)	40.07 (12.75)
Purpose Score-Essay 1	2.02 (.65)	2.77 (.65)	2.57 (.73)
Language Usage Score-Essay 1	2.25 (.42)	2.83 (.51)	2.66 (.56)
Purpose Score-Essay 2	1.43 (.60)	2.32 (.76)	2.07 (.82)
Language Usage Score-Essay 2	2.13 (.50)	2.78 (.53)	2.59 (.60)
Purpose Score Composite	1.72 (.49)	2.55 (.58)	2.32 (.67)
Language Usage Score Composite	2.19 (.38)	2.81 (.44)	2.62 (.51)
	N = 140	N = 375	N = 556

Means and (Standard Deviations) of All CAAP Examinees

	Blacks	Whites	Total
Writing Skills	32.65 (12.01)*	45.53 (12.86)*	43.40 (13.23)*
Purpose Score-Essay 1	2.07 (.82)+	2.71 (.75)+	2.53 (.80)+
Language Usage Score Essay 1	2.22 (.69)	2.84 (.63)	2.65 (.67)
Purpose Score-Essay 2	1.61 (.78)	2.35 (.80)	2.14 (.85)
Language Usage Score Essay 2	2.11 (.69)	2.79 (.60)	2.58 (.67)
Purpose Score Composite	1.84 (.69)	2.53 (.63)	2.34 (.65)
Language Usage Score Composite	2.16 (.64)	2.82 (.55)	2.62 (.60)
	*N = 858 +N = 388	*N = 3352 +N = 970	*N = 2000 +N = 1379

Note: More examinees took the multiple-choice CAAP Writing Skills Test than the CAAP Writing (Essay) Test. The difference in N-counts reflects this.

Table 2

t-Test Results for Black and White Differences on the CAAP Tests for Examinees Who Took the Complete Test Battery

	t-Value*	Probability	Effect Size
Writing Skills	14.78	.000	1.41
Purpose Score-Essay 1	11.76	.000	1.17
Language Usage Score-Essay 1	11.96	.000	1.13
Purpose Score-Essay 2	12.46	.000	1.17
Language Usage Score-Essay 2	12.75	.000	1.24
Purpose Score Composite	14.90	.000	1.42
Language Usage Score Composite	14.71	.000	1.41

* N = 140 Blacks
375 Whites
df = 513

Table 3

Frequency Distributions of Scores on Essay One and Kolmogorov-Smirnov Z-test (K-S Z) Results for Black and White Examinees Who Took the Complete Test Battery

	Black		White			
Purpose Score	Percent	Cum. Percent	Percent	Cum. Percent	K-S Z*	Probability
not ratable	2.7	2.7	0.4	0.4	5.07	.000
1.0	17.2	19.9	5.1	5.5		
1.5	5.5	25.4	2.8	8.3		
2.0	40.2	65.6	23.0	31.4		
2.5	15.2	80.9	14.6	45.9		
3.0	14.5	95.3	38.4	84.3		
3.5	2.0	97.3	7.8	92.1		
4.0	2.7	100.0	7.9	100.0		

	Black		White			
Language Usage Score	Percent	Cum. Percent	Percent	Cum. Percent	K-S Z*	Probability
not ratable	3.9	3.9	1.2	1.2	4.91	.000
1.0	2.3	6.3	.3	1.5		
1.5	1.6	7.8	.6	2.2		
2.0	55.5	63.3	17.3	19.5		
2.5	12.9	76.2	14.2	33.6		
3.0	21.9	98.0	54.4	88.0		
3.5	.8	98.8	6.4	94.4		
4.0	1.2	100.0	5.6	100.0		

* df = 7

Table 4

Frequency Distributions of Scores on Essay Two and Kolmogorov-Smirnov Z-test (K-S Z) Results for Black and White Examinees Who Took the Complete Test Battery

	Black		White			
Purpose Score	Percent	Cum. Percent	Percent	Cum. Percent	K-S Z*	Probability
not ratable	2.7	2.7	0.5	0.5	5.30	.000
1.0	43.8	46.5	13.7	14.3		
1.5	13.3	59.8	5.6	19.9		
2.0	23.4	83.2	30.9	50.8		
2.5	7.8	91.0	15.9	66.7		
3.0	6.3	97.3	23.0	89.7		
3.5	.4	97.7	5.9	95.7		
4.0	2.3	100.0	4.3	100.0		

	Black		White			
Language Usage Score	Percent	Cum. Percent	Percent	Cum. Percent	K-S Z*	Probability
not ratable	3.9	3.9	1.2	1.2	5.19	.000
1.0	6.6	10.5	0.5	1.7		
1.5	4.7	15.2	0.5	2.3		
2.0	51.2	66.4	18.7	21.0		
2.5	15.6	82.0	17.3	38.3		
3.0	16.8	98.8	53.0	91.2		
3.5	.8	99.6	5.1	96.3		
4.0	.4	100.0	3.7	100.0		

* df = 7

Table 5

Correlations Among CAAP Scores for Black Examinees (N=140)

	PR1	LU1	PR2	LU2	PRC	LUC
Writing Skills (WS)	.21	.34	.14	.32	.23	.40
Purpose Score-Essay 1 (PR1)		.32	.23	.30	.80	.38
Language Usage Score-Essay 1 (LU1)			.16	.36	.28	.79
Purpose Score-Essay 2 (PR2)				.12	.76	.14
Language Usage Score-Essay 2 (LU2)					.28	.86
Purpose Score-Composite (PRC)						.34

Table 6

Correlations Among CAAP Scores for White Examinees (N=375)

	PR1	LU1	PR2	LU2	PRC	LUC
Writing Skills (WS)	.27	.41	.19	.37	.27	.46
Purpose Score-Essay 1 (PR1)		.48	.35	.33	.79	.48
Language Usage Score-Essay 1 (LU1)			.23	.43	.42	.84
Purpose Score-Essay 2 (PR2)				.41	.85	.38
Language Usage Score-Essay 2 (LU2)					.45	.85
Purpose Score-Composite (PRC)						.51

Table 7

Reliability Estimates for the CAAP Writing Skills Test and the CAAP Writing (Essay) Test

	Blacks	Whites
Writing Skills	.90	.93
Purpose Score	.37	.52
Language Usage Score	.53	.60

Interrater Reliability Estimates for the CAAP Writing (Essay) Test

	Blacks	Whites
Purpose Score-Essay 1	.84	.81
Language Usage Score-Essay 1	.81	.76
Purpose Score-Essay 2	.76	.83
Language Usage Score-Essay 2	.82	.78

Appendix A

Interpretation of CAAP Writing (Essay) Test Scores

Purpose Score

4 ***Excellent***. These papers take a clear position on the issue described in the writing assignment and support that position with an elaborated argument that consists of well-developed reasons which are focused on the specific concerns of the audience identified in the assignment.

3 ***Good***. These papers take a clear position on the issue described in the writing assignment and support that position with an argument that consists of several reasons which are focused on the specific concerns of the audience identified in the assignment, but which are not sufficiently developed to constitute and elaborated argument.

2 ***Weak***. These paper take a position on the issue described in the writing assignment and support that position with a brief argument consisting of only one or two undeveloped reasons related to the specific concerns of the audience identified in the assignment.

1 ***Poor***. These papers either do not take a position on the issue described in the writing assignment, or they take a position but support that position with only one undeveloped reason related to the specific concerns of the audience identified in the assignment.

Language Usage Score*

4 ***Excellent***. These paper contain few errors in sentence structure, grammar, punctuation, or style, one or two of which may cause some awkwardness of expression, but none that cause any significant awkwardness of expression or any confusion in meaning. The sentence patterns are varied and the prose is fluent. In general, control of the language is excellent.

3 ***Good***. These papers contain some errors in sentence structure, grammar, punctuation, or style, one or two of which may cause some significant awkwardness of expression or confusion in meaning. The sentence patterns are varied and the prose is fluent. In general, control of the language is accurate and clear, but not excellent.

2 ***Weak***. These papers may contain several errors in sentence structure, grammar, punctuation, or style resulting in significant awkwardness of expression or confusion in meaning. Or, these papers may lack variety in sentence patterns or fluency. In general, control of the language is weak.

1 ***Poor***. These papers contain a number of errors in sentence structure, grammar, punctuation, or style resulting in significant awkwardness of expression or confusion in meaning. In general, control of the language is poor.