

RACE, GENDER AND OMISSIONS ON STANDARDIZED ACHIEVEMENT TESTS

by

Robert L. Pour

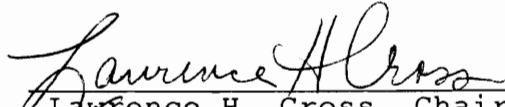
Dissertation submitted to the Faculty of the  
Virginia Polytechnic Institute and State University  
in partial fulfillment of the requirements for the degree of


DOCTOR OF PHILOSOPHY

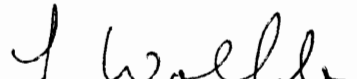
in

Educational Research and Evaluation

APPROVED:

  
\_\_\_\_\_  
Lawrence H. Cross, Chairman

  
\_\_\_\_\_  
Timothy E. Keith

  
\_\_\_\_\_  
Lee M. Wolfle

  
\_\_\_\_\_  
Harold W. Mick

  
\_\_\_\_\_  
Robert B. Frary

April, 1991  
Blacksburg, Virginia

RACE, GENDER AND OMISSIONS ON  
STANDARDIZED ACHIEVEMENT TESTS

by

Robert L. Pour

Committee Chairman: Lawrence H. Cross  
Educational Research and Evaluation

(ABSTRACT)

The purpose of this study was to examine the effects of race and gender on omissions in multiple choice tests.

Modest but significant ( $p \leq .05$ ) correlations were observed between gender and omissions and between race and omissions on the mathematics subtest of the Tests of Achievement and Proficiency, a standardized achievement test administered to all eleventh graders in the Commonwealth of Virginia. Item characteristics (difficulty, discrimination and an index of differential item functioning) were used as independent variables in regression equations for the male/female case and the black/white case. In both cases item difficulty was the only significant ( $p \leq .05$ ) predictor of omissions.

Principal components analysis was used to create composite variables characterizing school divisions. These composites together with race (proportion of black students) were used as independent variables in a regression equation with omissions as the dependent variable. Race was the only variable which was a significant predictor of omissions.

## Acknowledgments

I hereby express my sincere appreciation to my advisor, Lawrence H. Cross, for his generous help as my teacher, and for the invaluable criticism and encouragement he provided during the writing of this dissertation.

I also wish to thank Dr. Robert Frary, for his assistance in dealing with the data set used in this study and for many helpful suggestions throughout the course of my writing.

In addition I thank the other members of my committee: Timothy Z. Keith, Dr. Lee M. Wolfle, and Dr. Harold W. Mick for the valuable time they expended and for the constructive suggestions they provided.

The data for the study were provided by the Virginia Department of Education. I am grateful to Dr. David R. Mott for his help in this regard.

## Table of Contents

Chapter 1:	Introduction .....	page 1
Chapter 2:	Literature Review .....	page 12
Chapter 3:	Methodology .....	page 21
Chapter 4:	Results .....	page 29
Chapter 5:	Discussion .....	page 36
References	.....	page 58

## CHAPTER I

### INTRODUCTION

Whether to guess, or not to guess, has been a question of concern to test takers and test makers ever since multiple-choice tests were introduced about the turn of the century. In an effort to decrease the influence of guessing on multiple-choice test scores and increase public acceptance of these tests, alternatives to number-right scoring were developed. The most popular procedure, commonly referred to as the correction for guessing, uses a scoring formula which subtracts a fraction of the number of incorrect answers from the number of correct answers. Just as important as the formula are the associated directions which encourage examinees to refrain from guessing unless at least one choice can be eliminated from consideration. Although there have been more than 70 years of research and debate regarding the merits of formula scoring, there is little solid evidence to suggest that formula scoring is superior to number-right scoring.

Under number-right scoring, examinees should be encouraged to answer all questions, even if their answers represent sheer guesses. However, to provide such directions may seem inappropriate to those who fear that to encourage guessing will unfairly benefit the less well

prepared examinees. Nonetheless, it is clearly to every examinee's advantage to provide an answer to every question under number-right scoring. Failure to point this out forcefully in the test directions may benefit testwise individuals who will recognize this on their own and penalize naive examinees who may feel obliged to "confess" their lack of knowledge by omitting questions about which they are unsure.

The potential impact of less than clear directions regarding guessing on multiple-choice came to light when data from a statewide testing program were being analyzed for another purpose. Riverside Publishing Company's Tests of Achievement and Proficiency (Scannell, 1986) are administered to all eleventh grade students across the Commonwealth of Virginia. Inspection of the data for 1988 revealed that in some school divisions there were large numbers of omitted items, and in other school divisions there were virtually no items omitted. Because number-right scoring was used, there should have been virtually no omissions. Moreover, higher rates of omissions were observed in school divisions having higher percentages of black student enrollment. Inspection of the test directions showed that examinees were not specifically advised to guess when unsure of an answer. Indeed, the directions were more nearly like directions appropriate for formula scoring.

These observations gave rise to the present study. At issue is not the less than adequate directions, per se, but whether the omissions possibly precipitated by these directions were race or gender related.

There is compelling evidence that omitting responses to items on a multiple-choice test is disadvantageous, regardless of how that test is scored. This harm is unavoidable if the test is scored number-right. If black examinees omit proportionally more than white examinees, the typically lower level of performance of black examinees is exacerbated by these omissions. If female examinees omit proportionally more than male examinees, test results may not accurately measure the mathematical achievement of female students. Results of standardized tests tend to have enduring effects on both individuals and on groups of students. Test scores are used to determine individual eligibility for scholarships and special programs, while programs themselves may be developed or implemented as a result of group performance. For these reasons, test validity must be of prime concern to test developers and test users. If high omissions are associated with group membership, validity of the test will be compromised.

This study investigated multiple-choice test omissions at three basic levels: the individual examinee level, the item level, and the school division level. At the

individual examinee level this question was asked: did black examinees omit more frequently than white examinees, and did female examinees omit more frequently than male examinees? At the item level, item characteristics were examined which could serve to explain the incidence of omissions for particular types of items. Finally, at the school division level, demographic features of the school that might contribute to an explanation of differing rates of omission among school divisions were examined. Omissions that occurred on items which were reached by the examinee are of primary interest. Items are classified as having been reached if they are followed by answered items; such omissions are referred to as embedded omissions. Trailing omissions (which are not followed by answered items) were studied only at the division level.

#### THE TESTS OF ACHIEVEMENT AND PROFICIENCY

The State of Virginia, as a component of its educational assessment program, requires all school divisions to administer Riverside Publishing Company's Tests of Achievement and Proficiency (TAP) to all eleventh graders. The complete battery consists of subtests on reading comprehension, mathematics, written expression, using sources of information, social studies, and science. Institutions use scores to make decisions about individuals,



in addition scores on the TAP are often used as a yardstick for comparing school divisions in terms of academic performance. Funding for remedial programs is also linked to school division performance on the TAP. The percentage of students in the bottom quartile is used as one measure of the need for remedial programs. Hence, unnecessarily high rates of omissions may have a detrimental effect on both school divisions and individuals.

In The Ninth Mental Measurement Yearbook, Keene (1985) states that the purpose of the TAP is to "provide efficient and comprehensive appraisal of student progress toward widely accepted academic goals in the basic skill areas" (p. 1611). The main emphasis of the tests is on the application of knowledge and skills rather than specific content. Content validity is considered quite adequate for these purposes. Furthermore, item response theory techniques were used to assess ethnic and gender bias of individual items. The reliability coefficient of the tests are reported to be at least .82 based on KR-20 estimates. Although the TAP was reported to be a relatively easy test to administer, no discussion of examinee instructions was given, nor was there mention of the examinee samples used to test for race and gender bias.

This study focuses on the mathematics subtest of the TAP. The analysis of the item data for this subtest

revealed omission rates which varied greatly across items, ranging from 0.01% to 16%. Across divisions, embedded omissions ranged from means of close to zero omissions to 1.4 omissions per examinee. Such variation may suggest faulty test instructions.

#### TEST INSTRUCTIONS

Individual test taking behavior is likely to be influenced by test directions. Hence, directions may have significant effects on test outcomes. The "Student Directions" given at the beginning of the TAP (p. 2) were as follows (*italics added*):

#### EARNING YOUR BEST SCORE

Some students receive lower scores on tests than they could receive, simply because they do not take the test in the most efficient manner. The information below is provided to help you earn your best score.

As you take the test, remember these points

1. If you are not absolutely sure about the answer to a question, but think you know the correct answer, mark a choice. You will earn your best score if you attempt all questions *for which you think you know the answers*. You will not lose any points for incorrect choices.
2. There are some questions on each test which you may not be able to answer. Do not linger over difficult questions; omit these and go on to easier ones. You *may* return to omitted questions at the end of the test *if there is time remaining*.

Since the TAP was scored using number-right scoring, the optimal score would be earned by answering all questions whether or not one knew the answer. A critical examination of the instructions presented above reveals a failure of the test publisher to explain clearly the best test-taking strategy. The above instructions leave doubt as to whether or not to omit items. The instruction identified above as 1 states that examinees also should attempt items for which they think they know the correct answer. However, it does not specifically state what examinees should do if they do not think they know the answer. In fact, it is to the examinee's advantage to answer all questions. The problem is further exacerbated by the instructions identified above as 2. Here the directions suggest omitting difficult items and returning to them only if there is time remaining. These instructions are more nearly in keeping with and indeed bear a striking resemblance to directions appropriate to tests using the correction for guessing.

The SAT is a test which is scored using correction for guessing. One of the test-taking tips in Taking the SAT (College Entrance Examination Board, 1990) states (p.6):

You can omit questions. Many students who do well on the SAT omit some questions. You can always return to questions you've omitted if you finish before time is up for that section.

Other test-taking tips given make perfectly clear the best strategy, given the scoring method. However, there is unquestionable ambiguity in the TAP directions as to whether guessing is advisable. Examinees taking a test scored by number-right scoring should receive unequivocal instructions to answer *all* items. Any other instructions may lead to misguided test behavior and a high rate of omissions, as apparently was the case for the TAP mathematics subtest to be analyzed in this study. Moreover, not only were omissions widespread, but they were generally highest among examinees in divisions with high percentages of black enrollment. On a number-right scored test, it is naive to omit any item, and this act is detrimental to any examinee.

#### STANDARDS FOR TEST INSTRUCTIONS

The Code of Fair Testing Practices in Education prepared by the Joint Committee on Testing Practices (1988) of the American Psychological Association, the American Educational Research Association, and the National Council on Measurement in Education offers guidelines for both test developers and test users. Item 18 of the code, relating to informing test takers, states (p. 2):

Test Developers or Test Users should provide test takers the information they need to be familiar with the coverage of the test, the types of question

formats, the directions, appropriate test-taking strategies, and strive to make such information equally available to all test takers.

The disregard for these guidelines in TAP directions could have served to increase the number of omissions, a situation which warranted investigation of possible patterns of omissions.

#### BLACK AND FEMALE STUDENTS AND MATHEMATICS

Mathematics, perhaps more than any other subject in the American curriculum, has been an area of poor performances by black students. Although mathematics is relatively free of the cultural artifacts that might estrange the black student, blacks tend to be under-represented in mathematics classes. This phenomenon begins in high school and worsens at higher levels of education. The TAP mathematics subtest, however, is sufficiently general so as to lessen the impact of differential coursework as an explanation for test results.

Ben-Shakhar & Sinai (1991) found that while females tend to earn better grades than males in high school mathematics classes, their scores on standardized mathematics tests tend to be lower and their rates of omission higher. Although female enrollment in mathematics classes decreases at higher levels, differential coursework

should not be a factor in a test as general in scope as the TAP.

More to the point of this study, omissive behavior

- a) may be greater for black examinees than for white examinees,
- b) may be greater for female examinees than for male examinees,
- c) may vary with item characteristics, such as item bias, item difficulty, and item discrimination, and
- d) may relate to characteristics of the school division.

The purpose of this study is to examine the extent to which the above relationships are evident in the TAP Mathematics subtest.

#### HYPOTHESES

A correlational analysis was used at the individual level to investigate the hypothesis that black examinees had higher rates of embedded omissions than did white examinees and that female examinees had higher rates of embedded omissions than did male examinees. The large number of omissions in divisions with larger percentages of blacks initially motivated the hypothesis regarding a difference for black and white examinees. Multiple regression was the

primary tool in the investigation of the relationship between item characteristics, race/gender, and omissions. At the school division level, this study investigated the hypothesis that there is a systematic relation between the average number of embedded omissions, trailing omissions, total omissions, and demographic variables such as size of school division, per capita expenditures, dropout rate, etc. It was expected that financial support, size, racial makeup, and ability of the student body might be significant factors in predicting rates of omission for a school division, since these factors had been observed to be significant predictors of test scores.

## CHAPTER II

### LITERATURE REVIEW

When examinees do not know the correct response to a multiple-choice item they may either guess or omit. If the test is scored number-right, it is clearly to the examinee's advantage to guess even if completely ignorant with respect to a test item. Formula scoring was introduced in the 1920s to discourage guessing. The correction formula for guessing is:  $\text{Corrected score} = R - W/(n-1)$ , where  $R$  is the number of items right,  $W$  is the number wrong, and  $n$  is the number of alternative responses per item. While this is not a study of formula scoring versus number-right scoring, a consideration of some studies involving formula scoring may serve to shed light on test-taking behavior on the TAP. Clearly, test-taking strategy is influenced by test directions that explain the scoring method. Hence, an understanding of the effect of scoring directions is germane to the study of omissions.

### STUDIES OF THE EFFECT OF SCORING METHOD ON TEST RESULTS

Why should guessing be discouraged? Formula scoring was apparently introduced for moralistic, not technical, reasons. The argument against guessing was based on a value judgment; specifically, the examinee who guesses is trying



to deceive the examiner and should be penalized. This approach was defended by maintaining that the technical arguments for correction for guessing were compelling.

Rowley and Traub (1977) point out that the primary reasons for correction for guessing via formula scoring were the improvement of the psychometric properties of the test scores through a decrease in error variance. Although theoretical studies of Mattson and Lord (1988) argued that formula scoring increases reliability and validity, empirical studies, more often than not, show small differences.

Test instructions that tend to discourage guessing have been shown to favor assertive (vs. submissive) personalities (Wiley & Trimble, 1936). Thus, reliability is potentially increased at the expense of validity, since reliable, but incidental, personality traits are influencing the scores.

Traub and Hambleton (1972) studied the effects of scoring instructions on the validity and reliability of multiple-choice tests. Their results demonstrate that offering a small reward for not guessing, rather than imposing a penalty for guessing, is more effective in reducing completely random guessing among college students. They also contend that instructions encouraging guessing, offering a reward for not guessing, or imposing a penalty for guessing affect the extent to which various personality

types fail to guess when they have eliminated an option or have a hunch. The TAP directions virtually invite the examinee to decide whether to guess or not, thereby confounding the validity and reliability for various groups. A TAP examinee with a low level of testwiseness could possibly interpret the instructions as providing a penalty for guessing.

Rowley and Traub (1977) investigated the relationship between scoring method on the test and test-taking strategy by administering the same test with differing statements as to method of scoring. Examinees were asked to indicate which of three guessing strategies would be best for the test. The findings indicate that number-right scoring possesses the advantage that results were not influenced by personality characteristics of examinees. Wiley and Trimble (1936) found that confidence in responses influenced scores, while Sherrifs and Boomer (1954) found that a penalty for guessing penalized students who are "characterized by introversion, rumination, anxiety, low self-esteem, and undue concern with the impression they make on others" (p. 82). Indeed one might expect low self-esteem among many blacks in a test-taking situation given the record of standardized test performances of blacks and whites.

Lord (1975) assumed that the difference between the answer sheet scored with number-right and formula scoring

(correction for guessing) is that the blanks on the latter are filled with random responses on the former. This assumption is equivalent to assuming that all guessing on multiple-choice tests is random. Rowley and Traub (1977) note that many examinees fail to differentiate between random guessing and informed guessing; some examinees exhibit a characteristic of testwiseness while others do not. That is, some examinees fail to answer items for which they possess at least partial knowledge. Other investigators, Cross and Frary (1977), Cureton (1966), and Rowley and Traub (1977), criticize directions against guessing on the grounds that students' use of partial knowledge, on items previously omitted under correction for guessing directions, results in a better than random chance of a correct response. Thus, guessing appears to be advisable for both formula scored and number-right scored tests.

The possession of partial information may influence an examinee's guessing strategy, and this influence may differ across ability levels. Angoff and Schrader (1984) proposed the *Invariance Hypothesis*, namely that formula scores are invariant with respect to guessing strategy. They offered empirical evidence, some of which tended to support the invariance hypothesis. Angoff and Schrader argue that partial information may offer a delusive advantage. This is

to say that partial information may lead the examinee to choose distractors. An obvious question is whether or not the disadvantage is greater for one racial group than for another. Angoff (1974) conducted an empirical study which suggested that partial information may help higher ability students while hindering lower ability students who tend to be fooled by distractors.

In contrast to the *Invariance Hypothesis*, the *Differential Effects Hypothesis* maintains that when using formula directions, particular students will omit items that they have a greater than chance probability of answering correctly. Albanese (1988) studied the effect of formula scoring on individual scores. An important finding was that both partial information and misinformation could significantly affect scores. When formula scoring instructions were followed, it was found that examinees who had omitted 15% of the items would experience a .5 SD increase in their score if they answered those omitted items with a .55 success rate. Thus a disadvantage is imposed on the cautious student by a penalty for guessing. Hence, following instructions not to guess would lower test scores.

Fischer (1988) investigated the effect of instructions for guessing on multiple-choice test performance for fifth-graders. A significant effect was found for children's ability to understand the penalty/reward instructions for

the test. More accurately it could be said that testwiseness was a significant factor of test performance. Although Fischer's subjects were fifth-graders, it is not implausible to expect similar results with eleventh graders. The ability to understand test instructions is likely to influence test performance. Fischer found that cautious behavior led to an increase in trailing omissions, those omissions following the last answered item. Angoff and Schrader (1984) found that there are 29.9% more trailing omissions for groups given formula scoring directions than for those given number-right directions.

#### PERFORMANCE OF BLACKS ON STANDARDIZED MATHEMATICS TESTS

Dossey, Mullis, Lindquist, & Chambers (1988) report that although black students have made significant advances in mathematics test performance during the past 13 years, there is still a considerable gap in performance of black and white students at ages 9, 13, and 17 years. The reasons for these differences are extremely complex social problems. I have not attempted to address the major causes of this performance differential, but argue that the small portion of this difference which is due to omissions is the most easily remedied.

## GENDER AND OMISSIONS

Much research has been conducted on gender differences in the area of mathematical ability and achievement. Maccoby and Jacklin (1974) provide a review of the literature which indicates a gender difference in the areas of visual-spatial and mathematical ability of between .4 and .5 SD at the high school level.

Ben-Shakhar and Sinai (1991) studied gender differences in omissions on multiple-choice tests. Citing McManis and Bell (1968), they argue that there is a greater tendency for risk taking among high school boys than girls. This risk-taking tendency by male students resulted in an increased rate of guessing by male examinees, a tendency which was not altered by permissive test directions.

## METHODS FOR DETECTING DIFFERENTIAL ITEM FUNCTIONING

The literature abounds in studies focusing on differential item functioning (DIF). They fall roughly into two categories: first, those that compare various methodologies used to detect DIF; and second, those that apply one method of detecting DIF in a particular test, and seek explanations for the presence of DIF in those items found to exhibit DIF. Comparative studies of item bias methods tend to apply several bias detection methods to one data set and seek to evaluate their findings in terms of

sample sizes required, cost of the procedure, time required to analyze data, and other such criteria. ANOVA, chi-square, item characteristic curves, delta plots, the standardization approach, and log linear models have all been used as tools to detect differential item functioning. Comparative studies of Perlman (1988), Dorans and Kulick (1986), Hambleton and Rogers (1989), and Camilli and Smith (1989) lead to the conclusion that for data sets of moderate size, there are only small differences in items found to exhibit DIF when using different methods of estimation.

Kulick and Hu (1989) point out that the Mantel-Haenszel item bias test is a chi-square method which has the decided advantage of being readily available on several mainframe statistical packages. Unlike the item characteristic method, it is non-iterative, thereby costing less to compute on large data sets. An additional advantage of the Mantel-Haenszel technique is the index of bias which is provided. Beck (1982) points out that earlier chi-square methods studied DIF by utilizing a matching criterion of three to five score intervals to classify examinees by ability level ability. Using so few score intervals tends to confound differences in true ability with any difference in functioning that item may exhibit. Researchers now recommend that each score level be used to classify individuals. With the Mantel-Haenszel approach, the large

number of ability levels does not significantly increase the cost or the time of analysis.

Kulick and Hu (1989) examined the relationship between differential item functioning and item difficulty on the Scholastic Aptitude Test. The significant negative correlation between DIF and item difficulty was found to be independent of the index of bias (the Mantel-Haenszel or the standardization approach). That relationship was strong for each racial and ethnic group (black, Hispanic, and Asian American), and there was a stronger relationship between differential item functioning and item difficulty on the verbal section than on the mathematical section. The more difficult items tended to favor the black focal group over the white reference group.

Zwick and Ercikan (1989) used the Mantel-Haenszel approach to analyze the NAEP history assessment for differential item functioning. Conditioning was done on score, and score plus historical periods studied, for a focal group of blacks and a reference group of whites. It was discovered that additional conditioning on historical periods studied did not decrease the number of items shown to exhibit differential item functioning. The use of NAEP sampling weights had no significant effect on DIF detected by the Mantel-Haenszel procedure.



## CHAPTER III

### METHODOLOGY

#### THE DATA

The data analyzed for the study were responses for approximately 67,000 examinees to the TAP mathematics subtest. The sample included virtually all Virginia 11th graders for the 1987-88 academic year. Students whose total score was less than what would be expected by chance alone (approximately 3500) were excluded from the study.

Embedded omissions are items that are omitted though they have been reached by the examinee. The rationale for examining embedded omissions is that an embedded omission occurs for reasons other than a lack of time to reach the item. In the case of an embedded omission, the item has been intentionally omitted after it presumably has been considered.

#### LEVELS OF ANALYSIS

Since the larger number of omissions were in divisions with greater percentages <sup>in</sup> of black students, I first examined the correlation between the number of embedded omissions and race. The Pearson-product moment correlation between embedded omissions and black/white group membership was calculated and tested for significance. Similarly, the

correlation between gender and embedded omissions was examined. The second stage of analysis made use of multiple regression to determine whether the rate of omission for individual items was a linear function of item difficulty, item discrimination, and differential item functioning. The analysis at the school division level made use of principal components analysis followed by multiple regression to model school division level omissions.

It is assumed that the reader is familiar with the methodology of multiple linear regression. The reader wishing a detailed explanation may consult Pedhazur (1982). A complete presentation of the Mantel-Haenszel procedure will be given and some aspects of principal components analysis will be presented. A detailed presentation of the methods of principal components analysis may be found in Tatsuoka (1988).

#### THE MANTEL-HAENSZEL PROCEDURE

When considering the problem of DIF, the performance of one group of interest, called the focal group  $F$ , is compared to another group of interest called the reference group  $R$ . The reference group may be regarded as the benchmark against which the focal group is compared. Performance of each group is compared on each test item to ascertain whether items function differently for the two groups. An essential

aspect of this comparison of item performance is that only like groups of examinees are compared. That is to say, for a comparison of item performance to be meaningful, groups must be matched on attributes which relate to the characteristic being measured by the item in question. Although there may be numerous attributes of an examinee which relate to item outcome, total score is most often chosen as the conditioning variable. For an unidimensional test, the total score is another measure of the construct that an individual item measures; in addition, it is readily available.

To aid in the explanation of the methodology involved in the detection of differential item functioning, it is useful to introduce tables summarizing performance of the two groups on an item being investigated. Table 1 symbolizes information for a matched group on a particular item.

---

Insert Table 1 about here

---

In Table 1  $A_j$  represents the number of members of the reference group answering an item correctly while  $B_j$  represents the number of members of the reference group answering the item incorrectly.  $C_j$  is the number of the focal group answering the item correctly, while  $D_j$  is the

number of the focal group answering the item incorrectly. The total number in the reference group and focal group are given by  $n_{Rj}$  and  $n_{Fj}$  respectively, while the total number of examinees in the  $j^{\text{th}}$  matched group is given by  $T_j$ . Now  $n_{1j}$  and  $n_{0j}$  represent the total number of correct and incorrect responses for the  $j^{\text{th}}$  matched group.

Table 2 presents a similar table indicating population parameters for data on the studied item.

---

Insert Table 2 about here

---

The Mantel-Haenszel Procedure is a non-parametric, non-iterative contingency table method used to estimate and test the association between two factors for  $k$  matched groups. It was developed by Mantel and Haenszel (1959) to study dichotomous outcomes of medical treatments for matched groups of patients. The Mantel-Haenszel procedure tests the null hypothesis of equal odds of success on a given item, symbolized as:

$$H_0 : p_{Rj}/q_{Rj} = p_{Fj}/q_{Fj} \quad j = 1, \dots, K.$$

verses a specific alternative hypothesis

$$H_1 : p_{Rj}/q_{Rj} = \alpha (p_{Fj}/q_{Fj}) \quad j = 1, \dots, K. \text{ for } \alpha \neq 1. \text{ We can see}$$

that the null hypothesis is equivalent to  $\alpha = 1$  in  $H_1$ . We call the parameter  $\alpha$  the common odds ratio in the  $K$   $2 \times 2$

tables since

$$\alpha = \frac{p_{Rj}}{q_{Rj}} / \frac{p_{Fj}}{q_{Fj}} = (p_{Rj} q_{Fj}) / (p_{Fj} q_{Rj}) \quad \text{for all } j = 1, \dots, K. \text{ The}$$

Mantel-Haenszel chi-square is calculated as:

$$(\sum_j (A_j - \sum_i E(A_j))^2) / \sum_j \text{Var}(A_j) \quad \text{where } E(A_j) = (n_{Rj} m_{1j}) / T_j.$$

Note that this is a *conditional expectation* given the marginal totals in Table 1. The variance of  $A_j$  is given by:  $\text{Var}(A_j) = (n_{Rj} n_{Fj} m_{1j} m_{0j}) / (T_j^2 (T_j - 1))$ . The Mantel-Haenszel chi-square statistic is often given a continuity correction given by:  $MH-CHISQ = ((|\sum_j A_j - \sum_j (E(A_j))| - \frac{1}{2})^2) / \sum_j \text{Var}(A_j)$ .

An important advantage of the Mantel-Haenszel method of detecting DIF over log linear models, or even the IRT models, is that it provides an estimate of the common odds ratio across the  $K$   $2 \times 2$  tables. The estimator of  $\alpha$  is given by:  $\hat{\alpha}_{MH} = (\sum \frac{A_j D_j}{T_j}) / (\sum \frac{B_j C_j}{T_j})$ . An odds ratio of 1 indicates

no DIF and is equivalent to the null hypothesis. Holland and Thayer (1985) suggest a log transformation to convert this ratio into a symmetrical scale. This transformation is given by:  $\Delta_{MH} = -2.35(\ln(\hat{\alpha}_{MH}))$ . This scale is referred to as the ETS "delta scale" where the optimal range of delta is

observed to be  $1/3 < \alpha < 3$  , which is equivalent to  $-2.6 < \Delta_{MH} < 2.6$ . Variance estimates of  $\Delta_{MH}$  have been developed by Breslo (1981), Hauck (1979), Flanders (1985) and Phillips (1987). Holland and Thayer (1988) suggest the estimate given by Phillips:

$$\text{Var}(\Delta_{MH}) = 1/2U^2 \sum_j [T_j^{-2} (A_j D_j + \hat{\alpha}_{MH} B_j C_j) (A_j + D_j + \hat{\alpha}_{MH} (B_j + C_j))]$$

Where  $U = \sum_j (A_j D_j) / T_j$ .

Although MH-CHISQ is a test statistic and  $\Delta_{MH}$  can be tested for being significantly different from zero, the application of this index of DIF has no statistical criterion for application to individual items. Its use requires a judgment as to what level of DIF is large enough to exclude an item. Zwick and Ercikan (1989) cite rules developed by ETS for interpreting the Mantel-Haenszel DIF index.

- i.) Items with an index not significantly different from zero ( $\alpha = .05$ ) or which have an absolute value less than 1 are considered free of DIF.
- ii.) Items which exhibit an index significantly different from zero and have an absolute value between 1 and 1.5, or an absolute value of at least 1 but not significantly greater than 1, should be considered as candidates for replacement, provided there is a sufficient step of items with a smaller DIF index.

- iii.) Items which exhibit a DIF index of which has absolute value greater than 1.5 and where the index is significantly greater than 1 should only be used in extreme cases.

The DIF analysis of this data will be done using SAS PROC FREQ with the Mantel-Haenszel chi-square option and an  $\alpha$  level of .05. The log transformation to the delta scale was done as a hand calculation.

#### DIVISION LEVEL ANALYSIS

The hypothesis that the level omission rate at the school division level is a function of demographic variables was investigated via multiple regression. I first applied principal components analysis to 19 division level variables, followed by varimax rotation to reduce the number of variables involved in the regression analysis. A four-factor solution was chosen based upon the fact that four eigenvalues were greater than 1. The principal components analysis was used as the basis of a simple index construction suggested by Kim and Mueller (1978). This construction consists of "summing all the variables with substantial loadings and ignoring the remaining variables with minor loadings. The scale created in this way is no

longer a factor scale but merely factor-based ... " (p. 70). The variables were first standardized before the index was constructed.

The testing directors of the two school divisions with virtually no trailing omissions were contacted by telephone. They indicated that students were instructed to answer all items. Since students in these divisions received special instructions, they were excluded from this study.



## CHAPTER IV

### RESULTS

#### RESULTS AT THE INDIVIDUAL EXAMINEE LEVEL

The correlation of embedded omissions with race was computed to be  $-0.057$  with a  $p$ -value of  $0.0001$  for  $n = 57877$ . Since white students were coded with a 2 while black students were coded with a 1, this correlation indicates that black examinees tend to have slightly, but statistically significant, higher levels of embedded omissions. The correlation of embedded omissions with gender resulted in a correlation coefficient of  $-0.033$  with a probability value  $< .001$  for  $n = 63915$ . Although female examinees tended to have higher levels of embedded omissions than male examinees, the gender/omission correlation was extremely small.

#### ITEM LEVEL RESULTS

Embedded omissions on the TAP mathematics subtest items varied from a .01% omission rate to almost 16% over individual items. Standard procedures were used to compute item difficulty (proportion answering an item correctly), item discrimination (point-biserial correlation between the

item score and the total score), and  $\hat{a}$  (the population estimate of a correct response likelihood ratio for the reference group to the focal group);  $\hat{a}$  was also computed for each item. From this  $\Delta_{MH}$  was calculated. In addition, the average number of embedded omissions for each item were tabulated. Correlations between these four variables are reported in Table 3 for the black/white student comparison and in Table 4 for the male/female comparison.

---

Insert Table 3 about here

---

Two of the correlations reported in Table 3 are relatively large in comparison to the others. The correlation between the percentage of omissions and the DIF index delta was -0.280, but was not significant ( $p > .05$ ). The correlation between item difficulty and percent omissions was -0.5432 and significant ( $p < .05$ ). The relation between difficulty and embedded omissions is to be expected since, as item difficulty increases, omissions increase.

In Table 4, which shows the male/female comparison, the correlation between  $p$ -value with the percentage of embedded omissions is relatively large in comparison with the other correlations reported. The correlation between delta and

embedded omissions is  $-0.044$ , a considerably smaller correlation than for the black/white comparison.

---

Insert Table 4 about here

---

The mean embedded omissions for each item was then regressed on item difficulty and item discrimination, and  $\Delta_{MH}$  for both a black/white as well as a male/female comparison.

Table 5 shows the results of regressing embedded omissions on the other three item characteristics for the black white comparison. As indicated by the  $t$ -values, only item difficulty made a significant ( $p < .001$ ) contribution to the equation, accounting for approximately 30 percent of the variation in omissions.

---

Insert Table 5 about here

---

It appears, therefore, that although the delta values are significantly correlated with percentage of omissions as shown in Table 3, the use of delta values does not improve the prediction of percentage of embedded omissions over and above the prediction possible on the basis of item

difficulty alone. Table 6 shows the results of regressing embedded omissions on the other three item characteristics for the male/female comparison.

---

Insert Table 6 about here

---

Once again the results show that only item difficulty makes a significant contribution to the prediction of embedded omissions, with item difficulty accounting for approximately 24 percent of the variation in omissions.

#### DIVISION LEVEL RESULTS

An observation of striking interest is the wide variability in omissions across school divisions. The following tables summarize the omission rates for divisions with more extreme values.

---

Insert Tables 7 and 8 about here

---

The rates of omission range from a low of 0.09 omissions per student in Bath County to 1.36 omissions per student in Madison County. Those cities and counties with a higher

percentage of black students tended to have higher omission rates than those that were predominately white divisions.

A data set was created in which the principal unit of analysis was the school division. Data for a 137 divisions were analyzed. Variables included were total enrollment, end of year enrollment, pupil teacher ratio, average teacher salary, percentage promoted to the 9<sup>th</sup> grade, TAP scores (mathematics, reading, written expression, using sources of information, social studies, and science), percentage of 9<sup>th</sup> graders graduating, percentage of students continuing their education, percentage dropouts 8-12, total population in the school division, local composite index (ability to pay), local percent contribution to cost of schooling, local amount per pupil, total expenditure per pupil, racial composition (percentage black), and rate of omissions (embedded, trailing, and total). All but the last four data items were taken from Facing Up-23 published by the Virginia Department of Education (1988). The means, standard deviations, and intercorrelations among all of these variables are presented in Appendix A.

The rotated factor pattern loadings of the principal components analysis are presented in Appendix B. The highlighted loadings indicate the variables defining each of

the four factors. All variables included in the analysis were standardized, and four composite variables were created by summing the standardized values for variables loading on a factor with a loading of .4 or greater.

A multiple regression of mean number of embedded omissions on composite variables 1 through 4 and percentage of black enrollment yielded the results shown in Table 9.

---

Insert Table 9 about here

---

Testing the model indicates that only the percentage of black enrollment was a significant ( $p < .05$ ) predictor of mean embedded omissions per division.

Separate multiple regressions of trailing omissions and total omissions on level composite variables 1 through 4 and percentage of black enrollment produced results similar to the regression involving embedded omissions.

---

Insert Tables 10 and 11 about here

---

Again, the percentage of black enrollment was the only significant predictor of trailing omissions and of total omissions. These results differ from the hypothesized

results of an effect due to composite variables 1, 2, 3 and 4. The following chapter considers the discrepancy between the results expected and those obtained.

## CHAPTER 5

### DISCUSSION

The central question of this study is whether or not embedded omissions on the TAP were systematic. That is to ask, were omissions related to group membership? The item characteristics: item difficulty, item discrimination, and DIF index were expected to be significant predictors of rate of omission. This relationship was anticipated for the black/white comparison as well as for the male/female comparison. This type of result would be consistent with the research of Kulick and Hu (1989). However, regression analysis demonstrated only item difficulty was a significant predictor of rate of embedded omissions on an item. With a correlation of  $-0.280$  between  $\Delta_{MH}$  (black/white) and embedded omissions, it is worth mentioning that a high DIF index is associated with low omission rates. While more difficult items were more frequently omitted, the study of Kulick and Hu (1989) found that the more difficult items sometimes tended to favor the focal group, the black examinee. Hence if this were true for the TAP, omitting could sometimes serve to penalize black examinees more than their white counterparts. Items which exhibit significant DIF are



functioning differentially for matched groups whose scores are in the middle of the score range much more than those at the extremes. That is to say that the differential functioning is not uniform. Although, overall, blacks had a higher rate of omission than did whites, there is no evidence from this study to suggest that omissiveness is associated with DIF. However, there is sufficient variation at the division level to warrant division level investigation.

The rate of embedded omissions for a test on which no items should be omitted represents a measure of testwiseness. My hypothesis for division level data was that larger, predominately white, higher ability, financially strong divisions contained more sophisticated test takers, and hence would tend to have fewer embedded omissions than their counterparts in smaller, less homogeneous, lower ability, financially weaker divisions. What interpretation can be made of the resulting significance of only the race variable in this model? The size of a school or of a community appears to contribute no decided advantage in terms of this aspect of testwiseness. Rate of embedded omissions is not related to size of the school division, nor are embedded omissions related to the

composite variable designated achievement (see Table 9). Although higher test scores tend to be associated with larger schools, embedded omissions appear, at the school division level, not to be significantly related to the composite achievement variable. Focusing on mean scores for divisions, however, obscures the fact that at the individual level, embedded omissions are significantly correlated with score. Hence, individual test outcomes can be affected by omissions.

The financial support composite variable also failed to be a significant predictor of embedded omissions. Larger school divisions tended to spend more per pupil, to be more able to contribute to the total amount spent, and to pay teachers higher salaries. None of these school characteristics, however, helped to explain behavior related to omissions.

Although the composite variables incorporate a great deal of division level information, they also are subject to severe limitations. Size alone fails to encompass many school characteristics, characteristics that might be important in explaining test taking behavior. An examination of the five divisions with the highest and lowest levels of omissions (Tables 6 and 7) reveals no clear

pattern of attributes, aside from race, which serve to explain even these extreme cases.

The only teacher-related variables which were available in this data set were pupil-teacher ratio and average salary. It is more likely that information such as teaching experience and education would be useful in modeling omissions. Variables which relate to parental influence may be important but are unavailable in this data set, as are measures of student attitudes.

Although the subject matter of the test is sufficiently broad, data on individual coursework, or when mathematics was most recently taken, would also be potentially valuable information. Division level data is perhaps too broad to capture the subtleties which relate to test-taking behavior.

The significance of race in predicting omissions should not be ignored by test publishers, test users, teachers and guidance counselors. Omission of reached items contributes to the already lower scores of blacks, a fact which suggests that a focus on test-taking strategy for blacks would produce scores which more realistically reflected achievement of that group. The ambiguity of the TAP directions is a violation of the spirit of The Code of Fair Test Practices in Education (Joint Committee on Testing

Practices, 1988). A clarification of test taking directions for the TAP is unquestionably the obvious first step in ensuring that omissions will not cloud future test results, even if bias was found to be unrelated to omissiveness in this study.

Table 1

Item Summary for the  $j^{th}$  Matched

Reference and Focal Group

		Item Score		
		1	0	
<b>Group</b>	R	$A_j$	$B_j$	$n_{Rj}$
	F	$C_j$	$D_j$	$n_{Fj}$
<b>Total</b>		$m_{1j}$	$m_{0j}$	$T_j$

Table 2.

Population Parameters for the  $j^{th}$   
Matched Reference and Focal Group

		Score on Item		
		1	0	Total
Group	R	$p_{Rj}$	$q_{Rj}$	1
	F	$p_{Fj}$	$q_{Fj}$	1

Table 3

Correlations among Item Characteristics on the

TAP Mathematics Subtest

Black White Comparison (n=48)

<b>Statistic</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>
<b>1 delta (B/W)</b>	1.000	0.250	0.076	-0.280
<b>2 p-value</b>		1.000	0.015	-0.543
<b>3 pt-biserial</b>			1.000	0.017
<b>4 %omit</b>				1.000
<b>MEAN</b>	0.001	0.575	0.418	3.281
<b>SD</b>	0.538	0.178	0.101	4.305

Table 4

Correlations among Item Characteristics on the TAP

Mathematics Subtest

Male Female Comparison (n = 48)

<b>Statistic</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>
<b>1 delta (M/F)</b>	1.000	0.332	0.122	-0.044
<b>2 p-value</b>		1.000	0.016	-0.543
<b>3 pt-biserial</b>			1.000	-0.017
<b>4 %omit</b>				1.000
<b>MEAN</b>	0.000	0.575	0.418	3.281
<b>SD</b>	0.516	0.178	0.101	4.305



Table 5

Omissions as a Function of Item Characteristics including a Black/White DIF Index

(n=48)

Independent Variable	Parameter Estimate	Standardized Estimate	Standard Error	t-value (b=0)	p	Seq. R <sup>2</sup>	Simple R <sup>2</sup>
Constant	9.657	0.000	3.203	3.02	0.005		
delta(B/W)	-1.257	-0.157	1.140	-1.10	0.277	0.079	0.079
p-value	-12.200	-0.505	3.436	-3.55	0.001	0.317	0.300
pt-biserial	1.534	0.359	5.893	0.26	0.796	0.319	0.000

Table 6

Embedded Omissions as a Function of Item Characteristics including a Male/Female

DIF Index (n=48)

<b>Independent Variable</b>	<b>Parameter Estimate</b>	<b>Standardized Estimate</b>	<b>Standard Error</b>	<b>t-value</b>	<b>p</b>	<b>Seq. R<sup>2</sup></b>	<b>Simple R<sup>2</sup></b>
Constant	9.704	0.000	3.597	2.70	0.011		
p-value	-12.860	-0.534	3.828	-3.36	0.002	0.244	0.244
pt-biser	2.140	0.052	6.214	0.34	0.733	0.249	0.008
delta-MF	0.986	0.127	1.245	0.79	0.434	0.263	0.002

Table 7

School Divisions with High Levels of Embedded Omissions

Name	Madison	Surry	Staunton	Isle of Wight	Rockbridge
#Students	1746	1195	2992	3900	2895
End Yr. #	609	516	1067	1362	4391
Pup/Tea	11.7	11.7	13.9	12.7	9.7
Salary	21303	25105	26017	24121	23817
%Promot	93.8	92	91.5	87.6	94
Reading	50	43	64	40	38
Math	47	45	56	39	40
Info	53	50	61	46	38
Social S.	53	49	69	48	38
Science	48	49	68	46	46
%9th Grad	80.1	78	75.8	60.4	69.9
%Cont Ed.	66.1	60	71.1	71.4	46.1
%Drop out	5.7	1.6	5.5	7.8	4.1
Total Pop	10300	6100	24783	22500	25600
Abil Pay	.463	1	.5064	.4725	.2595
Local %	41.2	68.9	42.4	42	16.1
Local \$	1545	3359	1491	1598	516
Total \$	3703	4876	3520	3800	3206
%Black	32	78	16	45	3
Score	25.26	25.02	27.59	23.78	23.94
Emb. Om.	1.36	1.23	1.17	1.16	1.03

Table 8

School Divisions with Low Levels of Embedded Omissions

Name	Bath	Craig	Patrick	Buena Vista	Waynesboro
#Students	876	694	2805	1225	2639
End Yr.#	344	295	1152	482	954
Pup/Tea	8.8	13.4	15.6	13.9	12.7
Salary	22179	25527	21906	25030	27550
%Promot	97.1	94.5	95.8	91.2	93.6
Reading	51	40	53	51	60
Math	48	36	55	55	63
Info	53	50	59	56	69
Social S.	46	40	57	50	61
Science	55	38	55	50	64
%9th Grad	75.6	94.6	81	83.3	75.2
%Cont Ed.	51.6	71.7	48	61.2	63.7
%Drop out	2.5	2.6	3.8	3.9	7.9
Total Pop	5500	3900	17500	6700	18577
Abil Pay	9644	.3961	.3277	.295	.5127
Local %	70.8	10.5	28.6	29.1	48.1
Local \$	4132	335	969	980	1852
Total \$	5834	3198	3383	3369	3847
%Black	0.03	0.0	0.12	0.02	0.1
Score	25.74	23.1	27.11	27.4	29.29
Emb Om	.09	.13	.19	.15	.1

Table 9

Embedded Omissions as a Function of School Division Characteristics

<b>Independent Variable</b>	<b>Parameter Estimate</b>	<b>Standardized Estimate</b>	<b>Standard Error</b>	<b>t-value (b=0)</b>	<b>p</b>	<b>Seq. R<sup>2</sup></b>	<b>Simple R<sup>2</sup></b>
<b>Constant</b>	1.042	0.000	0.184	5.67	0.000		
<b>Race</b>	-0.324	0.319	0.105	-3.100	0.002	0.052	0.052
<b>Achievement</b>	0.007	0.173	0.005	1.46	0.147	0.078	0.002
<b>Financial</b>	-0.001	-0.022	0.006	-0.20	0.840	0.078	0.009
<b>Size</b>	0.004	0.057	0.007	0.60	0.550	0.081	0.010
<b>Persistence</b>	0.029	0.107	0.024	1.22.	0.223	0.093	0.008

Table 10

Trailing Omissions as a Function of School Division Characteristics

<b>Independent Variable</b>	<b>Parameter Estimate</b>	<b>Standardized Estimate</b>	<b>Standard Error</b>	<b>t-value (b=0)</b>	<b>p</b>	<b>Seq. R<sup>2</sup></b>	<b>Simple R<sup>2</sup></b>
<b>Constant</b>	2.281	0.000	0.472	4.83	0.000		
<b>Race</b>	-0.643	-0.239	0.269	-2.39	0.018	0.113	0.113
<b>Achievement</b>	-0.020	-0.190	0.012	-1.65	0.101	0.114	0.034
<b>Financial Size</b>	0.029	0.201	0.015	1.94	0.054	0.143	0.016
<b>Persistence</b>	0.015	0.072	0.019	0.78	0.435	0.147	0.005
	-0.073	-0.100	0.062	-1.17.	0.243	0.157	0.012

Table 11

Total Omissions as a function of school characteristics

<b>Independent Variable</b>	<b>Parameter Estimate</b>	<b>Standardized Estimate</b>	<b>Standard Error</b>	<b>t-value (b=0)</b>	<b>p</b>	<b>Seq. R<sup>2</sup></b>	<b>Simple R<sup>2</sup></b>
<b>Constant</b>	3.321	0.000	0.586	5.66	0.000		
<b>Race</b>	-0.966	-0.291	0.334	-2.89	0.005	0.116	0.116
<b>Achievement</b>	-0.013	-0.101	0.015	-0.87	0.386	0.117	0.019
<b>Financial</b>	0.028	0.157	0.018	1.50	0.136	0.137	0.012
<b>Size</b>	0.019	0.076	0.023	0.82	0.416	0.141	0.008
<b>Persistence</b>	-0.045	-0.050	0.077	-0.59.	0.558	0.144	0.004

Appendix A

Table A-1

Correlations, Means and Standard Deviations for Division Level Variables  
(n=137)

Correlations

	#STUDS	END_YR#	PUP/TEAC	AVSALARY	%PROMOT	READING
#STUDS	1.00	0.99	0.24	0.49	0.05	0.28
END_YR#	0.99	1.00	0.26	0.48	0.09	0.28
PUP/TEAC	0.24	0.26	1.00	0.11	0.21	0.20
AVSALARY	0.49	0.48	0.11	1.00	0.10	0.48
%PROMOT	0.05	0.09	0.21	0.10	1.00	0.31
READING	0.28	0.28	0.20	0.48	0.31	1.00
MATH	0.34	0.34	0.14	0.55	0.25	0.91
WRITTEN	0.31	0.31	0.10	0.48	0.10	0.84
INFORMAT	0.33	0.33	0.16	0.51	0.27	0.94
SOCIALST	0.30	0.30	0.16	0.51	0.19	0.93
SCIENCE	0.29	0.29	0.18	0.47	0.35	0.92
%9THGRAD	0.07	0.10	0.13	0.08	0.42	0.14
%CONT ED	0.06	0.06	0.12	0.10	-0.03	0.22
%DROPOUT	-0.01	-0.04	-0.19	-0.03	-0.39	-0.18
TOT POP	0.97	0.96	0.17	0.55	-0.03	0.25
ABIL PAY	0.10	0.10	-0.32	0.44	0.16	0.31
LOCAL%\$	0.21	0.21	-0.34	0.54	0.16	0.41
LOCAL\$	0.22	0.22	-0.37	0.66	0.13	0.35
TOTAL\$	0.21	0.20	-0.39	0.68	0.04	0.21
RACE	0.05	0.07	0.26	0.03	0.50	0.51
TROMITS	0.07	0.04	-0.00	0.08	-0.18	-0.23
EMBOMIT	0.19	0.08	0.06	0.11	-0.16	0.01



Table A-1 (Cont'd)

	Correlation						
	MATH	WRITTEN	INFORMAT	SOCIALST	SCIENCE	%9THGRA	
#STUDS	0.34	0.31	0.33	0.30	0.29	0.07	
END_YR#	0.34	0.31	0.33	0.30	0.29	0.10	
PUP/TEAC	0.14	0.10	0.16	0.16	0.18	0.13	
AVSALARY	0.55	0.48	0.51	0.51	0.47	0.08	
%PROMOT	0.25	0.10	0.27	0.19	0.35	0.42	
READING	0.91	0.84	0.94	0.93	0.92	0.14	
MATH	1.0	0.85	0.92	0.88	0.91	0.16	
WRITTEN	0.85	1.00	0.88	0.87	0.84	0.03	
INFORMAT	0.92	0.88	1.00	0.91	0.92	0.18	
SOCIALST	0.88	0.87	0.91	1.00	0.88	0.09	
SCIENCE	0.91	0.84	0.92	0.88	1.00	0.18	
%9THGRAD	0.16	0.03	0.18	0.09	0.18	1.00	
%CONT_ED	0.17	0.20	0.19	0.26	0.13	-0.02	
%DROPOUT	-0.18	-0.05	-0.14	-0.10	-0.21	-0.60	
TOT_POP	0.32	0.30	0.30	0.29	0.27	0.00	
ABIL_PAY	0.32	0.31	0.30	0.30	0.26	0.02	
LOCAL%\$	0.44	0.41	0.44	0.43	0.37	0.03	
LOCAL\$	0.39	0.34	0.36	0.38	0.31	0.07	
TOTAL\$	0.29	0.23	0.24	0.27	0.21	0.05	
RACE	0.45	0.28	0.46	0.39	0.55	0.29	
TROMITS	-0.23	-0.10	-0.20	-0.14	-0.19	-0.08	
EMBOMIT	-0.05	0.12	0.07	0.09	0.00	-0.11	

Table A-1 (cont'd)

## Correlations

	%CONT_ED	%DROPOUT	TOT_POP	ABIL_PAY	LOCAL%\$	LOCAL\$
#STUDS	0.06	-0.01	0.97	0.10	0.21	0.22
END_YR#	0.06	-0.04	0.96	0.10	0.21	0.22
PUP/TEAC	0.12	-0.19	0.17	-0.32	-0.34	-0.37
AVSALARY	0.10	-0.03	0.55	0.44	0.54	0.66
%PROMOT	-0.03	-0.39	-0.03	0.16	0.16	0.13
READING	0.22	-0.18	0.25	0.31	0.41	0.35
MATH	0.17	-0.18	0.32	0.32	0.44	0.39
WRITTEN	0.20	-0.05	0.30	0.31	0.41	0.34
INFORMAT	0.19	-0.14	0.30	0.30	0.44	0.36
SOCIALST	0.26	-0.10	0.29	0.30	0.43	0.38
SCIENCE	0.13	-0.21	0.27	0.26	0.37	0.31
%9THGRAD	-0.02	-0.60	0.00	0.02	0.03	0.07
%CONT_ED	1.00	0.04	0.06	0.18	0.16	0.10
%DROPOUT	0.04	1.00	0.06	-0.03	0.06	0.01
TOT_POP	0.06	0.06	1.00	0.15	0.27	0.30
ABIL_PAY	0.18	-0.03	0.15	1.00	0.87	0.84
LOCAL%\$	0.16	0.06	0.27	0.87	1.00	0.94
LOCAL\$	0.10	0.01	0.30	0.84	0.94	1.00
TOTAL\$	0.03	0.02	0.32	0.71	0.81	0.96
RACE	-0.09	-0.29	-0.01	-0.05	-0.04	-0.07
TROMITS	0.12	0.04	0.11	0.10	0.07	0.11
EMBOMIT	0.18	0.21	0.13	0.06	0.06	0.06

Table A-1 (Cont'd)

	Correlations			
	TOTAL\$	RACE	TROMIT	EMBOMITS
#STUDS	0.21	0.05	0.06	0.10
END YR#	0.20	0.07	0.04	0.08
PUP/TEAC	-0.39	0.26	-0.00	0.05
AVSALARY	0.68	0.03	0.08	0.11
%PROMOT	0.04	0.50	-0.18	-0.16
READING	0.23	0.51	-0.23	0.01
MATH	0.29	0.45	-0.23	-0.05
WRITTEN	0.23	0.28	-0.11	0.12
INFORMAT	0.24	0.46	-0.20	0.07
SOCIALST	0.27	0.39	-0.14	0.09
SCIENCE	0.21	0.55	-0.19	0.00
%9THGRAD	0.05	0.29	-0.08	-0.11
%CONT ED	0.03	-0.09	0.12	0.18
%DROPOUT	0.00	-0.29	0.04	0.21
TOT POP	0.32	-0.01	0.11	0.13
ABIL PAY	0.71	-0.05	0.10	0.06
LOCAL%\$	0.81	-0.04	0.07	0.06
LOCAL\$	0.96	-0.07	0.11	0.06
TOTAL\$	1.00	-0.14	0.16	0.07
RACE	-0.14	1.00	-0.34	-0.23
TROMITS	0.16	-0.34	1.00	0.50
EMBOMIT	0.07	-0.23	0.50	1.00

Table A-1 (Cont'd)

	MEAN	STANDARD DEVIATION
# STUDENTS	7359.76	13807.84
END-YEAR#	2737.64	55395.06
PUPIL/TEACHER RATIO	12.54	1.64
AVERAGE SALARY	24614.70	2905.57
% PROMOTED	93.25	2.93
READING	52.25	9.06
MATHEMATICS	49.63	9.35
WRITTEN	57.17	8.07
INFORMATION	52.23	8.86
SOCIAL STUDIES	55.63	9.12
SCIENCE	56.06	9.83
% 9TH GRADE GRADUATING	75.00	11.07
% CONTINUING EDUCATION	65.44	48.53
% DROPOUT	4.79	1.83
TOTAL POPULATION	41955.35	73507.43
ABILITY TO PAY	0.46	0.17
LOCAL % \$	38.00	13.50
LOCAL \$	1554.00	900.11
TOTAL \$	3787.45	705.31
% BLACK	25.31	22.95
EMBEDDED OMISSIONS	0.48	0.23
TRAILING OMISSIONS	1.12	0.62

Appendix B

Table B-1  
Rotated Factor Loadings

Variable	Achievement	Financial	Size	Persistence
READING	<b>0.936</b>	-0.114	0.111	-0.201
MATH	<b>0.901</b>	-0.172	0.180	-0.179
WRITTEN	<b>0.905</b>	-0.128	0.148	0.003
INFORMAT	<b>0.927</b>	-0.136	0.164	-0.189
SOCIALST	<b>0.936</b>	-0.145	0.142	-0.086
SCIENCE	<b>0.905</b>	-0.092	0.134	-0.247
PUP/TEAC	0.207	<b>0.561</b>	0.316	-0.264
AVSALARY	0.392	- <b>0.505</b>	0.481	-0.089
ABIL_PAY	0.189	- <b>0.863</b>	0.007	-0.051
LOCAL&\$	0.291	- <b>0.898</b>	0.110	-0.018
LOCAL\$	0.202	- <b>0.953</b>	0.157	-0.059
TOTAL\$	0.086	- <b>0.918</b>	0.190	-0.027
#STUDS	0.158	-0.055	<b>0.971</b>	-0.008
END_YR#	0.157	-0.049	<b>0.966</b>	-0.054
TOT_POP	0.141	-0.148	<b>0.961</b>	0.075
%PROMOT	0.144	-0.059	-0.001	- <b>0.759</b>
%9THGRAD	0.065	-0.039	0.038	- <b>0.864</b>
%DROPOUT	-0.078	-0.064	0.022	<b>0.794</b>
%CONT_ED	0.290	-0.041	0.003	0.117

## References

- Allbanese, M. K. (1988). The projected impact of the correction for guessing on individual scores. Journal of Educational Measurement, 25, 149-157.
- Angoff, W. H. (1974). The evaluation of differences in test performance of two or more groups. Educational and Psychological Measurement, 34, 807-816.
- Angoff, W. H. (1989). Does guessing really help?. Journal of Educational Measurement, 26, 323-336.
- Angoff, W. H., & Ford, S. F. (1988). Item-race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 10, 95-106.
- Angoff, W. H., & Schrader, W. B. (1984). A study of hypothesis basic to the use of rights and formula scores. Journal of Educational Measurement, 21, 1-17.
- Beck, R. A. (1982). Handbook of methods for detecting test bias. Baltimore: Johns Hopkins Press.
- Ben-Shakhar, G., & Sinai, Y. (1991). Gender differences in multiple-choice tests: The role of differential guessing tendencies. Journal of Educational Measurement, 28, 23-35.
- Choppin, B. H. (1974). The correction for guessing on objective tests (IEA Monograph Studies, No. 4.). Stockholm: The International Association for The Evaluation of Educational Achievement.
- College Entrance Examination Board. (1990). Taking the SAT. Princeton: Educational Testing Service.
- Cureton, E. E. (1966). The correction for guessing. The Journal of Experimental Education, 4, 44-47.

- Cross, L. H., & Frary, R. B. (1977). An empirical test of Lord's theoretical results regarding formula scoring of multiple-choice tests. Journal of Educational Measurement, 14, 313-321.
- Dossey, J. A., Mullis, I. V., Lindquist, M. M., & Chambers, D. L. (1988). The mathematics report card (Report No. 1-M-01). Princeton: Educational Testing Service.
- Fischer, F. E. (1988). Effects of instructions for guessing on multiple-choice test performance. Educational Research Quarterly, 12, 6-9.
- Hambleton, R., & Rogers, H. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. Applied Measurement In Education, 2, 313-324.
- Joint Committee on Testing Practices (1988). Code of fair testing practices in education. Washington, DC: American Psychological Association.
- Keene, J. M., Jr. (1985). The Test of Achievement and Proficiency. In James V. Mitchell, Jr. (Ed.), The Ninth Mental Measurements Yearbook (pp. 1610-1611). Lincoln, Nebraska: University of Nebraska Press.
- Kim, J., & Mueller, C. W. (1978). Factor analysis: Statistical method and practical issues. New York: McGraw-Hill.
- Kulick, E., & Mu, P. G. (1989). Examining the relationship between differential item functioning and item difficulty (Report No. 89-5). New York: College Board.
- Lord, F. M. (1979). Formula scoring and number-right scoring. Journal of Educational Measurement, 12, 7-11.

- Maccoby, E. E., & Jacklin, C. N. (1974). The psychology of sex differences. Stanford: Stanford University Press.
- McManis, D. L., & Bell, D. R. (1968). Risk-taking by reward-seeking, punishment-avoiding, or mixed orientation retardates. American Journal of Mental Deficiency, 73, 267-272.
- Pedhazur, E. J. (1982). Multiple regression in behavioral research (2nd ed.). New York: Holt Rinehart & Winston.
- Perlman, C. L. (1982). Investigating the stability of four methods for estimating item bias. Unpublished manuscript. Chicago Public Schools, Department of Research and Evaluation, Chicago.
- Rowley, G. L., & Traub, R. E. (1977). Formula scoring, number-right scoring and test-taking strategy. Journal of Educational Measurement, 14, 15-21.
- Scannell, D. P. (Ed.) (1986). Tests of achievement and proficiency. Chicago: Riverside Publishing Company.
- Shepard, L. A., Camilli, G., & Williams, D. M. (1985). Validity of approximation techniques for detecting item bias. Journal of Educational Measurement, 22, 77-105.
- Sherriffs, A. C., & Boomer, D. S. (1954). Who is penalized for guessing? Journal of Educational Psychology, 45, 81-89.
- Tatsuoka, M. M. (1988). Multivariate analysis (2nd ed.). New York: Macmillan.
- Virginia Department of Education (1988). Facing up-23. Richmond: Author.



Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. Journal of Educational Measurement, 26, 55-66.

**ROBERT L. POUR**  
Route 4, Box 547  
Abingdon, VA 24210  
(703) 944-4514

**EDUCATION**

Virginia Tech, Blacksburg, VA  
August, 1991  
Ph.D., Educational Research & Evaluation

University of Virginia, Charlottesville, VA  
Master of Arts, Mathematics, May, 1982

Bowling Green University, Bowling Green, OH  
Bachelor of Science, Education, June, 1968

**EXPERIENCE**

DEPARTMENT OF MATHEMATICS  
EMORY AND HENRY COLLEGE, Emory, VA  
Department Chair and Assistant Professor  
1984 - present

DEPARTMENT OF MATHEMATICS  
WASHINGTON AND LEE UNIVERSITY, Lexington, VA  
Instructor  
1982-1984

DEPARTMENT OF MATHEMATICS  
UNIVERSITY OF VIRGINIA, Charlottesville, VA  
Teaching Assistant  
1981-1982

DEPARTMENT OF MATHEMATICS  
PIEDMONT VA COMMUNITY COLLEGE,  
Charlottesville, VA  
Instructor  
1980-1982

THE TANDEM SCHOOL  
Charlottesville, VA  
Mathematics Teacher  
1980-1981  
AMHERST COUNTY HIGH SCHOOL, Amherst, VA

Mathematics Teacher  
1978-1979

BUCKINGHAM JUNIOR HIGH SCHOOL, Dillwyn, VA  
Mathematics Teacher  
1977-1978

NELSON CO. SENIOR CITIZENS PROGRAM, Shipman, VA  
Outreach Worker  
1976-1977

LIBERATION INSTITUTE, Celina, OH  
Mathematics Teacher/Counselor  
1971-1972

NORTHERN ILLINOIS UNIVERSITY, DeKalb, IL  
Graduate Teaching Assistant  
1970-1971

JOHN F. KENNEDY AMERICAN SCHOOL, Naples, Italy  
Head Mathematics Teacher  
1968-1969

#### **AFFILIATIONS**

MATHEMATICAL ASSOCIATION OF AMERICA

SIGMA XI

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION

SOUTHWEST VIRGINIA COUNCIL OF TEACHERS OF  
MATHEMATICS

A handwritten signature in black ink, appearing to read "Robert L. Long". The signature is fluid and cursive, with a large initial "R" and "L".