DOCUMENT RESUME

ED 382 689 TM 023 230

AUTHOR	Zhu, Daming; Thompson, Tony D.
TITLE	Gender and Ethnic Differences in Tendencies To Omit Responses on Multiple-Choice Tests Using Number-Rigbt Scoring.
PUB DATE	Apr 95
NOTE	40p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 18-22, 1995).
PUB TYPE	Reports - Research/Technical (143) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Academic Achievement; Black Students; *Ethnic Groups; High Schools; *High School Students; Hispanic Americans; Mathematics Tests; *Multiple Choice Tests; Racial Differences; Responses; *Scores; *Sex Differences; Test Results; White Students
IDENTIFIERS	ACT Assessment; *Number Right Scoring; *Omitted Responses

ABSTRACT

This study attempted to control differences in achievement when examining omitting tendencies of examinees. Test data of randomly sampled examinees (7 samples of 2,000 examinees each) from one national administration of the ACT Assessment were used. The number of responses omitted per examinee was examined over all examinees and over only those who omitted responses. The relationship between test scores and number of omits was negative and weak. More examinees at lower levels omitted responses (and more of them), but some at higher levels also omitted a surprising number of responses. No significant differences in tendencies to omit were found between males and females. Statistically significant differences in the tendency to omit were found among examinees of different ethnic groups, especially between African origin and Caucasian origin samples and between Hispanic origin and Caucasian origin examples, but the differences were all less than one omit on average on the mathematics test. Controlling for potential achievement level differences between ethnic groups by using covariates resulted in only a minimal reduction in omitting tendencies between the groups, suggesting that this is not the explanation for differences between ethnic groups. Ten tables summarize study findings. (Contains 22 references.) (SLD)

GENDER AND ETHNIC DIFFERENCES IN TENDENCIES TO OMIT RESPONSES ON MULTIPLE-CHOICE TESTS USING NUMBER-RIGHT SCORING

Daming Zhu and Tony D. Thompson

American College Testing (ACT)

April, 1995

Paper Presented at the annual meeting of the
American Educational Research Association
in San Francisco, CA

BACKGROUND

Given the purposes of standardized multiple choice tests and the ways test scores are generally used (such as admission to colleges, scholarship awards, hiring decisions, etc.), it is critically important that test scores accurately reflect examinees' achievement. When number-right scoring is used the test score is computed based simply on the number of correct answers; omitted responses are scored as incorrect. Under this scoring system and with accompanying test directions usually encouraging answering every question, examinees who omit test items will generally disadvantage themselves when other examinees guess randomly on some items[1]. The test scores are fair with regard to guessing and obtaining chance scores only when every examinee attempts all the items (Lord, 1975; Wood, 1976; Frary, 1988; Ebel & Frisbie, 1991).

Nonetheless, for various reasons, some examinees do not answer all items (Davis, 1967; Sabers & Feldt, 1968; Traub, et al, 1969; Traub & Hambleton, 1972; Angoff & Schrader, 1984; Grandy, 1987; Ben-Shakhar & Sinai, 1991). Numerous research findings indicate that on a multiple-choice test using number-right scoring, some examinees respond to all items even if they have to guess blindly on some items, while other examinees leave items unanswered despite test directions that encourage guessing. This affects the accuracy and interpretation of the reported scores produced by number-right scoring. Furthermore, examinees with differing personality traits probably differ in their guessing behaviors (Votaw, 1936; Slakter, 1968a; Slakter, 1968b; Abu-sayf, 1979; Albanese, 1988) and guessing tendencies may also differ for examinees of different gender or cultural origins (Grandy,

[1] It is assumed here that the test in question has been designed to be a power test. In the case of a test purposely designed to be speeded, not-reached items at the end of the test may be indicative of a lower-ability examinee, although even in this case an examinee should quickly guess at items at the end of the test if the test is scored number-right.

1987; Ben-Shakhar & Sinai, 1991). If large group differences exist, measurement error will certainly increase and possibly bias estimates of achievement for certain examinee sub-populations.

Although many studies have investigated such topics as test scoring procedures and differential item functioning, little research has been directed toward examining group differences in tendencies to omit items on standardized tests when number-right scoring is used. In one study examining omitting, Grandy (1987) reported that female examinees (as compared to male examinees), non-whites, particularly blacks, and those with lower test scores left more items unanswered on the GRE General Test. After the "corrected-for-not-guessing" scores (the originai score plus the chance score based on the number of existing omits in each case) were entered into the regression model, white/non-white and gender variables continued to make statistically significant contributions in accounting for the variations in the number of omits. But in a practicai sense the effects Grandy found that attributed to gender and ethnicity were very small: averaging less than 0.04 omits for gender and 0.03 to 0.06 omits for ethnicity on the three tests of the GRE General Test.

In another study, Ben-Shakhar and Sinai (1991) used Ziller's index of guessing (Ziller, 1957) to study gender differences in the tendency to omit or to guess on multiple-choice tests. Ziller's index (G) is given by,

$$G=\frac{W_2}{W_2+O} \tag{1}$$

where O is the number of total omits, and W_2 is the estimated number of guesses, defined as,

$$W_2 = W_1 + \frac{W_1}{K-1} \qquad (2)$$

where W_1 is the number of wrong answers and K is the number of alternative responses per item. The study also used a modification of Ziller's index (following Angoff & Schrader, 1981) by defining O as the number of trailing omits rather than the number of total omits. Based on their results, they concluded that females tend to omit more items than males. However, the validity of Ziller's index is questionable because it assumes that every wrong answer results from random guessing. In addition, the number of omits on the tests used in their study may have been confounded by the examinees' relevant knowledge/achievement levels.

The current study seeks to control differences in achievement when examining omitting tendencies of examinees. It is expected, of course, that high achievement examinees will omit fewer items than low achievement examinees. The research question of interest is whether group-level differences in omitting tendencies can be explained by potential achievement differences between the groups. In this study, examinees representing males and females and four ethnicity categories were randomly selected from a national administration of a large scale standardized test battery. The study focused on the following: 1) differences in omit rates between gender and ethnic groups; 2) differences in omit rates between low, medium, and high scoring examinee groups; 3) differences in omit rates between gender and ethnic groups with achievement differences between the groups controlled.

METHOD

Test data of randomly sampled examinees from the one national administration of the ACT Assessment were used in this study. Seven samples of 2,000 examinees each were selected. One sample was selected from each gender (male and female), and one sample of each of the following ethnicity groups: Afro-American/Black, Asian-American and Pacific Islander, Caucasian-American/White, and Mexican-American and other Hispanic origins (these ethnicity groups are referred hereafter as African origins, Asian origins, Caucasian origins, and Hispanic origins, respectively). The gender and ethnicity identity were based upon the examinees' self-reported information provided on the examinees's registration forms for taking the ACT Assessment[2]. In addition, a nationally representative sample (hereafter referred to as the national sample) of examinees was selected so that the proportions of gender and ethnic groups matched those of the overall test administration population.

The primary purpose of the tests in the ACT Assessment is to assess examinees' academic achievement in order to provide information for college and university admission and placement decisions. The test is a standardized test battery using number-right scoring and is administered nationwide in the United States. Directions for taking the tests are printed on the front cover of the ACT Assessment test booklet, which examinees are told to read in the standard oral instructions announced by the test room supervisor. Included in the test directions is the information regarding scoring and guessing exactly as:

> Only responses marked on your answer document will be scored. Your score on each test will be based only on the number of questions you answer correctly. You will NOT be penalized for guessing. *HENCE IT IS TO YOUR ADVANTAGE TO ANSWER EVERY QUESTION.*

The ACT Assessment test battery contains four curriculum-based tests in the content areas of English, mathematics, reading, and science reasoning, with 75, 60, 40, and 40 items,

[2] It is recognized that the categories listed above do not constitute an ideal system for categorizing individuals into ethnic groups and that any one of these categories represents a diverse and ever-changing mixture of individuals and sub groups. It was, however, the best data available to us.

respectively. The English Test, the Reading Test, and the Science Reasoning Test each consists of several reading passages and attendant passage-based items. Each item in the three passage-based tests has four answer choices. There are no passages in the Mathematics Test and the items in the test are discrete, each containing five answer choices. Four individual test scores, a composite score (the rounded arithmetic mean of the four test scores), and several subscores are provided on each examinee's score report. The reported ACT test scores are scaled and range from 1 to 36 for the four test scores and the composite score. The scale scores are converted from the corresponding raw scores based on an equating to previous test forms.

In the sampled data of this study, a few examinees did not complete the whole test battery (i.e., did not take all four tests) and were excluded from the analyses. The final national sample consisted of 1,998 examinees. The samples of female, male, African origin, Asian origin, and Caucasian origin all had 1,999 valid cases. The Hispanic origin sample size remained at 2,000.

Examinees' omitted responses on each test and on the test battery were summed. For the purposes of the study, embedded omits were not differentiated from trailing omits, as the focus of the study was to examine the overall omit rate. It would be interesting for a latter study to examine group differences in embedded omit rates and trailing omit rates.

Omitting tendencies for the gender and ethnic groups were examined on all four of the content-area tests and the test battery. When attempting to control for examinee group achievement differences, we focused on the Mathematics Test. To control examinee group differences in high school math knowledge, two estimates of examinee's achievement level in mathematics were computed to be used as covariates. The two estimates were both based on an examinee's responses to the first-x items on the Mathematics Test. The values of x were determined based on the extent of trailing omits found in the sample data so that at least 99% of the examinees did not start their trailing omits before or at the *xth* item on the test. The value of x was 52 for female and male comparison and the value of x was 47 for the ethnic group comparisons. One estimate of math achievement was the number of items answered correctly by the examinee in the first-x items on the Mathematics Test. The second estimate was an achievement index calculated by using item response theory (IRT) three parameter logistic model in the BILOG program (Mislevy & Bock, 1990) which was also based on the examinee's responses to the first-x items. Two other variables, external to the test, that were related to math achievement, ^xaminees' self-reported years of high school math courses taken (ranging from 0 to 8 with each increment of 1 indicating a half year), and examinees' self-reported high school math course grade (ranging from 0 to 4 in the increment of 1) were also used as covariates.

To examine the extent of omits for examinees at w, medium, and high test performance levels, the examinees in the national sample were assigned to one of the three groups according to their composite test scores. The grouping criteria were: high achievement—ACT composite score ?4 or higher (top 27% examinees on the national norms for reporting ACT Assessment scores during the 1993-94 testing year), medium achievement—ACT composite score 18 to 23 (middle 47% examinees on the national norms); and low achievement—ACT composite score 17 or lower (bottom 26% examinees on the national norms).

The analyses were conducted with the computer statistical package SAS (SAS Institute, Inc., 1985). The probability level for significance of difference was set at .05. The conducted analyses included frequency distributions and descriptive statistics, Pearson correlations, analysis of variance (ANOVA) and analysis of covariance (ANCOVA) In the post hoc multiple comparisons among the adjusted mean o,nits of the ethnic samples, Scheffé method for ANCOVA (Huitema, 198C; Hays, 1988) was employed. Since no ready procedures of the Scheffé multiple comparisons for ANCOVA is available in SAS, the computations of the test statistic of the Scheffé test on the adjusted mean omits in this study were performed using SAS, EXCEL (Microsoft Version 4.0 for Windows), and some manual calculations.

RESULTS AND DISCUSSION

Omitted Responses on the Tests

The number of responses omitted per examinee was examined in two ways: over all examinees in a sample and over only those examinees who omitted responses in a sample. The former gives an overall view of the number of omits in a sample and the latter makes a more accurate picture of the extent of omits over the examinees who actually omitted.

Omits over All Examinees and over only Examinees with Omits

Table 1 includes a summary of the number of omits for all examinees and for only those examinees in the national sample with omits. For all examinees, the means of number of omits on the individual tests ranged from 0.6 to 1.1 with standard deviations ranging from 2.5 to 4.3. The mean omits was 3.0 and the standard deviation was 10.2 on the test battery. The all-examinee mean omits was quite small, which can be explained by noting that 75.9 percent of the examinees did not omit items. The statistics of omits based on only those examinees with omits showed that 24.1 percent of the examinees in the national sample had omits. They omitted an average of 12.3 responses with a standard deviation of 17.9 and a

Table 1. Omits in National Sample by Test and Battery

		Examinees	
		All	With Omits
English Test		100.0%	11.5%
(75 Items)	Mean Omits	1.1	9.4
	SD of Omits	4.3	8.9
	Median Omits	0.0	6.0
	Max. Omits	43	43
Math Test		100%	10.7%
(60 Items)	Mean Omits	0.6	5.9
	SD of Omits	2.9	7.1
	Median Omits	0.0	2.0
	Max. Omits	32	32
Reading Test		100.0%	11.7%
(40 Items)	Mean Omits	0.7	5.7
	SD of Omits	2.6	5.3
	Median Omits	0.0	4.0
	Max. Omits	25	25
Science Test		100.0%	9.5%
(40 Items)	Mean Omits	0.6	6.1
	SD of Omits	2.5	5.8
	Median Omits	0.0	4.0
	Max. Omits	26	26
Battery		100.0%	24.1%
(215 Items)	Mean Omits	3.0	12.3
	SD of Omits	10.2	17.9
	Median Omits	0.0	4.0
	Max. Omits	101	101

maximum of 101 omits on the test battery. The number of examinees with omits on individual tests ranged from 9.5% to 11.7% and their mean omits on individual tests ranged from 5.7 to 9.4 and the maximum omits ranged from 25 to 43.

These results show that when assessing the impact of omits on the test scores, it is important to look at the average number of omits based on only those who did omit responses. Even though the average examinee omitted very few items on the tests examined, some examinees omitted many items.

Omits at Different Achievement Levels

Table 2 displays the correlation coefficients among the test scores and the number of omits. The relationship between the omits and scores on each test as well as on the test battery were all negative and fairly weak. The Pearson correlation coefficients between the number of omits and test scores were -.28 on the English Test, -.11 on the Mathematics Test, -.24 on the Reading Test, -.19 on the Science Reasoning Test, and -.22 on the test battery. The Pearson correlation coefficients for the number of omits between any two tests ranged from .53 to .74.

Looking at the number of omits on the test battery at three achievement levels in the national sample (Table 3), the percentages of examinees with omits and the average number of omits on the test battery decreased from low achievement level (31.2%, 20.3), to medium achievement level (25.0%, 10.7), and to high achievement level (17.3%, 4.5). The results showed that a considerable number of examinees in the high achievement group still omitted responses and the mean number of omits on the test battery for these examinees was 4.5 with a standard deviation of 5.1 and a maximum number of omits of 24. Thus, although low

Table 2. Correlation Coefficients between Test Scores* and Number of Omitted Responses

	Omits on Test					Test Score				
	English	Math	Reading	Science	Battery	English	Math	Reading	Science	Battery
Omits on English	--	0.53	0.55	0.53	0.84	-0.28				-0.22
Omits on Math		--	0.60	0.66	0.82		-0.11			-0.12
Omits on Reading			--	0.74	0.83			-0.24		-0.20
Omits on Science				--	0.84				-0.19	-0.17
Omits on Battery					--					-0.22

* Raw scores for each test and scale score for test battery

Table 3: Omits for Examinees with Omits in Different Achievement Level Groups* in National Sample

		Achievement Level*			
		High	Medium	Low	Total
Number of Examinees		618	905	475	1,998
Examinees with Omits on Battery (N):		107	226	148	481
	(% of Group):	17.3	25.0	31.2	24.1
English Test	Mean Omits	1.1	3.6	8.3	9.4
(75 Items)	SD of Omits	2.3	6.2	10.4	8.9
	Max. Omits	11	36	43	43
Math Test	Mean Omits	1.1	3.6	8.3	5.9
(60 Items)	SD of Omits	2.3	6.2	10.4	7.1
	Max. Omits	17	28	32	32
Reading Test	Mean Omits	1.0	2.5	4.4	5.7
(40 Items)	SD of Omits	2.2	4.1	6.1	5.3
	Max. Omits	11	24	25	25
Science Test	Mean Omits	1.0	2.0	4.0	6.1
(40 Items)	SD of Omits	2.0	3.9	6.5	5.8
	Max. Omits	13	19	26	26
Battery	Mean Omits	4.5	10.7	20.3	12.3
(215 Items)	SD of Omits	5.1	14.4	24.4	17.9
	Max. Omits	24	77	101	101

* The grouping was according to examinees' ACT Composite scores. High achievement: ACT Composite score >= 24; Medium achievement: 18 <= ACT Composite score <= 23; Low achievement: ACT Composite score <=17.

achievers tended to omit more frequently than high achievers, in a few cases even high-achievenient examinees omitted a substantial number of items.

The correlational and achievement level results reveal that although the tendency to omit and the examinee's test performance are inversely related, the relationship between them is only moderate. Many medium and high achievers omit a substantial number of items.

Tendencies to Omit Responses between Genders and among Ethnic Samples

As in the national sample, the number of omits in the gender samples and the ethnic samples were examined over all examinees as well as over only those examinees with omits. To study the differences in their tendencies to omit responses while controlling for achievement level, the numbers of omits on the Mathematics Test were compared between gender samples and among ethnic samples. Three main comparisons were performed. The number of omits were first compared by ANOVA, and then by ANCOVA using two different sets of covariate variables in which examinees' achievement estimates, years of high school math courses taken, and high school math course average grades were controlled as covariates. The related results are described in the following sections.

Omits and Mean Omits Adjusted by Achievement between Genders

Table 4 displays the summary statistics of omits in female and male samples. The female all-examinee mean omits on the test battery was 2.9 and the males' was 2.6. The female sample had more examinees (27.6%) with omits on the test battery compared to the male sample (21.4%), but male examinees with omits omitted more (an average of 11.9 and a maximum of 128) than did their female counterparts (an average of 10.5 and

Table 4: Omits in Female Sample and in Male Sample

		Female Sample (N=1999)		Male Sample (N=1999)	
		All Examinees	With Omits	All Examinees	With Omits
English Test		100.0%	11.4%	100.0%	10.7%
(75 Items)	Mean Omits	1.0	8.3	1.0	9.1
	SD of Omits	3.7	7.5	4.0	8.8
	Median Omits	0.0	6.0	0.0	6.0
	Max. Omits	39	39	37	37
Math Test		100.0%	11.9%	100.0%	8.2%
(60 Items)	Mean Omits	0.7	5.9	0.6	6.7
	SD of Omits	2.9	6.5	2.9	7.8
	Median Omits	0.0	3.0	0.0	3.0
	Max. Omits	33	33	36	36
Reading Test		100.0%	14.9%	100.0%	10.0%
(40 Items)	Mean Omits	0.7	4.6	0.6	5.7
	SD of Omits	2.3	4.1	2.4	5.5
	Median Omits	0.0	3.0	0.0	3.0
	Max. Omits	20	20	32	32
Science Test		100.0%	11.0%	100.0%	7.7%
(40 Items)	Mean Omits	0.6	5.1	0.5	6.1
	SD of Omits	2.2	4.4	2.4	6.4
	Median Omits	0.0	4.0	0.0	3.5
	Max. Omits	23	23	28	28
Battery		100.0%	27.6%	100.0%	21.4%
(215 Items)	Mean Omits	2.9	10.5	2.6	11.9
	SD of Omits	8.6	13.6	9.9	18.7
	Median Omits	0.0	5.0	0.0	4.0
	Max. Omits	88	88	128	128

a maximum of 88) on the test battery. The trend was the same on individual tests. When the number of omits on the test battery between the female sample and male sample were compared with ANOVA, the results indicated that the difference was not statistically significant. The probability of such difference occurring by chance was at *P*=.2401. For the number of omits on the Mathematics Test between the female sample and male sample, the ANOVA results also showed that they did not differ significantly (*P*=.0916).

Since no differences between female and male mean omits were found on the Mathematics Test, the planned analysis of examining male and female omit rates on the Mathematics test while controlling for possible achievement level differences by use of the covariate variables described earlier is not provided in detail here. It is sufficient to state that when achievement level differences were controlled for, females and males omit rates were still not statistically different.

Omits and Adjusted Mean Omits among Ethnic Samples

Omits among Ethnic Samples

Among ethnic samples (Table 5), Caucasian origin had the lowest percent of examinees with omits on the test battery (21.1%), followed by Asian origin (25.8%), then Hispanic origin (37.4%). African origin had the highest percentage of examinees with omits (41.1%). The average number of omits for examinees with omits in each of the four samples on the test battery (in the order of African origin, Asian origin, Caucasian origin, and Hispanic origin) were 16.7, 12.2, 9.9, and 17.0, respectively, and the maximum number of omits in each group was 118, 145, 77, and 122. The mean number of omits for only those examinees with omits on each test followed basically the same pattern among the four

Table 5. Omits by Ethnic Origin Samples

		African Origin (N=1,999)		Asian Origin (N=1,999)		Caucasian Origin (N=1,999)		Hispanic Origin (N=2,000)	
		All Examinees	With Omits	All Examinees	With Omits	All Examinees	With Omits	All Examinees	With Omits
English Test		100.0%	24.2%	100.0%	11.8%	100.0%	93.5%	100.0%	22.5%
(75 Items)	Mean Omits	2.7	11.0	1.1	9.6	0.7	7.2	2.4	10.8
	SD of Omits	6.6	9.4	4.3	9.0	3.0	6.9	6.3	9.2
	Median Omits	0.0	9.0	0.0	7.0	0.0	4.0	0.0	9.0
	Max. Omits	44	44	50	50	30	30	40	40
Math Test		100.0%	19.2%	100.0%	12.2%	100.0%	8.7%	100.0%	18.4%
(60 Items)	Mean Omits	1.3	7.0	0.7	5.6	0.4	5.1	1.4	7.6
	SD of Omits	4.2	7.2	2.9	6.4	2.2	5.6	4.5	7.8
	Median Omits	0.0	4.0	0.0	2.0	0.0	3.0	0.0	5.0
	Max. Omits	35	35	37	37	33	33	43	43
Reading Test		100.0%	24.2%	100.0%	12.7%	100.0%	10.2%	100.0%	19.7%
(40 Items)	Mean Omits	1.7	6.9	0.7	5.5	0.5	4.9	1.4	7.0
	SD of Omits	4.1	5.6	2.6	5.3	2.0	4.3	3.7	5.4
	Median Omits	0.0	6.0	0.0	4.0	0.0	3.0	0.0	6.0
	Max. Omits	27	27	32	32	19	19	25	25
Science Test		100.0%	16.5%	100.0%	10.2%	100.0%	8.5%	100.0%	16.1%
(40 Items)	Mean Omits	1.2	7.2	0.6	6.3	0.5	5.6	1.1	7.1
	SD of Omits	3.5	5.7	2.7	6.0	2.2	5.1	3.5	5.8
	Median Omits	0.0	6.0	0.0	4.0	0.0	4.0	0.0	6.0
	Max. Omits	26	26	31	31	33	33	27	27
Battery		100.0%	41.1%	100.0%	25.8%	100.0%	21.1%	100.0%	37.4%
(215 Items)	Mean Omits	6.8	16.7	3.1	12.2	2.1	9.9	6.4	17.0
	SD of Omits	15.6	20.6	10.7	18.4	7.4	13.6	15.2	20.9
	Median Omits	0.0	9.0	0.0	4.0	0.0	4.0	0.0	8.0
	Max. Omits	118	118	145	145	77	77	122	122

samples with the exception that African origin sample had a slightly higher mean number of omits (11.0 and 7.2) on both the English Test and the Science Reasoning Test than Hispanic origin did (10.8 and 7.1). The ANOVA of the number of omits on test battery among ethnic samples showed that the overall differences were statistically significant at *P*=.0001 level. The results of ANOVA on number of omits on the Mathematics Test also showed that the overall differences among the four groups were significant at *P*=.0001. The results of the Scheffé post hoc tests indicated that significant differences existed between all pairs of samples at *P*=.05 except between African origins and Hispanic origins and between Asian origins and Caucasian origins (Table 6). The results from the analyses with number of omits on the test battery were similar.

Mean Omits Adjusted by Achievement among Ethnic Samples

The differences on number of omits on the Mathematics Tes among the four ethnic examinee samples were tested using two analysis of covariance analyses designed to control for possible group differences in achievement level. As described earlier in the paper, the first ANCOVA used the following covariates: the number-right score to the first 47 items, the self-reported number of years of high school math, and the self-reported high school math grade point average. The second ANCOVA used the same covariates substituting an IRT achievement estimate for the number-right score.

For every ethnic sample, the number of cases was reduced because of missing or invalid values of the covariate variables. The number of remaining cases was 1,810 in the African origin sample, 1,901 in the Asian origin sample, 1,869 in the Caucasian origin sample, and 1,860 in the Hispanic origin sample.

Table 6. Absolute Value of Differences in Mean Omits between Samples and Results of the Scheffe Tests (whole samples)

	African Origin	Asian Origin	Caucasian Origin	Hisp. Origin
	N=1,810	N=1,901	N=1,869	N=1,860
African Origin	---	0.674*	0.898*	0.070
Asian Origin		---	0.224	0.787*
Caucasian Origin			---	1.011*
Hispanic Origin				---

* Indicating the difference between the two groups was significant at .05 chance level.

The ANOVA of the number of omits using the samples reduced in size indicated that the among-group differences were significant at P=.0001 on the Mathematics Test. And Scheffé's follow-up test results revealed that the chances for the existing difference in the number of omits to occur on the Mathematics Test were similar to \ hat were found in the whole sample analyses: not significant at P=.05 level between Asian origin and Caucasian origin samples and between African origin and Hispanic origin samples, but significant between all other ethnic origin sample pairs.

After the effects of three covariates were taken into consideration in the ANCOVA; the effect of the ethnic origin was still significant (P=.0001) in the results of both runs, one with the first-47-item raw scores (Table 7) and the other with the IRT achievement estimate based on responses on the same 47 items (Table 8). In the results of the ANCOVA using the first-47-item raw scores, the probabilities of the significance of the covariate effects were P=.0001 for first-47-item raw scores, P=.0003 for years of math courses, and P=.1158 for high math course grade. With IRT estimates as one of the three covariates, the significance levels of the covariate effects were P=.0001, P=.0029, and P=.3954, respectively.

Since the overall differences among the ethnic samples weic still significant after the number of omits were adjusted for the covariates, the Scheffé method was used in the follow-up tests to test the significance of differences in adjusted mean omits between all pairs of ethnic origin samples. For the mean omits adjusted for difference in the first-47-item raw scores and the other two covariates, the results of the tests identified that the differences were significant at .05 probability level between the African origin and the Caucasian origin samples, between the Caucasian origin and the Hispanic origin samples, and between the Asian origin and the Hispanic origin samples; but the differences were not significant between

Table 7. Math Achievement Estimate (First-47-Item Raw Score), Years of High School Math Courses Taken, High School Math Course Grade Average, and Mean Omits on Math Test before and after Adjustment by Ethnicity Samples

		Mean (Standard Deviation) of Covariate Variables			Mean Omits on Math Test	
Sample	N	First-47-Item Raw Score	Years of Math Course	High School Math GPA	Pre-adjusted	Adjusted
African Origin	1,810	19.11 (9.50)	3.49 (0.73)	2.56 (0.98)	1.338	1.129
Asian Origin	1,901	33.44 (10.17)	3.80 (0.48)	3.25 (0.87)	0.664	0.836
Caucasian Origin	1,869	29.11 (10.37)	3.66 (0.59)	3.02 (0.94)	0.440	0.510
Hispanic Origin	1,860	25.11 (10.65)	3.61 (0.64)	2.80 (0.99)	1.451	1.408
Prob. of Significance		.0001	.0003	.1158	.0001	.0001

Table 8. Math Achievement Estimate (First-47-Item IRT Score), Years of High School Math Courses Taken, High School Math Course Grade Average, and Mean Omits on Math Test before and after Adjustment by Ethnicity Samples

		Mean (Standard Deviation) of Covariate Variables			Mean Omits on Math Test	
Sample	N	First-47-Item IRT Score	Years of Math Course	High School Math GPA	Pre-adjusted	Adjusted
African Origin	1,810	-0.60 (0.82)	3.49 (0.73)	2.56 (0.98)	1.338	1.172
Asian Origin	1,901	0.61 (0.87)	3.80 (0.48)	3.25 (0.87)	0.664	0.803
Caucasian Origin	1,869	0.23 (0.84)	3.66 (0.59)	3.02 (0.94)	0.440	0.493
Hispanic Origin	1,860	-0.10 (0.87)	3.61 (0.64)	2.80 (0.99)	1.451	1.416
Prob. of Significance		.0001	.0018	.3954	.0001	.0001

the African origin and the Asian origin samples, between the African origin sample and the Hispanic origin samples, and between the Asian origin and the Caucasian origin samples (Table 9). The follow-up tests on mean number of omits adjusted using the IRT achievement estimates as one of three covariates produced the same between-sample significance results (Table 10) as those before controlling for the covariates effects as described earlier, that is, not significantly different at .05 probability level between the Asian origin and the Caucasian origin samples and between the African origin and the Hispanic origin samples, but significant between other sample pairs.

As shown in Table 10, after the number of omits were adjusted for the differences in the three covariate variables, the ranges of the adjusted mean omits with the first-47-item raw score as one of the covariates (0.898) and with the first-47-item IRT achievement estimate as one of the covariates (0.923) in the ANCOVA were both found to be smaller than the range of the mean omits before being adjusted (1.011). Together with the fact that the covariate effects in the ANCOVA were significant, this reduction indicated that, at least to a limited extent, the differences in mean omits among the ethnic samples were associated with differences in achievement level among the samples.

Discussion on Differences in the Tendencies to Omit Responses

Three relevant factors, examinees' math achievement level (estimated using their number-right scores and IRT scores on partial ACT Mathematics Test items), years of high school math courses taken, and high school math course grade, were used as covariates in the ANCOVA on the between sample differences in the number of omits. With these covariates controlled, the differences in the number of omits between gender samples and among ethnic

Table 9. Absolute Value of Differences in Mean Omits Between Samples Adjusted by Using First-47-Item Raw Score as One of the Covariates and Results of the Scheffe Tests

	African Origin	Asian Origin	Caucasian Origin	Hisp. Origin
	N=1,810	N=1,901	N=1,869	N=1,860
African Origin	---	0.293	0.619*	0.279
Asian Origin		---	0.326	0.572*
Caucasian Origin			---	0.898*
Hispanic Origin				---

* Indicating the difference between the two groups was significant at .05 chance level.

Table 10. Absolute Value of Differences in Mean Omits Between Samples Adjusted by Using First-47-Item IRT Score as One of the Covariates and Results of the Scheffe Tests

	African Origin	Asian Origin	Caucasian Origin	Hisp. Origin
	N=1,610	N=1,901	N=1,869	N=1,860
African Origin	---	0.366*	0.679*	0.244
Asian Origin		---	0.310	0.613*
Caucasian Origin			---	0.923*
Hispanic Origin				---

* Indicating the difference between the two groups was significant at .05 chance level.

samples should be less confounded with the possible differences in their knowledge of high school mathematics and thus be better indicators of differences in their tendencies to omit responses on the Mathematics Test.

The results of the different analyses on the difference of the number of omits between female sample and male sample were similar. The differences between female and male omit rates were very small on all the tests and for the overall test battery and the differences were not statistically significant. Nor were statistical differences found for the Mathematics test after controlling for the three covariates given above. Therefore, the null hypothesis that no difference exists between female and male examinees in the tendencies to omit items could not be rejected.

The results of this study relating to the tendency to omit between genders did not agree with what Grandy (1987) and Ben-Shakhar and Sinai (1991) reported based on their studies. Grandy's regression model included the variable gender as a statistically significant predictor of the number of omits on the GRE tests. Ben-Shakhar and Sinai concluded, after comparing the number of total omits, number of trailing omits, and two indexes for measuring guessing, Ziller's index (Ziller, 1957) and the modified Ziller's index (Angoff & Schrader, 1981), that males tend to guess more than females. However, in Grandy's study, the difference in mean omits between genders was less than 0.04 items per test on any of the three tests. The statistical significance of the factor being a predictor was based on a total sample of 55,656 examinees, of which 51.8% were females and 48.2% were males. With an average difference of 0.04 omitted items on a test, the practical meaningfulness of the results is questionable. Ben-Shakhar and Sinai used the number of omits and the original and modified Ziller's index measures between genders in their conclusions. As discussed in their report, the guessing indexes were based on the unrealistic assumption that every error made by the examinees was a result of pure guessing. Furthermore, although the value of the test statistics they computed based on the different guessing measures were all toward the direction which indicates males tend to guess more than females, the significance test results often did not agree among the different guessing measures used, i.e., while a difference in one guessing measure was found to be statistically significant, the other guessing measure based on the same data was often not. Other factors may also contribute to the differences in results of this study and others, such as the nature of the tests, the age of the examinee populations, the cultural and ethnic background of the examinee sample used.

For ethnic origin, the differences of mean omits between some pairs of samples differed statistically on both the battery and the Mathematics Test. Although no significance test was performed for the number of omits on the other tests, the numbers in omits showed the same trend.

Before adjusting for the covariates, the overall between group differences in the number of omits on the Mathematics test were statistically significant for both ANOVA using the whole sample data or the reduced sample data. After achievement level was controlled, both ANCOVAs (one used the first-47-item raw score and the other used the first-47-item IRT score as one of the covariates) among the ethnic groups produced results showing significant differences. However, the follow-up test results did not agree completely. Having adjusted for the covariates using two different partial test scores, the results of the significance tests on differences between sample groups all remained the same as those before except for the comparison between the Asian origin sample and the African origin sample. The significant differences between the two samples were only found in the mean number of omits adjusted using the IRT achievement estimate and not for that adjusted for the first-47-item raw scores.

It is hard to determine which set of results is more accurate. The above results could be interpreted as meaning that the differences in tendencie to omit responses were large enough to be statistically significant between some pairs of samples (such as between the African origin and the Caucasian origin samples, the Asian origin and the Hispanic origin samples, and the Caucasian origin and the Hispanic origin samples), but were too small to be significant between some other groups (between the African origin sample and the Hispanic origin sample, for example), and were just around the boarder line to have statistical significance between some other samples (like between the African origin and the Asian origin samples), Since the covariate variables used in the analysis were themselves estimated indicators of examinee achievement level in high school mathematics, the use of different covariates could cause the significance test results to vary. Thus, the outcome of the significance tests between some samples, between the African origin and the Asian origin samples, for example, could change in different analyses depending on the type and accuracy of the control variables used.

As the results suggested, the statistically significant differences in tendencies to omit responses likely exist among examinees with different ethnic origins, especially between the African origin and the Caucasian origin samples and between the Caucasian origin and the Hispanic origin samples. Grandy also reported ethnicity (white and non-white) as a statistically significant predictor of number of omits on the GRE tests, though the actual weights of the predictor, which meant differences of 0.03 to 0.06 omit on a test, were quite small.

While the results rejected the null hypothesis that the tendencies to omit responses are not different among examinees with different ethnic origins, the practical meaningfulness of the differences warrants discussion. With each sample size approaching 2,000 in the analysis of this study, small differences in a variable between samples could relatively easily result in statistical significance of the test statistics. The largest observed statistically significant between-sample differences in mean omits on the Mathematics Test among the ethnic samples were between the Caucasian origin and the Hispanic origin samples: 1.011 before any adjustment, and 0.898 and 0.923 after being adjusted for the two different sets of covariates accordingly. The statistically significant differences between other ethnic samples were all smaller. The values of the differences after the adjustment, when converted to average chance scores, all equal less than a 1/5 raw score point (there are five answer choices for each math item). Thus, although the average omit rate for some ethnic groups is greater than that for other groups, the difference seems to have little effect on the average scores for the groups. Still, some ethnic groups (especially African origin and Hispanic origin) have many more examinees that omit items compared to others. And, for certain individuals, omitting has a large effect on the score they receive.

Another issue is whether the adjustments made in the number of omits were enough to reflect each sample's true tendency to omit responses. As Huitema (1980) discussed, "even if the appropriate variables are included in the analysis, measurement error associated with the measurement of these variables will generally lead to an underadjustment of the means....It may *reduce* bias on *Y* that is predictable from the covariates, but it will generally not *eliminate* bias." The differences in the adjusted mean number of omits on the Mathematics Test among ethnic samples were all smaller than their corresponding differences in the mean number of omits before the adjustment. This supports the notion that the differences in number of omits were confounded by differences in achievement levels. When controlling for the selected covariate variables, the ANCOVA reduced the confounding in number of omits. However, the reduction in the differences in the number of omits was small. This indicates either that the covariates selected did not sufficiently account for differences in achievement level, or more likely, that differences in the number of omits between ethnic groups cannot largely be explained by differences in the groups' achievement levels.

CONCLUSION

The relationship between examinees' test scores and number of omits was negative and weak, with correlations ranging from -.11 to -.28. Although more examinees at the lower achievement levels omitted and they omitted more responses, some examinees at the high achievement level also omitted a surprising number of responses.

No significant differences were found in tendencies to omit responses between female and male examinees. Statistically significant differences in the tendencies to omit responses were found among examinees of different ethnic origins, especially between the African origin and the Caucasian origin samples and between the Hispanic origin and the Caucasian origin samples. However, the differences found between the ethnic groups were all less than one omit on average on the Mathematics Test.

Finally, controlling for potential achievement level differences between ethnic groups by the use of covariates resulted in only a minimal reduction in omitting tendency differences between the groups. Thus, our results indicate that differences between ethnic groups in the tendency to omit responses cannot solely be explained by differences in the groups' achievement levels.

REFERENCES

Abu-sayf, F. K. (1979). The scoring of multiple-choice tests: A closer look. *Educational Technology, 19(6)*, 5-15.

Albanese, M. A. (1988). The projected impact of the correction for guessing on individual scores. *Journal of Educational Measurement, 25(2)*, 149-157.

Angoff, W. H. and Schrader, W. B. (1981). *A Study of Alternative Methods for Equating Rights Scores to Formula Scores* (ETS Research Report No. 81-8). Princeton, NJ: Educational Testing Service.

Angoff, W. H. and Schrader, W. B. (1984). A study of hypotheses basic to the use of rights and formula scores. *Journal of Educational Measurement, 21(1)*, 1-17.

Ben-Shakhar, G. and Sinai, Y. (1991). Gender differences in multiple-choice tests: the role of differential guessing tendencies. *Journal of Educational Measurement, 28(1)*, 23-35.

Davis, F. B. (1967). A note on the correction for chance success. *The Journal of Experimental Education, 35(3)*, 42-47.

Ebel, R. L. and Frisbie, D. A. (1991). *Essentials of Educational Measurement, 5th Edition..* Englewood Cliffs, NJ: Prentice-Hall, Inc.

Frary, R. B. (1988). Formula scoring of multiple-choice tests (correction for guessing). *Educational Measurement: Issues and Practice, Summer*, 33-38.

Grandy, J. (1987). Characteristics of examinees who leave questions unanswered on the GRE General Test under rights-only scoring (GRE Board Professional Report No. 83-16P). Princeton NJ: Educational Testing Service.

Hays, W. L. (1988). *Statistics, 4th Edition.* New York, New York: Holt, Rinehart and Winston, Inc.

Huitema, B. E. (1980). *The Analysis of Covariance and Alternatives.* New York: John Wiley & Sons.

Lord, F. M. (1975). Formula scoring and number-right scoring. *Journal of Educational Measurement, 12(1)*, 7-11.

Mislevy, R. J. and Bock, R. D. (1990). *BILOG III user's guide* [Computer program manual]. Mooresville IN: Scientific Software, Inc.

Sabers, D. L. and Feldt, L. S. (1968). An empirical study of the effect of the correction for chance success on the reliability and validity of an aptitude test. *Journal of Educational Measurement, 5(3)*, 251-258.

SAS Institute, Inc. (1985). *SAS user's guide: Statistics* [Computer program manual]. Cary, NC: Author.

Slakter, M. J. (1968a). The penalty for not guessing. *Journal of Educational Measurement, 5(2)*, 141-144.

Slakter, M. J. (1968b). The effect of guessing strategy on objective test scores. *Journal of Educational Measurement, 5(3)*, 217-222.

Traub, R. E., Hambleton, R. K. and Singh, B. (1969). Effects of promised reward and threatened penalty on performance of a multiple-choice vocabulary test. *Educational and Psychological Measurement, 29*, 847-861.

Traub, R. E. and Hambleton, R. K. (1972). The effect of scoring instructions and degree of speededness on the validity and reliability of multiple-choice tests. *Educational and Psychological Measurement, 32*, 737-758.

Votaw, D. F. (1936). The effect of do-not-guess directions upon the validity of true-false or multiple-choice tests. *Journal of Educational Psychology, 27*, 698-703.

Wood, R. (1976). Inhibiting blind guessing: the effect of instructions. *Journal of Educational Measurement, 13(4)*, 297-307.

Ziller, R. C. (1957). A measure of the gambling response-set in objective tests. *Psychometrika, 22(3)*, 289-292.