INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

University Microfilms International
A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
313/761-4700 800/521-0600

Order Number 9121496

Conceptualizations of test bias and adverse impact: Implications of recent policy proposals

Tesh, Anita Star, Ed.D.

The University of North Carolina at Greensboro, 1990

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

CONCEPTUALIZATIONS OF TEST BIAS AND ADVERSE IMPACT: IMPLICATIONS OF RECENT POLICY PROPOSALS

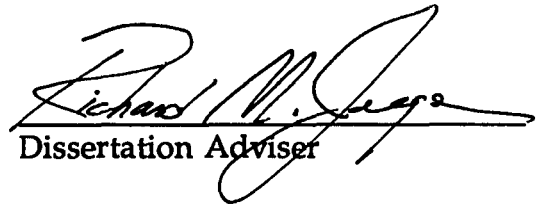
by

Anita S. Tesh

A Dissertation Submitted to
the Faculty of the Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Education

Greensboro
1990

Approved by



Dissertation Adviser

APPROVAL PAGE

This dissertation has been approved by the following committee of the Faculty of the Graduate School at The University of North Carolina at Greensboro.

Dissertation Adviser _____

Committee Members _____

_____

_____

_____

5 November 1990
Date of Acceptance by Committee

5 November 1990
Date of Final Oral Examination

TESH, ANITA S., Ed.D. Conceptualizations of Test Bias and Adverse Impact: Implications of Recent Policy Proposals. (1990) Directed by Dr. Richard M. Jaeger. 297 pp.

The purpose of this study was to investigate selected implications of recent proposals for the imposition of a set of test assembly procedures called the "Golden Rule procedures" in a wide variety of testing situations. These test assembly procedures stipulate that items selected for test inclusion should exhibit differences in performance between groups below a specified level, and should not be more difficult than specified, for either majority or minority groups.

This study used data from the performances of 1807 examinees on four standardized tests completed during their eighth and tenth school grades. Synthetic tests composed of items from the original tests which conformed to the stipulations of the Golden Rule procedure were created. The effects of applying the procedures on the adverse impact and racial bias of test use were examined through comparisons between and among the properties of and results of using original tests and corresponding synthetic tests.

In this study, application of the Golden Rule procedures reduced the disparate impact of test use: the mean total scores of black and white examinees were more similar on the synthetic tests than on the corresponding original tests. However, this reduction in disparate impact was accompanied by impairment of important psychometric properties of the tests. Using Golden Rule procedures lowered the internal-consistency reliability, the average item difficulty level, the average item-total score correlation, and the predictive validity of the tests. No evidence was found that application of the procedures consistently resulted in reduction of test bias. When psychometric properties were examined separately by race,

application of the procedures (1) decreased the similarity of test reliabilities for black examinees and white examinees in some cases, and increased it in others; (2) lowered tests' predictive validities for black examinees and white examinees, with more pronounced impact on white examinees. (In some cases this resulted in an increase, and in other cases a decrease, in the difference between predictive validities for black examinees and white examinees.) For one original test that exhibited differential regression, application of the procedures increased, rather than decreased, the difference between the slopes for black examinees and white examinees.

ACKNOWLEDGEMENTS

My gratitude is extended to Dr. R. M. Jaeger and the members of my dissertation committee, Dr. Lloyd Bond, Dr. Dale Brubaker, Dr. Christian Busch, and Dr. David Ludwig, for their guidance and patience. Thanks also to Dr. Steve Gabrielson of the Georgia Assessment Project for allowing me access to the student performance data used in this dissertation. Finally, and above all, heartfelt thanks to my family and friends, who have supported and tolerated me through the prolonged parturition of this dissertation and graduate degree.

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

Fairness in the distribution of the primary goods of society, including wealth, jobs, access to education, power, and status, is a principal concern in this country. One of the reasons that issues of distributive justice have never been fully resolved is the fundamental and deep-seated ambivalence in our society about the very nature of equity and the meaning of justice (Hartigan & Wigdor, 1989). While most of us hold strong convictions about what constitutes fairness in the allocation of specific societal resources, as Hartigan and Wigdor (1989, p. 30) state: "few of us could lay claim to a systematic, coherent theory of social justice." Most of us react from an inconsistent and unexamined value system which is a mixture of philosophical ideals, normative ideology, and self-interest. As a meritocracy, our society is committed to detecting and rewarding merit, regardless of its source or cause, through open competition among individuals. On the other hand, as a Jeffersonian democracy we are committed to protecting the rights and privileges of the disadvantaged and members of minority groups from victimization or exploitation by the majority (Adler, 1981).

We have historically viewed standardized tests as powerful tools to help us detect merit. Performance on standardized tests is the basis for the allocation of a wide variety of societal goods, including jobs, access to education, recognition, and status. Yet the fairness of standardized test use for certain identifiable subgroups of the population has long been questioned by professionals in the fields of measurement and education (c. f. Thorndike,

1971a, 1971b), and by the lay public, including legislators. Bond (1981, p. 55) has stated that the existence or nonexistence of biases in testing is "a crucial scientific, social, and political issue."

It is widely recognized that black examinees, on average, score lower than white examinees on many achievement and aptitude tests. Hispanic examinees, particularly those for whom English is not a first language, and American Indian examinees also typically score lower than white examinees. Women, on average, score lower than men on some tests, such as those involving mathematics. Since standardized tests are used as a basis for decision making in many educational and employment situations, these differences in average performance create the potential for substantial adverse impact on minorities and women. There is concern, in particular, about the adverse effect on minority children of the current widespread use of standardized tests in public schools (First & Cardenas, 1986; National Commission on Testing and Public Policy (NCTPP), 1990). Concern over this adverse impact has led to a variety of actions, including efforts to define the term "bias" precisely, the development and refinement of techniques to detect biased items in a test, and the establishment of laws, regulations and guidelines concerning discriminatory practices involving testing (Berk, 1982; Cole & Moss, 1989; Reynolds & Brown, 1984).

One strategy for attempting to make tests fair to minorities that has received considerable attention in the measurement community and from the public is the *Golden Rule strategy* (Haney & Reidy, 1987). In essence, the Golden Rule strategy is a set of test assembly procedures intended to minimize or eliminate the disparate impact of test use by stipulating that the

items selected for test inclusion should exhibit small differences in performance between groups, and should not be excessively difficult for either minority or majority groups. (The exact stipulations of the Golden Rule strategy are provided and discussed in a subsequent chapter of this study.) At times, the Golden Rule strategy has been described as a mechanism to reduce test bias, while at other times, it has been described as a mechanism to reduce adverse or disparate impact (Haney & Reidy, 1987). As Flaugher (1978) noted, many members of the public see these two issues as synonymous. Professionals in the fields of measurement and testing, however, see disparate impact and bias as two distinct issues.

Although the Golden Rule strategy was initially intended to apply to two licensure examinations for insurance agents developed by Educational Testing Service (ETS), it has gained widespread popular support and interest. Since 1985, bills requiring use of variations of the Golden Rule strategy have been introduced in the state legislatures of California, New York, Wisconsin, and Texas. An out-of-court settlement in Alabama incorporated the Golden Rule strategy. John Weiss, Executive Director of the National Center for Fair and Open Testing (Fairtest) has stated that the procedures of the Golden Rule strategy should be applied as often and as widely as possible (Weiss, 1987). The legislative impact of the Golden Rule strategy is described and discussed in a subsequent chapter of this dissertation.

Professionals in the fields of measurement and testing have expressed serious reservations about widespread application of the Golden Rule strategy or variants of it. It has been stated that widespread application of the Golden Rule strategy would "result in severe adverse consequences for those

individuals and educational institutions that objective tests are designed to serve" (Jaeger, 1987). These reservations center on the potential effects of the Golden Rule strategy on the psychometric properties of tests (cf. Jaeger, 1987; Linn and Drasgow, 1987; Marco, 1988; Shepard, 1987). The results of existing research on the effects of application of the Golden Rule strategy on the psychometric properties of tests are summarized in the next chapter of this dissertation, however these effects have not yet been thoroughly examined using empirical procedures.

Purposes and Procedures of this Study

This study will examine the philosophical and policy contexts of issues of adverse impact and preferential treatment, and the philosophical, policy, and scientific contexts of issues of test bias, as they relate to the Golden Rule strategy. Empirically, two major research questions will be investigated in this study. The first is: Is application of the Golden Rule strategy effective in reducing the adverse impact of test use? The second major research question is: Is application of the Golden Rule strategy effective in reducing test bias? In the next chapter of this study, these two major research questions are derived from the professional literature, together with a series of subsidiary research questions that serve to elaborate and define the major research questions, and to relate them to the psychometric properties of tests.

The empirical research questions of this study will be investigated using data on the performances of 1807 students on the Georgia Eighth-Grade Criterion Referenced Tests in Reading and Mathematics in the Spring of 1986, and data on the same students' subsequent performances on the Georgia Basic Skills Tests in the Fall of 1987. The Georgia Basic Skills Tests, administered at

the beginning of the tenth grade, also have Mathematics and Reading components.

Investigation of the empirical research questions of this study required the creation of synthetic tests which were composed only of those items from the original tests which conformed to the stipulations of the Golden Rule strategy. The first major research question of this study, which addresses the effectiveness of the Golden Rule strategy in reducing the adverse impact of test use, was empirically examined by investigating whether the differences between the mean scores for black and white examinees were smaller on synthetic tests composed only of items which conform to the specifications of the Golden Rule strategy, than on the original, unmanipulated tests.

The second major research question of this study addressed the effectiveness of application of the Golden Rule strategy in reducing test bias. As Cole and Moss (1989) and others have noted, the question "Is this test biased?" can not receive a simple yes-or-no answer. Rather, a series of connected concerns relating to the fairness of the test to all examinees must be examined in answer to the question (Cole and Moss, 1989; Hackett, Holland, Pearlman, & Thayer, 1987). Recognizing that all investigation of bias relates to the search for evidence of differential construct validity, Cole and Moss (1989) suggest that the evidence pertinent to bias be grouped into five categories: 1) internal test structure; 2) external test relationships; 3) content and format; 4) test administration and scoring; and 5) constructs in context. These five categories proposed by Cole and Moss are not seen as different types of bias, but as different areas in which evidence about bias can be examined, as part of the unitary process of construct validation.

For the empirical investigation of the second major research question of this study, evidence within each of the five categories proposed by Cole and Moss (1989) was examined to the extent that it was feasible and appropriate, given the nature of the tests and the data at hand. This involved a series of examinations of the characteristics of, and relationships between and among, examinees' performances on the synthetic tests which conform to the stipulations of the Golden Rule strategy, and on the original standardized tests. Under the category of internal test structure, as proposed by Cole and Moss (1989), five subsidiary research questions were examined. These were:

1. What is the effect of applying the Golden Rule strategy on the internal-consistency reliability of tests?
2. Does application of the Golden Rule strategy increase the similarity of the reliability of tests for black and white subgroups?
3. What is the effect of applying the Golden Rule strategy (without specific protection of content representativeness) on the content representativeness of tests?
4. What is the effect of applying the Golden Rule strategy on the average item difficulty (as defined in classical true-score theory) of tests?
5. What is the effect of applying the Golden Rule strategy on the average item by total-score correlation of tests?

Under the category of external test relationships, three subsidiary research questions were investigated. These were:

1. What is the effect of applying the Golden Rule strategy on the overall predictive validity of tests?
2. Does application of the Golden Rule strategy increase the similarity of test-criterion correlations for black and white examinees?
3. Does application of the Golden Rule strategy increase the similarity of regression equations for black and white examinees (i.e., does it reduce differential prediction)?

Only limited evidence in the category of test content and format was examined in this study. The items that were excluded from the tests by applying the Golden Rule strategy were reviewed by the researcher for content similarities. It was beyond the scope of this study to compose panels of experts to review these items for discriminatory language or content. The administration and scoring of tests is not affected by applying the stipulations of the Golden Rule strategy, so no evidence in this category was examined empirically in this study. One subsidiary research question was addressed within the category of constructs in context. This was:

1. Does application of the Golden Rule strategy to both a test and the criterion it is intended to predict reduce the dissimilarity of correlations between the test and criterion, and reduce the dissimilarity regression equations, for black and white examinees?

The exact procedures for investigation of each of these research questions are provided in a subsequent chapter of this study.

Organization of the Remainder of this Study

The remaining chapters present more fully the details of this dissertation study. Chapter II summarizes relevant literature on the background of the Golden Rule strategy, policy and legislative impacts of the Golden Rule strategy, research related to the Golden Rule strategy, and an explication of the details of the procedures of the Golden Rule strategy. Chapter II also contains a summary and synthesis of literature related to the philosophical and policy contexts of issues of adverse impact and preferential treatment, and the philosophical, policy, and a summary of scientific contexts of issues of test bias, as they pertain to the Golden Rule strategy. Various conceptualizations of test

bias and adverse impact are used to derive a set of hierarchical research questions for the empirical component of the study.

Chapter III of this study describes methodology for examination of the hierarchical research questions, as well as details on the sample and the instruments used. Chapter IV reports the results of data analyses, and Chapter V provides a discussion of the results of the data analyses, in light of the philosophical and policy contexts of issues of adverse impact and preferential treatment, and the philosophical, policy, and scientific contexts of issues of test bias, as they relate to the Golden Rule strategy.

CHAPTER II

REVIEW OF RELEVANT LITERATURE

The purpose of this chapter is to review, summarize and integrate relevant literature pertaining to the philosophical and policy contexts of issues of adverse impact and preferential treatment, and the philosophical, policy, and scientific contexts of bias, as they relate to the Golden Rule strategy.

This chapter begins with a review of the background of the Golden Rule strategy, followed by an explication of the details of the procedures of the Golden Rule strategy. A summary of policy and legislative impacts of the Golden Rule strategy is presented next, followed by a summary of existing research findings and professional judgements and opinions concerning the consequences of applying the procedures of the Golden Rule strategy. The philosophical context of the Golden Rule strategy is then examined through an investigation of conceptualizations of justice and equality. This is followed by an examination of the social policy context of the Golden Rule strategy. An examination of the scientific context of the Golden Rule strategy is presented next, through an investigation of conceptualizations of bias and adverse impact. The final section of this chapter synthesizes the literature examined to derive a series of hierarchical research questions related to application of the procedures of the Golden Rule strategy. These research questions were used to guide the empirical component of this dissertation.

Background of the Golden Rule Strategy

In October of 1975, the Illinois Department of Insurance began use of a new insurance agent licensing examination, developed by Educational Testing Service

(ETS). In 1976, the Golden Rule Insurance Company and five individuals who had failed the Life and Accident & Health portions of the Illinois insurance licensing examinations brought suit against the Illinois Department of Insurance and ETS, alleging that the insurance agent licensing examinations were not sufficiently related to the knowledge, skills, and abilities needed by insurance agents, and that the tests intentionally discriminated against test-takers on the basis of race and were racially biased, (three of the five individual plaintiffs were black). ETS was named in the suit because it had designed the licensure examinations and had administered such examinations in Illinois for more than a decade.

J. Patrick Rooney, Chief Executive Officer of the Golden Rule Insurance Company at the time the suit was filed, stated that the passing rate on the new insurance agent licensing examination was initially much lower than the passing rate had been for the test it replaced (reportedly about 31% of examinees passed the new test, as compared to 60-70% for the previous test). He further stated that use of the new test was having a devastating impact on the insurance industry. According to Rooney, the passing rates for black examinees on the new tests were even lower, causing Golden Rule's black regional managers in Chicago "simply to give up on trying to get new black candidates into the insurance business" (Rooney, 1987a, p. 10). His impression was that the new examination effectively excluded blacks from the occupation of insurance agent. The Illinois Department of Insurance expressed concern over the low passing rates, but Rooney (1987a, p. 10) states that questions about racial impact "were met with stonewalling." This lack of concern by the Department of Insurance over racial exclusion led to the Golden Rule Insurance Company's decision to sue. In May,

1976, ETS modified the test such that overall passing rates rose to the 70-75% range, and industry complaints quieted, while the issue of disparate impact remained unaddressed. At this point, the Department of Insurance had not been collecting ethnic-group-membership data on examinees, and ETS had not performed a job analysis of the test or pretested items for their effect on minorities (Rooney, 1987a).

The suit brought by *Golden Rule Insurance Company, et al.* was dismissed twice by trial court. The case was initially dismissed when ETS developed a revised insurance agent licensing examination for the state of Illinois, and discontinued use of the form involved in the original suit (*Golden Rule Insurance Company, et al. v. Duncan, et al.*, 1978). The suit was subsequently refiled, covering both the original and revised test forms. The suit was then dismissed on the grounds that it failed to cite intentional discrimination, and that ETS, as a private contractor, was not subject to the constitutional mandates regarding equal protection and due process (*Golden Rule Insurance Company, et al. v. Mathias, et al.*, 1979). This dismissal was unanimously reversed and remanded for trial by a state appellate court in 1980 (*Golden Rule Insurance Company, et al. v. Mathias, et al.*, 1980, p. 11), although that court expressed "considerable dubiety as to whether plaintiffs' allegations can be sustained at trial." Equally important, according to Werner (1988), the plaintiffs had never offered to show that the tests were not job-related, and had never identified a single "biased" item, despite the fact that they were given access to thousands of ETS documents and hundreds of actual test items.

On November 20, 1984, after eight years of negotiations, the case was resolved in a voluntary out-of-court settlement (*Golden Rule Insurance Company, et*

al. v. Washburn, et al., 1984). The key provision of the settlement concerned procedures for selecting test items that were intended to reduce disparate impact and, by implication, racial bias, in two of the insurance agent licensing exams (Faggen, 1987; Werner, 1988). This set of procedures for test assembly has come to be called the "Golden Rule strategy" (Haney & Reidy, 1987), the "Golden Rule procedures" (Linn & Drasgow, 1987), or simply the "Golden Rule" (Werner, 1988). In the popular lexicon, the name "Golden Rule" has Biblical associations, and signifies egalitarian concern for others. In the present context, however, it merely, or perhaps ironically, reflects the name of the litigant in a lawsuit (Werner, 1988).

The settlement agreement contains no admission of wrongdoing by ETS or the Illinois Department of Insurance. No recompense for damages was awarded the plaintiffs; in fact, the plaintiffs were required to contribute to the funds needed to perform some of the analyses specified by the settlement agreement (*Golden Rule Insurance Company, et al. v. Washburn, et al.*, 1984). ETS has stated that it agreed to settle the case because the plaintiffs ultimately gave up many of their original unacceptable demands, the final terms of the settlement were not inconsistent with ETS practice, and the prospect of several more years of costly litigation seemed wasteful. ETS has stressed that their agreement to settle out of court does not constitute an admission of racial bias in test construction or other wrongdoing (Anrig, 1987b, 1988; Werner, 1988). However, Rooney (1987a, 1987b) of the Golden Rule Insurance Company maintains that Gregory Anrig of ETS initiated the effort to settle out of court when the Social Security Administration, which has racial information on all persons with social security numbers, agreed to process examinee passing data by race for use in the trial.

When, as required by the terms of the settlement, ETS reported passing rates by race for the last year of testing prior to implementation of the settlement's test assembly procedures, it was found that only 59% of black examinees passed the Life Insurance examination, and only 41% of black examinees passed the Accident and Health Insurance Examination. The passing rates for white examinees on these two tests were 83% and 74%, respectively. Differences between the mean scores of blacks and whites exceeded 14 standard deviation units. Rooney (1987b) claims that ETS agreed to settle out of court in part to avoid disclosure of passing rates by race during the previous nine years of testing. He stresses that the stipulations of the settlement were worked out by ETS experts in conjunction with Golden Rule experts; they were not imposed on ETS by the court.

Requirements of the Golden Rule Agreement

The settlement agreement between the Golden Rule Insurance Company, the Illinois Department of Insurance, and ETS, has several significant provisions. The provisions of central interest to this dissertation study involve what have come to be called the "Golden Rule procedures," a set of test assembly procedures ETS and the Department agreed to use in assembling new forms of the Life Insurance and Accident and Health Insurance tests. These procedures require that all potential test items be categorized into two types: Type I items and Type II items. Type I items are those for which: a) the correct-answer rates for blacks, whites, and all examinees are not less than 40% at the .05 level of statistical significance, and, b) the correct answer rates of black examinees and white examinees differ by no more than 15% at the .05 level of statistical

significance. All other items are classified as Type II items. In test assembly, the Golden Rule agreement specifies that Type I items be used exclusively if they are available in sufficient numbers, and that, among Type I items, those with the smallest difference between proportions correct of black and white examinees be used first. Type II items are to be used only if Type I items do not exist in sufficient numbers to satisfy the constraints of the test plan, and, to the extent that it is necessary to use Type II items, those with the smallest difference between proportions correct of black and white examinees are to be used first. The agreement also contains provisions for not using Type I items for causes such as breach in security of the item, or duplication of content with a previously selected item. In such cases, the decision not to use a Type I item is to be discussed with an advisory committee, described below (*Golden Rule Insurance Company, et al. v. Washburn, et al.*, 1984, p. 10). These procedures clearly convey that it is ultimately desirable to have tests constructed solely of Type I items.

The Golden Rule settlement agreement had provisions beyond those involving test assembly. ETS and the Department agreed to collect racial, ethnic, and educational data, on a voluntary basis, from all examinees. They agreed to include specified numbers of items for pretesting in each test form. They also agreed to publish annual reports containing numbers of examinees, percentage of examinees passing each test, and mean scaled scores on each test, by racial or ethnic subgroup. The correct answer rates and item by scaled-score correlations also were to be reported for black examinees and all examinees combined. ETS agreed that the tests would comply with certain professional and reading-level standards. The Department agreed to establish an advisory committee composed of persons knowledgeable of the fields of psychometrics and

insurance to assist in developing tests and reviewing test results. Finally, ETS agreed to disclose one form of each test every other year (*Golden Rule Insurance Company, et al. v. Washburn, et al.*, 1984). Ironically, the settlement required that, in constructing future tests, ETS should adhere to the 1985 *Standards for Educational and Psychological Testing* (American Psychological Association [APA] et al., 1985), yet these standards are discordant with procedures of the sort specified in the settlement, and specifically recognize that differential item or test performance does not constitute evidence of bias (p. 26-27).

Further Impact of the Golden Rule Strategy

In the initial out-of-court settlement, it was agreed that the Golden Rule strategy would be applied to only two of the four tests of the Illinois Insurance Licensing Program (*Golden Rule Insurance Company, et al. v. Washburn, et al.*, 1984). However, the agreement has since had impact far beyond insurance licensure in Illinois. J. Patrick Rooney (1987a), Chief Executive Officer of the Golden Rule Insurance Company, maintains that the provisions of the Golden Rule agreement are laudable and "should be imposed upon ETS in every possible situation." He further expressed the belief that the procedures "should have the effect over time of reducing unnecessary racial differences" on tests (p. 12). Groups concerned with fairness in testing practice, and some legislators, have advocated the general use of these or similar procedures to reduce or eliminate what they view as test bias (cf. Rooney, 1987a; Weiss, 1987). The procedures have been called a "practical procedure currently available to make tests as fair as possible" (Weiss, 1987, p. 24). Emory University Professor Martin Shapiro told the New York Times, "Once you have this method, not to use it is to knowingly use a more

discriminatory procedure" (Weiss, 1987, p. 24). The fact that ETS agreed to settle the Golden Rule case out of court, rather than insisting on a trial, has been taken by some (cf. Rooney, 1987a, 1987b) as an admission that the ETS was engaging in unlawful and discriminatory testing practices in Illinois, despite ETS's insistence that this is not the case (Anrig, 1987b).

Proponents of the Golden Rule strategy have argued that, in contrast to more technically complex methods of bias detection, it is "understandable" to the layman, lawmaker, or judge (Rooney, 1987a). While some of its proponents say that they "...recognize that group score differences reflect a host of causes, including genuine knowledge differences, test-taking abilities, as well as the inclusion on tests of irrelevant and biased questions " (Weiss, 1987, p. 25), they nonetheless feel that "the purpose of the Golden Rule reform is to help assure that biased test questions are removed from exams" (Weiss, 1987, p. 25) and regard the procedures as "an important milestone in the quest for fair, unbiased testing" (Rooney, 1987a, p. 12).

The Golden Rule settlement has also influenced other legal actions. Another out-of-court settlement in Alabama (*Allen v. Alabama State Board of Education,* 1985), following a class action suit charging racial bias in Alabama's teacher certification tests, resulted in imposition of an even more stringent variation of the Golden Rule provisions. The provisions of the settlement required that preference be given to items for which the difference between proportions of correct answers for black and white examinees be no more than 5%. The provisions did not cite the need to ensure content representativeness of a test as a legitimate basis for inclusion of items for which the difference between proportions of correct answers for black and white examinees was more than 5%.

This agreement also required formation of panels of black educators to review all tests for racially biased content or language.

Since 1985, legislation has been proposed in the state legislatures of California, New York, Wisconsin, Massachusetts, and Texas to require widespread application of variants of the Golden Rule strategy with tests used for a variety of purposes, including admissions, placement, certification and licensure (Faggen, 1987; McAllister, 1987; Werner, 1988). Appendix A of this study contains a list of legislative initiatives identified by this researcher which suggest regulation of standardized testing through stipulations similar to those of the Golden Rule settlement. Some of these bills propose only reporting and/or disclosure requirements similar to those of the Golden Rule settlement. Often, the bills propose more stringent stipulations for item selection than those in the Golden Rule settlement. One legislative proposal in Texas (Senate Bill 29) required that, in tests used for admission, placement, or advancement in an educational program or institution, *no* test items be used for which the correct answer rates for any two racial or ethnic groups differ by more than 15%, or which fewer than 30% of any racial or ethnic group answer correctly. Another Texas bill (Senate Bill 28), aimed at admissions tests for teacher education programs, *prohibited* the use of items for which the correct answer rates of *any* two ethnic groups differ by more than 10%, or which fewer than 40% of *any* racial or ethnic group answer correctly. Yet another Texas bill (House Bill 1377) required special scrutiny and review of any items on tests used for admission to teacher education programs for which any racial or ethnic subgroup had correct answer rates significantly different from the average for all examinees taking the test. Two bills requiring variants of the Golden Rule reporting and disclosure

procedures were introduced in New York in 1987, S-3623/A-5601 and S-3614/A-5582. The second of these bills, and a subsequent bill, New York Bill S-3614 /A-5582, required panel review for any licensure test item for which the correct answer rates for white and minority examinees, or male and female examinees, differed by more than 10%; if the panel did not agree that the question was free of bias, its use was prohibited. Several bills proposing use of variants of the Golden Rule procedures were introduced in the California legislature in the 1986-87 session (Faggen, 1987). One of these (California Assembly Bill no. 4046) stated as its purpose, to "neutralize cultural differences" in licensing examinations for various professions and occupations, and suggested the imposition of test assembly stipulations identical to those in the Golden Rule settlement, except that it required consideration of correct answer rates for four ethnic minorities, not just blacks. California Assembly Bill 4045, also introduced in 1986, called for postsecondary admission and placement tests to be assembled by selecting first those items with the least difference in correct answer rates between whites and several minority groups. Bills introduced in New York in 1986 (S-8985/A-11023 and S-9020/A-11029) also called for application of the Golden Rule test assembly procedures using data for minority groups other than blacks. A bill proposed in Wisconsin in 1986 (Assembly Bill 855) for professional licensure examinations required that preference be given to items for which the correct answer rates of white examinees and minority examinees differed by no more than 15%, but did not specify minimum correct answer rates.

To date, none of these proposed bills has been enacted. However, these cases suggest that segments of the public view the Golden Rule procedures, or

variants of them, as sensible ways to address discriminatory testing practices and to reduce test bias.

Research and Professional Judgements Concerning the Golden Rule Strategy

It has been argued that the Golden Rule strategy, in contrast to more technically complex methods of bias detection, is "understandable" to the layman, lawmaker, or judge (Rooney, 1987a). Yet one recurrent concern of measurement professionals is that the Golden Rule strategy reflects a basic misconception about the nature of bias (cf. Jaeger, 1987; Linn & Drasgow, 1987; Shepard, 1987; Werner, 1988). Many different meanings have been attached to the word "bias". Among people without training in measurement, test bias is often seen as synonymous with differences between minority and majority average scores on the test. (By this definition, measurement of height in inches would be considered "biased" against females, since they are shorter, on average, than males.) Measurement professionals argue that the Golden Rule provisions are based upon this definition of bias (cf. Bond, 1987; Jaeger, 1987; Linn & Drasgow, 1987; Werner, 1988). Measurement professionals emphasize that differences in the distributions of scores of groups might validly reflect differences between the groups on the construct the test is designed to measure, and that bias is present only if the test or items on the test systematically underestimate the performance of one group, measure different constructs in the different groups, or are otherwise less valid for one group than another (Angoff, 1982; Bond, 1981; Cole & Moss, 1989; Shepard, 1982, 1987). Measurement professionals have stressed that, although differences between average performances of groups on tests or test items may be a cause for concern, it is not prima facie evidence of bias, and thus use of the Golden Rule procedures might

sacrifice important psychometric properties of a test without true benefit to any group (cf. Bond, 1987; Jaeger, 1987; Linn and Drasgow, 1987; Marco, 1988; Shepard, 1987; Werner, 1988).

The concerns cited above have led professionals in the fields of measurement and testing to express serious reservations about widespread application of the Golden Rule strategy or variants of it. Both the American Educational Research Association (AERA) and the National Council on Measurement in Education (NCME) have expressed strong opposition to use of the Golden Rule strategy or similar procedures in test assembly. In January 1987, the President of ETS, Gregory Anrig, publicly stated that he had made a mistake in agreeing to the out-of-court settlement with the Golden Rule Insurance Company, and stressed that the procedures were never intended to have application beyond the two insurance tests for which they were specified (Anrig, 1987a). Anrig has since confirmed his opinion that the settlement was a mistake, and has further stated that the test assembly procedures specified in the settlement "have proved cumbersome to use and have not improved the tests or apparently reduced the performance gap between black and white examinees" (Anrig, 1987b, p. 24). Writing as president of NCME, Richard M. Jaeger (1987) stated that requiring widespread application of the Golden Rule procedures would "result in severe adverse consequences for those individuals and educational institutions that objective tests are designed to serve" by producing tests for which scores were less valid and reliable for both majority and minority students. Linn and Drasgow (1987) also state that application of the Golden Rule strategy would undermine test reliability by favoring items with poor discriminating power, and would distort the construct validity of tests. In letters

to the Texas House and Senate Education Committees in March of 1987, AERA President Lauren Resnick wrote that although legislation proposing use of the Golden Rule procedures clearly embodied a well-meaning attempt to reduce ethnic and racial bias in educational tests, as a professional body AERA "stands strongly against any proposed solutions which are inconsistent with the research-based knowledge accumulated to date on the problem" and concluded that the Golden Rule provisions "are clearly inconsistent with that knowledge" (Faggen, 1987, p. 6). Similarly, in letters to the New York State Senate's Education Committee, NCME President Richard M. Jaeger and Association of American Medical Colleges President Robert G. Petersdorf stressed that proposed procedures mirroring the Golden Rule were inadequate to detect the presence of bias in a test, evaluate the extent of any bias, or address any bias present in a test (Faggen, 1987: Jaeger, 1987).

Writing from a "minority perspective," Bond (1987) agrees that the Golden Rule procedures are fundamentally misguided. However, he stresses that the error lies in the specification of fixed and arbitrary values for minimum acceptable correct answer rates and for maximum acceptable difference between groups' correct answer rates for test items, not in the principle of introducing issues of equity into item selection practices. For professionally developed tests, many more items are written than are included in the final test. Selection among items addressing the same content is typically based either upon item by total-score correlations, (for tests intended to measure a unidimensional construct), or upon item by criterion-score correlations, (for tests intended to predict some external criterion). Sometimes both of these criteria are used. Licensure and certification tests are generally not assumed to be unidimensional, and no

external predictive criterion typically exists. Content validity is the predominant concern for such tests, and item selection is typically based upon expert opinion. Bond proposes that statistics on differential item performance could validly play a part in final item selection under some circumstances. He opposes the robotic application of rigid formulas of any sort for selecting among items, and suggests that statistics on differential performance could be considered when selecting between items with essentially equal correlations to total test scores or appropriate external criteria. The *principle* of considering differential group performance is not at fault in the Golden Rule procedures: it is the imposition of inflexible and arbitrary rules for item selection which is at fault (Bond, 1987).

The Golden Rule stipulations also find very limited support in the *Standards for Educational and Psychological Testing* (American Psychological Association, [APA], American Educational Research Association, & National Council on Measurement in Education, 1985). The *Standards* (p. 25) state: "When selecting the type and content of items for tests and inventories, test developers should consider the content and type in relation to cultural backgrounds and prior experiences of the variety of ethnic, cultural, age, and gender groups represented in the intended population of test takers." The *Standards* go on to say that, when the relevance of such factors is in question, test developers should establish a review process for item material. These statements are consistent with the spirit of the Golden Rule settlement, although the settlement stipulations go far beyond what the *Standards* suggest or require. Incongruous with the settlement, however, the *Standards* (p. 25) specifically state that "differential response rates do not necessarily invalidate such items or scales based upon them. However, the developer's aim should be to maximize scale validity and, within this

constraint, the developer should strive to minimize the potential misrepresentation of interests for major groups in the population that is served." The *Standards* (p. 26) further state that, when previous research indicates that disparate impact may exist for a test, the researchers should investigate this as soon as is feasible, and that such investigations should be directed toward determining whether group differences in performance in fact reflect bias. Thus, the *Standards* recognize that differential performance is not synonymous with bias, and place paramount importance on validity. Elimination of test items solely on the basis of differential performance is inconsistent with the *Standards,* as is the imposition of an arbitrary minimum difficulty level for test items.

Werner (1988) summarized many of the concerns raised regarding application of Golden Rule-like procedures:

> 1. The Golden Rule procedures inappropriately lower standards of competence. The at-least-40% criteria will make tests easier, despite inadequate technical or public policy justification for doing so. The proportion of lower-level items (i.e. items testing only recognition and recall) will be increased by the procedures.
> 2. The procedures may fail to minimize test performance differences among groups of test takers.
> 3. The procedures underestimate the complexities of controlling for test question difficulty and validity in test assembly.
> 4. The procedures cannot distinguish a biased question from an unbiased question which validly assesses knowledge or skills possessed in differing levels by different groups.
> 5. Better methods of bias detection are available. Rather than being based upon average performance differences between groups, these methods address whether an item performs differently for individuals of equal ability.
> 6. The procedures fail to account for factors that can usefully explain group differences in test performance, such as examinee preparedness.
> 7. The procedures are incompatible with professional standards on testing, such as the 1985 *Standards for Educational and Psychological Testing.*

8. The procedures would probably require excessive investment in test item construction, and would waste scarce test development resources.

9. Under the procedures, examinees would be deceived about the use to be made of the personal data requested of them. Legislative efforts mirroring the Golden Rule require that examinees receive explicit assurance that the personal data they submit would be used only for research and statistical purposes. Using this data to determine the acceptability of test questions violates these assurances.

10. The demographic data on which the procedures depend would be subject to systematic distortion for tests which are administered repeatedly to the same examinees.

11. The procedures could ultimately stimulate actions harmful to the groups it is intended to help. Many tests are used to limit entry in professions or occupations to qualified applicants. Many of these same occupations currently have multiple paths to eligibility, such as two or four year educational programs, military or work training, and so on. These alternative routes often provide access to the careers for minorities and women, however, examination success rates from each of the routes are not generally equal. If test assembly procedures limited the ability to control entry into the occupation or profession through testing, governing boards might begin to limit entry pathways to those known to produce more homogeneously prepared applicants.

To date, empirical research related to the consequences of applying the Golden Rule procedures remains scant. Linn and Drasgow (1987) found that only 25 of 85 items on a form of the SAT Verbal subtest would satisfy the Golden Rule stipulations for Type I items. This has serious implications for item pool requirements, if tests are limited largely or exclusively to Type I items, as several legislative initiatives have proposed. Their findings also suggested that, for the test investigated, application of the Golden Rule criteria would lead to rejection of a large number of difficult items upon which blacks and whites performed similarly. They concluded that application of the Golden Rule procedures could actually increase, rather than decrease, the difference in average score between black and white examinees. Consistent with this suggestion, Anrig (1987b, p. 26) has reported, based upon application of the Golden Rule procedures in the

original Illinois insurance licensure tests for which they were intended, that "the results to date do not suggest that the Golden Rule procedures increase the black examinee passing rate." Linn and Drasgow (1987, p. 17) conclude that "the Golden Rule procedure threatens to undermine the most important characteristics of sound tests," their reliability and validity.

Using Item Response Theory (IRT) methodology and data from high school students, Marco (1987, 1988) conducted an investigation of the effect of applying four different sets of test assembly rules on the reliability (as defined by IRT methodology) and content of simulated SAT Verbal and Mathematics tests. These four sets of test assembly rules corresponded to test assembly procedures proposed in legislative bills in California and New York. For the first test form, items with proportions correct of less than chance level for either black or white examinees were excluded, and items with the least difference in proportions correct between groups were used first. Content and statistical specifications were not protected in this form. In the second form items with proportions correct of less than .40 for black or white examinees were excluded, and items with the least difference between groups were used first. Content and statistical specifications were again not protected in this form. The assembly procedures for the third form were identical to those of the first form, except that content representativeness and average item difficulty were controlled. For the fourth form, items for which the correct answer rates for both black and white examinees were greater than .40, and the difference between answer rates was less than .15, were used before other items, and items with the least difference in correct answer rates between groups were used first. Content representativeness was protected in this form, thus it corresponds to the Golden Rule stipulations,

however, Marco does not report how many Type II items were actually used in the two tests of this form. Marco's findings suggest that application of these test assembly procedures are likely to have severe impact on the item pool for relatively difficult tests like the SAT: less than 17% of the items in this study met the Golden Rule criteria for Type I items. He also found that, for the SAT, use of the test assembly procedures described above substantially reduced the content representativeness of the test unless content representativeness was specifically protected in test assembly. His results also suggest that application of the Golden Rule procedures might result in (a) the elimination of items of middle difficulty, since these items often had the largest differences in proportions correct for black and white examinees, (b) lower average item-total score correlations, and (c) lower reliabilities for both blacks and whites. (Marco estimated the test's reliabilities using IRT methodology, which, unlike classical test theory, is based upon the assumption that the standard error of measurement (SEM) of a test varies with score level (Dorans, 1984). The reliability was estimated by [1 - (average SEM^2)/(standard deviation)^2] (Marco, 1988)). The reliabilities observed for the unaltered Verbal SAT were 0.89 for black examinees and 0.90 for white examinees. The reliabilities for the Form Four Verbal SAT (which included an unknown number of Type II items) were 0.82 and 0.85 for black and white examinees respectively. For the unaltered Mathematics SAT, the reliabilities for black and white examinees were 0.88 and 0.91, respectively; while for the Form Four Mathematics SAT they were 0.87 and 0.91. The other synthetic test forms showed similar patterns for reliabilities, when compared to the unaltered test, with the exception of Form Two, which was found to be slightly more reliable for black examinees for both the Verbal and the

Mathematics SAT than were the corresponding unaltered tests. Marco also found that application of variants of the Golden Rule procedures were not consistently successful in decreasing average score differences between black and white examinees, since the procedures eliminated many items upon which the groups performed similarly. Some variants of the procedures Marco investigated actually resulted in increased differences in average score between black and white examinees. In the form of the SAT investigated by Marco (1988), the number of useable items in the item pool decreased dramatically when test assembly procedures prohibited use of items with proportions correct for black or white examinees below .40. The proportions correct for either black or white examinees, or the total group, was less than .40 for 58% of the SAT Verbal items and for 52% of the SAT Mathematics items. When both criteria for Type I items were imposed, only 15% of the SAT Verbal and 15% of the SAT Mathematics items qualified as Type I items (Marco, 1987).

Consistent with the finding of Linn and Drasgow (1987), Marco (1987: 1988) also found a curvilinear relationship between the proportions of correct answers for black examinees on test items and the differences between the proportions correct for black and white examinees on the items: items which were easiest and items which were most difficult for blacks had smaller differences in proportion correct between the two groups than did items of middle difficulty for blacks. This finding contradicts the implicit assumption of the Golden Rule procedures that elimination of difficult items will minimize the difference between the scores of blacks and whites. The finding also implies that selecting items with the smallest differences in proportions correct will tend to eliminate items of middle difficulty, yet there items typically have high point-biserial correlations, and

contribute substantially to a test's reliability and measurement efficiency (Marco, 1988). These findings are consistent with the predictions made by Bond (1987), Jaeger (1987), and Linn and Drasgow (1987) concerning the potential effects of application of the Golden Rule procedures on the psychometric properties of tests.

In a study pertinent to investigation of the effects of application of the Golden Rule procedures when constructing tests, Hackett, Holland, Pearlman, and Thayer (1987) investigated the effects of manipulating score differences for black and white examinees, using data from four experimental sections of a graduate-level admissions test. They constructed two sets of deliberately "biased" tests: one set constructed to maximize the difference between scores for blacks and whites, which are therefore "biased"* against blacks; and one set altered so as to minimize differences between scores for blacks and whites, which may be considered either less "biased" against blacks, or "biased" against whites, depending on one's viewpoint. Each set contained two tests. In constructing these tests, content representativeness and average item difficulty level were specifically protected. Only results relating to the tests designed to minimize score differences between groups will be discussed here, as these have implications for the Golden Rule test assembly procedures. In considering the results of this study, the reader should be aware that, although over a thousand white examinees took each manipulated and unaltered test form, the numbers of black examinees taking the tests were much smaller. Only 64 black examinees took the test forms altered to minimize score differences between groups.

---

*The term "biased" is surrounded with quotation marks here because the manipulated construct was actually differential impact on total score. The biasing effects of these test construction practices were not investigated in the study cited.

Hackett, Holland, Pearlman, and Thayer (1987) found that the average item-test point-biserial correlations for the tests constructed to minimize the score difference between black and white examinees were lower than those of the unaltered test forms. The average item-test point-biserial correlations for the two tests constructed to minimize the score difference between black and white examinees were 0.38 and 0.26, while the average point-biserial correlations for the corresponding unaltered tests were 0.46 and 0.33, respectively. As would be expected, the reliabilities were also lower for the tests constructed to minimize the score difference between black and white examinees than for the unaltered test forms. Alpha coefficients of 0.71 and 0.58 were observed for the two tests constructed to minimize the score difference between groups, while the average coefficients for the corresponding unaltered tests were 0.77 and 0.69, respectively. When reliabilities were examined separately for blacks and whites, the unaltered tests were about equally reliable for both groups (0.72 and 0.74 for blacks and whites, respectively, for one test; and 0.64 and 0.63, respectively, for the other). The reliabilities for the test forms manipulated to minimize score differences between groups were lower than those for the unaltered tests, but were again similar for blacks and whites (0.65 and 0.66, respectively, for one test; and 0.58 and 0.52, respectively, for the other). The test forms manipulated to minimize score differences between groups also showed slightly lower correlations with each of six other (unmanipulated) test section scores than did the corresponding unaltered tests. Interestingly, the test forms manipulated to minimize score differences between groups had slightly *higher* correlations, on average, with self-reported undergraduate grade-point average (GPA) than did the unaltered test forms. However, the correlations observed for the altered tests were within the

range of correlations observed for the unaltered tests (0.21 and 0.27 for the manipulated forms, versus averages of 0.18 and 0.25 for the unaltered tests). Also, the researchers noted that the non-response rate for self-reported GPA was substantial. Interpretation of these findings is ambiguous at best, since the test was not intended to predict undergraduate GPA. The researchers further noted that the test forms manipulated to minimize differences between groups were not demonstrably better than the unaltered test in any way, (including the opinion of expert content reviewers), and their assembly required substantial increases in the size of the item pool.

As the above studies illustrate, research to date on the consequences of applying the Golden Rule procedures has been limited to studies involving college or graduate-level admissions tests. The professional literature does not contain reports of investigations of the effects of applying the full Golden Rule procedures on the regression equation relating a test to another variable, on the differential regression, if any, of a test, or on the predictive or concurrent validity of a test.

Philosophical Context of the Golden Rule Strategy: Conceptualizations of Justice and Equality

As expressed in the introductory chapter of this study, there is a fundamental and deep-seated ambivalence in our society about the very nature of equity and the meaning of justice (Hartigan & Wigdor, 1989). This ambivalence is reflected in differing, and sometimes internally inconsistent, opinions on what constitutes just and proper action in specific circumstances. The controversies and diversity of opinion surrounding the issues of widespread use of the Golden Rule procedures, or variants of them, can best be understood

in light of the philosophical context of the issues. This philosophical context, and, to a limited extent, its historical derivation, are explored in the next section of this chapter.

Discussion of the nature of justice is the central theme of two dialogues of Plato -- the *Republic* and the *Gorgias*. The dispute explored in these two works is of such universal scope and fundamental character that all other discussions about the nature of justice can take place only after one or the other of these two extreme positions is abandoned. The dispute is between the proponents of might and the proponents of right -- between those who think that might makes right and that justice is expediency, and those who think that power can be wrongly as well as rightly exercised, and that justice cannot be measured by utility (Adler & Gorman, 1952).

Identifying power with right had a long history in Plato's era. Spinoza also voiced the opinion that "everything has by nature as much right as it has power to exist and operate." This thesis has two main corollaries. For the stronger, it means that they have the right, as far as they have the might, to extract from the weaker whatever serves their interests. Their laws or demands cannot be unjust. The thesis also means that the weak can only do injustice, by failure to obey the laws of their rulers. The weak cannot suffer injustice. For the weak, justice is also expediency, because they are likely to be made to suffer if they follow their own interests rather than their rulers' interests (Adler & Gorman, 1952).

Those who hold the opinion that might does not make right are confronted with two alternatives: either the principle of justice must be considered to be antecedent to the governing political state and its laws, or the determination of what is just or unjust must be considered entirely subject to the

state, and derived from its constitution and laws. The first position holds that there exist such things as natural justice and natural rights; the second position holds that justice and rights are merely and entirely political, derived from man-made laws. Under this second definition, the government itself, or specific laws, cannot be judged unjust (Adler & Gorman, 1952).

The position that there are no natural rights, and that there is no justice antecedent to the formation of social contracts, has been widely espoused. Thomas Hobbes, for example, was an adherent of this position, as seen in his 1651 work *Leviathan* (Hobbes, 1966). John Locke, however, believed that there exists a system of natural justice, as described in his 1689 work *Two Treatises of Government,* and that "being all equal and independent, no one ought to harm another in his life, liberty, or possessions" (Locke, 1964). According to Locke, government must be limited by mans' natural rights, and the function of the state is to protect human rights (Titus, 1970). These ideas were seminal influences on the Declaration of Independence, the Constitution, and the dominant ideology of the United States. Thomas Jefferson, in particular, advanced the Lockeian concept of the state as a contractual agreement among free and equal individuals, who had natural rights which the government should protect, and upon which it should not encroach (Hartigan & Wigdor, 1989; Titus, 1970). The beliefs that justice is more than the enforcement of laws, and that men are equal in that they all have some natural rights, is firmly entrenched in the dominant ideology of our society (Kluegel & Smith, 1986). Yet there is not consensus about the nature of justice, men's equality or their natural rights, or the limits, if any, that should be placed upon personal liberty.

Adhering to a position that man has natural rights and that justice antecedes government, Adler (1981) identifies the concept of justice as one of the six great ideas which have shaped western thought. He states, in fact, that the idea of justice is of paramount importance, as it regulates two other of the great ideas, liberty and equality. Rawls (1971) also identifies justice as the primary issue of social institutions. Justice, liberty and equality all ultimately derive from the idea of goodness: to act justly is to do what is good.

Of the three great ideas, liberty, equality, and justice, Adler (1981) states that only justice is an unlimited good. The primacy of justice among moral and political ideals is also held by Locke, Mills, and Rawls (Sandel, 1982). Adler (1981, p. 137) states: "One can want too much liberty and too much equality -- more than is good for us to have in relation to our fellowmen, and more than we have any right to. Not so with justice. No society can be too just: no individual can act more justly than is good for him or his fellowman." Adler (1981) further states that failure to understand the need for limitations upon and balance between liberty and equality leads to serious errors and irresolvable conflict. Those who place a supreme value upon liberty, whom Adler terms "libertarians," seek to maximize liberty at the expense of equality. They not only want an unlimited amount of freedom, they are willing to invoke irremediable inequality of conditions, in which some portion of society suffers serious deprivations, in order to achieve it. The only equality these "libertarians" support is equality of opportunity to compete, because this facilitates freedom of enterprise on the part of those who, favored by superior endowments or attainments, can make the best use of the competition to beat their fellowmen in the race of life. If put into practice, this would result in what Thomas Hobbes

(1966, p. 113) called the "war of each against all," a state of affairs he also described as "nasty, brutish, and short" However, this position of allowing some individuals to maximize their interests at the possible expense of others is also at the heart of utilitarianism, as espoused by Jeremy Bentham and others, which takes the position that justice consists in providing the greatest good for the greatest number. At the other extreme are those whom Adler terms "egalitarians," who not only regard equality of conditions as the supreme value, but are set upon achieving it even if it infringes in many ways on individual liberty, especially on the freedom of enterprise, exercised under equality of opportunity. In their view equality of opportunity to compete for the goods of society will inevitable result in inequality of conditions, thus is unacceptable. Marx, in advocating "from each according to his ability, to each according to his need," (Adler & Gorman, 1952, p. 854) was, of course, taking an "egalitarian" position. The conflict between these two extreme positions can only be reconciled by recognizing that neither liberty nor equality is an unlimited good, and that both can be maximized harmoniously only when maximization is regulated by justice. It should be noted here that justice stands in a different relation to liberty and equality. With respect to liberty, justice imposes a limitation on the amount of individual freedom that it allows, if the exercise of free will is to be just rather than unjust. With respect to equality, justice imposes a limitation on the kind and degree of the equality it requires if the community is to deal justly with all its members (Adler, 1981).

A variety of positions on what constitutes the proper balance of liberty and equality in justice have been espoused. Adler (1981) claims that our society, as a Jeffersonian democracy, entitles members to equality of status, treatment,

and opportunity, but not to equality of conditions. He states that we are also entitled to political liberty, and liberty of action, as it does not infringe on the rights of others.

Harvard professor John Rawls (1971) has written extensively on the nature of justice and its relation to the distribution of the goods of society. Rawls concludes, in part, that justice is synonymous with fairness in the dealings of individuals with one another as well as in actions taken by society in dealing with its members. Fairness here consists of treating equals as equals, thus the good of a minority cannot be sacrificed to benefit a majority. According to Rawls, just action is that action which will advance the interests of the disadvantaged the most. Injustice to any group or individual, in Rawl's view, is only acceptable if it is necessary to avoid an even greater injustice. Opportunities to compete should be open to all, but merit alone is not sufficient grounds for preference, according to Rawls. Although it is generally assumed that those with the greatest merit will be most productive, efficiency alone is not a sufficient grounds for the allocation of societal goods, unless it can be shown that such efficiency will most benefit the most disadvantaged sections of society (Rawls, 1971). This position is reminiscent of Aristotle's insight that if all human beings were associated in a community of friends there would be no need for justice. Adler (1981) has criticized Rawls, stating that, since Rawls discusses no natural right except equality, the Rawlsian concept of justice is inadequate. For instance, Adler maintains that Rawls would not consider a draconian punishment, such as execution for parking offenses, as unjust so long as the punishment was applied consistently to all members of society. This is probably an inappropriately narrow interpretation, however, since Rawls primarily addressed distributive,

rather than retributive or procedural, justice. In addition, this criticism does not abrogate Rawls's imperative that special consideration be afforded the disadvantaged sections of society. The great strengths of Rawls's concept of distributive justice are its straightforward assertion of the fundamental value of human happiness and well-being, and its utility for the derivation of rules or procedures by which questions of ethics and social policy can be resolved (Wolff, 1977).

Even among those who agree with Rawls that some considerations must be guaranteed minorities and the disadvantaged, there is conflict over the very nature of equality (Klugel & Smith, 1986). This is because equality is not a unitary concept. Three distinct types of equality can be distinguished: equality of treatment, equality of opportunity to compete, and equality of outcome. Guaranteeing equality of outcome regardless of contribution in economic spheres is generally regarded as unacceptably communistic in our society. However, the call for a guarantee of equality of outcome in employment selection and certain other arenas has been voiced by some members of society, and loudly denounced by others. Since the abolition of slavery, and the civil rights and feminist movements of the 1960's, acceptance of the right to equality of treatment has grown, but has still not been universally adopted. This can be clearly seen in actions ranging from the routine subjection of the poor and powerless to a degree of rudeness and depersonalization by police, medical personnel, and other public service workers that would not be tolerated by the rich or powerful. Our society is generally committed most strongly to equality of opportunity to compete. This type of equality can be acceptable to both "libertarians" or utilitarians, and "egalitarians." For many, fairness has to do with the rules of competition, rather

than the distribution of societal goods resulting from such competition. This concept derives from the utilitarian philosophy of economic liberalism, or "laissez-faire" economics, as developed by Adam Smith (Hartigan & Wigdor, 1989). The Founding Fathers of this country were, in most cases, deeply imbued with the laissez-faire spirit of individual freedom, minimal government, and the pursuit of rational self-interest. It is the very heart of capitalism, as is the pervasive belief that competition in a free-market economy is in the best interests of both individuals and society as a whole. Equality in this country is most often deemed the right to compete freely with others for society's economic and other rewards (Hartigan & Wigdor, 1989; Kluegel & Smith, 1986).

Policy Context of the Golden Rule Strategy: Issues of Merit and Preferential Treatment

The widely held beliefs of our society about the nature of justice and fairness, as described above -- that society is made up of individuals who should be treated equally under the law and that all should be allowed to compete freely for societal goods, have resulted in a firmly entrenched merit system in most spheres of our society. As Hartigan and Wigdor (1989, p. 32) state: "The concept of meritocracy has had great social approval over the years -- to the extent that we tend to forget that it is a construct and not a description of objective reality." Many never question the basic tenets of the American meritocracy. Hartigan and Wigdor (1989, p. 32) have delineated these tenets:

> 1. The goods of society should be awarded to individuals on the basis of merit.
> 2. The qualification that merits reward in the allocation of jobs is talent (ability, experience), not family connection, social class, political loyalty, virtue, need, or other criteria that are irrelevant to job performance.

3. Social, economic, and political structures should be designed to allow open competition among individuals.

4. A system of open competition and selection on the basis of competence satisfies both fairness and efficiency because every individual has the same chance to realize his or her potential regardless of birth or wealth and because all individuals will end up in the positions most suited to their talents.

5. Such a system is just because everyone has equal opportunity to compete for positions and is rewarded as he or she deserves.

There has been growing recognition over the years that the freedom to compete equally does not solely imply an absence of barriers based upon group affiliation; it also implies equality of life chances. This recognition is the foundation of our free public education system and of college scholarships for the needy. However, our society is so invested in the principles of a free-market economy that we fall far short of providing the same cultural and material advantages to all persons, particularly children, as would be necessary to truly foster talent or ability equally in all groups. There has been a persistent tendency in our social policies to promote minimum levels of social welfare, in the hope that "a decent minimum would help people to a better start in the race of life." Although social policy has been minimalist in practice, it has typically been conceived and described in egalitarian terms (Cohen & Haney, 1980, p. 5). Particularly in American public education, there is a great difference between rhetoric and practice: "Although we have a rhetorical commitment to a standard of excellence, the practical operation of the system has neither provided nor been expected to provide for universal attainment of excellence" (Jaeger & Tittle, 1980, p. 2). This disparity often creates confusion over the true nature and intent of social programs, and engenders disappointment and anger over the results (Cohen & Haney, 1980). This state of affairs is far from justice as conceived by Rawls and others who feel that special measures should be taken to ensure that

all members enjoy a share of the benefits of society. Edmonds (1979, p. 15), for example, would have us "measure our progress as a social order by our willingness to advance the equity interests of the least among us."

Less powerful segments of our society, notably blacks and, to a lesser extent, women, have been the victims of enormous systematic injustice for decades. Recognition of this injustice, and of the inequities of outcomes and life chances that still exist among these groups, has led some to advocate preferential treatment for these groups. Advocates of preferential treatment believe that compensation is due for this historical inequality, and that "the long history of unequal treatment has left blacks as a group so educationally, economically, and psychologically disadvantaged that, without special preference, they will be condemned by our newly color-blind society to remain de facto second-class citizens," (Hartigan & Wigdor, 1989, p.36). Preferential treatment for minorities will, of course, sometimes work against the interests of individual members of the advantaged majority, who will not be competing with minority group members on equal (i.e., free), and therefore beneficial, terms. This does not constitute an injustice for Rawls or others who reject total "libertarianism" or utilitarianism. However, opponents of preferential treatment are equally convinced that it is unjust to members of the majority to alter the rules of equal (i.e., free) competition, and feel that preferential treatment of any sort destroys the foundations of justice and equality, by creating "reverse discrimination," (cf. Hartigan & Wigdor, 1989; Menacker, 1987). They also argue that the individual beneficiaries of preferential treatment practices may themselves never have been the victims of discrimination, and those who incur the various costs of the preferential treatment program may never have practiced discrimination.

Further, Kluegel and Smith (1986) state that, since the middle 1960's, there has been a steady increase in the proportion of white Americans who believe that blacks do not currently face inequality of opportunity, and who attribute the difference between the socioeconomic status of blacks and whites to failings of blacks as individuals, particularly to lack of motivation.

The basic ambivalence in our society about preferential treatment, and the dichotomization of views of justice as it relates to disadvantaged subgroups, has led to a series of inconsistent and self-contradictory practices and federal policies regarding adverse impact. As Klugel and Smith (1986, pp. 1-2) state:

> It has often been remarked that Americans' attitudes about social welfare and inequality-related policies have an inconsistent and sometimes contradictory quality. The most recent example is, of course, the sharp change in direction of federal policy associated with the Reagan administration's goals of curtailing many of the redistributive programs developed since the New Deal. There are other examples as well. Americans generally accept the idea that blacks and other minorities have suffered from discrimination and maintain an abstract commitment to equal opportunity -- coexisting with widespread opposition to specific policies to implement equal opportunity (e.g. busing to desegregate schools, affirmative action programs). Although Americans highly value equal citizenship rights and democratic politics in the abstract, in practice the right of the wealthy to wield disproportionate economic and political power is unchallenged.

This ambivalence is illustrated nowhere so clearly as in the policies regarding employment selection. Discrimination in employment on the basis of race, religion, national origin or gender is prohibited by the Civil Rights Act of 1964. Section 703(j) of Title VII of the Act, however, states specifically that the Act does not require any employer to give preferential treatment to any group. Congress appeared to approve some uses of preference, however, when it amended Title VII in 1972 (Local 93, International Association of Firefighters v.

City of Cleveland, 478 U.S. 501, 543 [1986] [Rhenquist dissenting]), but did not take an unambiguous position, leaving employers open to suit by minorities if they do not give preferential treatment, and by members of the majority if they do (Hartigan & Wigdor, 1989).

Although Title VII was designed to protect the rights of individuals, Congress recognized that litigation was not an easily accessible avenue of redress for victims of widespread and deeply entrenched discrimination. Hence, Title VII not only gave individuals the right to sue employers on grounds of discrimination, it empowered the Attorney General to bring civil suit if an employer appeared to engage in a pattern of discriminatory practice. These "pattern of practice" suits were based largely upon work-force statistics. In 1966, the Equal Employment Opportunity Commission (EEOC), created by the Civil Rights Act, took a position interpreting Title VII to prohibit not only intentional discrimination, but also any practices that had an unintentional adverse impact on protected groups. This position was reaffirmed in 1971, in Griggs v. Duke Power Co. (401 U.S. 424 [1971]). These opinions, however, specifically state that there is no requirement for preferential treatment of minorities, and that applicant qualifications should be the controlling factor in employee selection. Thus, there is an important unspoken assumption of these federal policies that there is an underlying comparability in the distributions of career preparation in our society. This is voiced in Teamsters v. United States (431 U.S. 324, 342 n.20 [1970]):

> Absent explanation, it is ordinarily to be expected that nondiscriminatory hiring practices will in time result in a work force more or less representative of the racial and ethnic composition of the population in the community from which employees are hired.

Empirical evidence, however, indicates that major groups of non-Asian minorities in the United States perform less well than members of the majority on a wide variety of ability tests and performance indicators. As Hartigan and Wigdor (1989, p. 42) state: "Those who take seriously the effects of the kind of extreme economic, educational, and cultural disadvantage experienced by most blacks even today do not find this information surprising." It would be naive to expect the removal of discriminatory hiring practices alone to result in balance in the work force. In fact, the National Research Council has concluded that, despite various social and affirmative action programs, the disparity between black and white Americans in terms of standards of living, health, and education has increased since the early 1970s (Jaynes & Williams, 1989). In 1968, the National Advisory Commission on Civil Disorders stated that we were "moving toward two societies, one black, one white -- separate and unequal;" the rate of this movement has increased rather than decreased in the past decade (Shapiro, 1989, p. 12).

The ambiguity of the national position on the meaning of justice as it relates to equality is further shown by the 1972 requirement that each federal department develop an affirmative action plan, while in the private sector it was not until 1978 that the Supreme Court recognized the legality of even voluntary affirmative action programs. Last year, the Supreme Court handed down a series of decisions which essentially reversed its earlier position prohibiting unintentionally discriminatory practices, by requiring that discriminatory intent on the part of the employer be shown (*Indep. Fed'n of Flight Attendants v. Zipes*, 1989; *Jett v. Dallas Indep. School District, 1989; Lorance v. AT & T Technologies, 1989; Martin V. Wilks*, 1989; *Patterson v. McLean Credit Union*, 1989; *Price-Waterhouse v.*

Hopkins, 1989; *Wards Cove Packing Co. v. Antonio*, 1989). These decisions were regarded by many as a severe blow to efforts to secure the rights of minorities and women (cf. Amaker, 1989; Coyle, 1990; Hemeryck, Butts, Jehl, Koch, & Sloan, 1990). Congress responded by proposing the Civil Rights Act of 1990, which, among other provisions, would have removed the burden of from plaintiffs to show intentional discrimination on the part of defendants in cases alleging discriminatory employment practices (Coyle, 1990; Hemeryck, et. al., 1990). The House (H.R. 4000) and Senate (S.R. 2104) versions of the Civil Rights Act of 1990 were passed, but the Act was subsequently vetoed by the President. Despite the direction of the 1989 Supreme Court decisions cited above, the Circuit and Supreme Courts have repeatedly held disparate impact to be an important starting point in evaluating whether there has been intentional discrimination (cf. *Columbus Board of Education v. Pennick*, 1979; *Diaz v. San Jose Unified School District*, 1984; *Personnel Administrator v. Feeney*, 1979; *United States v. LULAC*, 1986; *Village of Arlington Heights v. Metropolitan Housing Development Corp.*, 1977). In decisions regarding allegations of discriminatory employment practices, courts have ordered a wide variety of race-conscious remedies, including the imposition of numerical hiring goals and timetables, promotion ratios, score adjustments, and alternative selection procedures. The court has repeatedly stressed that affirmative action plans are uniformly intended to be temporary and remedial (Hartigan & Wigdor, 1989).

Perhaps because the government has not been able to establish a uniform policy on discrimination and affirmative action, it has devoted a great deal of energy to scrutinizing the instruments, such as employment tests, that are the proximate cause of adverse impact. The *Uniform Guidelines on Employee Selection*

Procedures, published by the EEOC in 1978, lays out detailed requirements for the validation of employment tests and procedures. Generally, differences in average performance between the races has been considered insufficient to establish a violation of Title VII (Hartigan & Wigdor, 1989). However, the unified government position concerning employee selection procedures, as stated by the Equal Employment Opportunity Commission, the Civil Service Commission, the Department of Labor, and the Department of Justice (1978), is that adverse impact of an employee selection practice results in the obligation that the employer show that the selection practice is validly related to successful job performance. These guidelines use a practical "rule of thumb," called the "80% rule," which states that if the selection rate for one group is less than 80% of the rate for the group selected most frequently, federal enforcement agencies will generally consider this as evidence of adverse impact, and impose strict scrutiny for discrimination. Allegations of discrimination based solely on differences in average performance between the races have also involved test users and manufacturers in costly litigation, and have resulted in out-of-court settlements that made significant concessions to the plaintiffs despite the opinion of a recent federal appeals court decision in a testing case (*United States v. LULAC,* 1986) that "an action does not violate the equal protection clause simply because the decisionmaker knows that it will have a disparate impact on racial or ethnic groups" (Anrig, 1987b, p. 25).

The allocation of jobs is an obvious and highly visible case of distribution of a societal good, invoking a definition of equality and of justice. Although the above discussion centered on discrimination in employee selection, an area of particularly intense governmental regulation, the ambiguities and conflicts

illustrated by that discussion have been present in other arenas of distributive justice, particularly those involving standardized tests. Similar questions regarding disparate impact have been raised in situations involving tests used for admission to educational institutions or programs, placement tests, and tests used in determining a student's matriculation or retention in grade.

Standardized testing in the schools is pervasive and on the increase. As Pipho (1985, p. 19) noted: "Nearly every large educational reform effort of the past few years has either mandated a new form of testing or expanded the use of existing testing." Many, including minority groups and organizations and professionals in the measurement community, are legitimately concerned about the use of standardized tests as they relate to equity and access to educational and ensuing life opportunities (Howe & Edelman, 1985; McAllister, 1987; NCTPP, 1990). First and Cardenas (1986, p. 6) have concluded that "test scores are being widely used for a variety of inappropriate purposes in making decisions about students, teachers, and state and local programming. The result, we think, is that testing is often having a harmful impact on education and particularly on the interests of minority and special needs students." The proportions of minority students enrolled in public schools are increasing, such that, by the year 2000, it is estimated that minority groups will represent a majority of the student populations in more than 50 major cities. However, educational attainment of blacks is declining. For example, the number and percentage of blacks earning doctoral degrees, (which require successful completion of innumerable standardized tests), has decreased from 9.2% in 1975 to 7.0% in 1986 (Frierson, 1990). Concern over these issues has led to the formation of a number of private organizations, such as Fairtest, to monitor the

activities of testing companies, and to attempt to promote the regulation of standardized testing through litigation and legislative initiatives (McAllister, 1987). Child advocacy organizations, such as the National Coalition of Advocates for Students which is composed of "twenty-four groups concerned with promoting maximum student access to appropriate educational experiences," have expressed increasing concern over the ill effects and educationally counterproductive nature of educational testing programs (First & Cardenas, 1986, p. 6). In his Presidential Address at the 1989 annual meeting of the National Council on Measurement in Education, the flagship organization in educational measurement, Irvin Lehmann presented one of the tasks of the organization to be to "undertake a reexamination of the philosophical and value-related assumptions underlying educational measurement. ...as a professional organization, NCME should undertake a reexamination of the intents of measurement" (Lehmann, 1990, p. 7).

Charges have been made that, due to their "narrow and rigid definitions both of when children should be able to perform particular skills and how they should be able to exhibit their knowledge," standardized tests devalue "the variety of strengths they [children] bring with them to school" (First & Cardenas, 1986, p. 7). "All differences become handicaps" (First & Cardenas, 1986, p. 7). Linn, Madaus, and Pedulla (1982) have criticized over-reliance on student competency tests in particular, arguing that the standards set for such tests had been shown to vary "substantially across the methods used to derive them and the types of judges used to recommend them" (Jaeger, in press, p. 10). This led Linn, Madaus, and Pedulla (1982) to conclude that the results of administration of a competency test might have more to do with the method used to set the

standards for the test than with the abilities of the group being tested. Jaeger (in press) concluded that the benefits of the use of competency tests are arguable, while the adverse impact of the use of such tests on racial minorities and the poor is quite clear.

Standardized testing has also been charged with inhibiting teaching, since teachers feel compelled to tailor their teaching to the objectives which will be tested (cf., Glickman & Pellegrini, 1988; NCTPP, 1990; Rottenberg & Smith, 1990; Romberg, Zarinnia, & Williams, 1989). This effect is often augmented by the fact that the teachers and schools, as well as their students, are judged on the basis of the students' test scores. There have even been instances of local programs to provide a salary bonus to teachers whose classes showed the greatest yearly gains on standardized test scores (First & Cardenas, 1986). Such misuses cause testing to be seen as a barrier to providing a curriculum based upon individual students' needs. This impact is seen as particularly severe for minorities, since tests are often used to make classifications or distinctions among students, such as tracking or promotion decisions, but "the classification is rarely followed by effective educational support for students who are identified as 'at risk,' in need of remedial help, or not ready for promotion" (First & Cardenas, 1986, p. 7). Without application of appropriate educational remedy, students classified as "lacking" in some way only suffer stigma. Disproportionately , such students are minorities. Allegations have been made that some schools boost their average test scores by using a variety of methods to remove low-scoring students from the test pool; and that some schools do little to prevent or follow-up dropouts, as their absence tends to raise the schools' average scores (First & Cardenas, 1986).

Such practices would particularly affect blacks, since they are disproportionately represented among dropouts and low-scoring students (Jaeger, in press).

Charges of discriminatory use of tests in education have also led to litigation, although somewhat less frequently than in cases involving employment practices, due in part to the fact that "the right to engage in a legitimate occupation is a liberty right that can be denied by the state only by reasonable standards to protect the public health, safety and welfare" (Reutter, 1985, p. 440); and in part to lack of clarity about the applicability and extent of constitutionally protected rights due to minors and students (Menacker, 1987). However, some educational cases have received great public attention, such as *Hobsen v. Hansen, Larry P. v. Riles,* and *Debra P. v. Turlington* (First & Cardenas, 1986). These suits have been characterized by the same contradictory positions as the employment cases discussed earlier.

The growing practice of requiring public school students to pass competency tests for graduation or promotion regardless of whether all required courses have been satisfactorily completed, and similar requirements applied to the professional practice of teaching, have met challenges based on violation of legally protected rights to equal protection and to due process, as well as violations of the related statutory provisions found in the Civil Rights Act of 1964, the Rehabilitation Act of 1973, and the Education for All Handicapped Children Act of 1975 (Menacker, 1987; Reutter, 1985).

In the case *Alba v. Los Angeles* (1983), probationary teachers who had failed a competency test required by their school district sued, charging that the testing practice was unfair for several reasons: they were not administered all sections of the test which had been advertised and for which they had prepared; the passing

score for the test had been arbitrarily raised; and they had not been informed of the alternative of passing a section of the National Teachers Examination in lieu of the competency test. The original trial court ruled that the school district had acted arbitrarily and capriciously, but this decision was subsequently reversed by the California appellate court, which ruled that the testing practice was not unfair (Menacker, 1987).

The due process issue of sufficient notice (among other issues) was raised by black students who failed Florida's high school competency test, who charged that, given the nature and content of the test, notice of the test should be given to students before they entered high school, not after, as had occurred in their case. The trial court concurred, as did the appellate court, which further decided that the district court had the responsibility to determine whether the test itself fair, and whether it fairly assessed knowledge and information actually taught in the schools. Subsequent investigation, in the case *Debra P. v. Turlington* (1981), led to the conclusion that the test in question was a reasonably fair and valid measure of the curriculum (Menacker, 1987).

A subsequent case involving teacher education shows similarly conflicting positions. In 1981, Texas adopted a law requiring use of a competency examination for admission into state-approved teacher education programs. By 1985, more than 18,000 students had attempted the examination. Passing rates at that time were: 73% for white students, 34% for Hispanics, and 23% for black students. There were several state-accredited institutions of higher education in which less than 10% of students passed the examination. Suit was brought, charging that the state had arbitrarily adopted cut scores on the test, and that it had failed to give adequate notice of the test, to establish the validity of the test,

or to organize any program of remediation despite clear knowledge of the impact of the test on minority students (*LULAC, GI Forum, & NAACP v. State of Texas,* 1985). A federal district court judge imposed a temporary restraining order, prohibiting use of the test until its merits had been examined in court. The U.S. Department of Justice then filed a brief in the U.S. 5th Circuit Court of Appeals in New Orleans, arguing that the restraining order should be overturned, stating that "you don't solve the educational problems of minority students by holding them to a double standard of education" (First & Cardenas, 1986, p. 8).

In *Board of Educ. of Northport-East Northport Union Free Dist. v. Ambach* (1982) and *Brookhart V. Illinois State Board* (1983) it was judged a violation of due process to deny regular high school diplomas to handicapped students, whose educational experiences were organized according to individualized educational plans, on the basis of a general competency test. The handicapped students had been given notice of three years and one-and-a-half years, respectively, in the two cases, and this was ruled insufficient. Challenges regarding the validity of the competency test were rejected in both cases, and no violations of equal protection were found (Menacker, 1987).

The issue of due process in competency testing was raised again in *Anderson v. Banks* (1982), in which a federal court in Georgia barred used of the California Achievement Test as a graduation requirement unless the district could prove that the test content corresponded to what was actually taught in the schools. District officials were eventually able to do this to the court's satisfaction, and use of the test was approved (Menacker, 1987).

Hobsen v. Hansen (1967), sought to apply the equal protection concept implicit in the Fifth Amendment to the question of whether equal protection was

denied to minority and poor students placed in the lowest school tracks because of low scores on standardized tests, which were alleged to be biased in favor of the white middle class. The decision reached by the presiding justice in the case agreed that this was a case where "constitutional rights hang in the balance," (Menacker, 1987, p. 217), that this use of standardized tests did constitute discrimination because the tests used were primarily relevant to "a white, middle-class groups of students," and that such use of the tests resulted in a system of assigning minority students to the least-desirable educational experiences, and thus was "completely inappropriate" (p. 218).

The decision in *Hobsen v. Hansen* (1967) "sent shockwaves through much of the public-school community, causing many districts to reassess their testing and tracking practices" (Menacker, 1987, p. 219). A subsequent case in California, *Diana v. State Board of Education* (Civil No. C-70-37 RFR N.D. Cal. [1970]), was brought by parents of Spanish-speaking children who had been classified as mentally retarded on the basis of intelligence-test results. This case also claimed violation of the equal protection clause, on the bases that, first, English was not the children's primary language, and, second, the tests were culturally biased against Mexican-American students. This case was resolved in an out-of-court settlement in which the California Board of Education agreed to restandardize the tests, and to test these students in their primary language as well as in English (Menacker, 1987).

Following the out-of-court settlement of *Diana v. State Board of Education* , the case *Larry P. v. Riles* emerged to influence testing practices from 1972 until 1984. This case, which questioned whether the major standardized intelligence tests and the policies and procedures associated with them discriminated against

black students classified as Educable Mentally Retarded primarily on the basis of the tests, was shuttled back and forth between federal district courts and appellate courts for years. Relying heavily on the testimony of experts, district court Justice Peckham eventually ruled that the tests were indeed culturally biased, that they played too powerful a role in placement decisions involving students classified as Educable Mentally Retarded, and that they had not been validated for such placement decisions. These factors were judged to be the causes of the disproportionately high classification of black students as Educable Mentally Retarded. The court ordered the California schools to stop using the tests for this purpose, to reevaluate black students currently classified as Educable Mentally Retarded, and to develop plans to *eliminate* racial disproportion in Educable Mentally Retarded class enrollments (Menacker, 1987).

The finding in the case of Parents *in Action on Special Education (PASE) v. Hannon* (1980) was in contrast to the decisions in *Hobsen v. Hansen, Diana v. State Board of Education,* and *Larry P. v. Riles,* In this case, which also dealt with the question of discrimination in the use of intelligence tests to classify black children as mentally retarded, Justice Grady of the seventh circuit district court concluded that the small number of items he judged to be biased were insufficient to seriously affect a student's classification. Rather than relying on the testimony of experts as Justice Peckham had, Justice Grady himself evaluated the items of the Stanford-Binet and Weshsler tests for bias. His ruling allowed the testing practices in question to continue; however, the Chicago School District entered into an out-of-court settlement in which they agreed to discontinue the use of the tests for classification of black children, in exchange for the plaintiffs' promise not

to appeal Justice Grady's decision (Menacker, 1987). Thus, as Menacker (1987, p. 222) states, "the issue of test bias remains cloudy, with contradictory decisions in two federal circuits."

Many of the uses and misuses of standardized tests described above are clearly at odds with the *Standards for Educational and Psychological Testing* (APA et al., 1985, pp. 41-54). The *Standards* deal directly and extensively with the obligation of test users to validate specific test uses, such as student promotion, retention or classification. They state that when tests are used for certification purposes in schools, students' prior instructional opportunities be demonstrated. When a decision or characterization will have a major impact on a student, they stress that the decision should not be based upon a single test score. They require that differential prediction of test scores be investigated when student numbers are sufficient to do so, and that explanations for poor test performance other than ability level, such as socioeconomic or cultural background, be considered before judgments are made. However, First and Cardenas (1986) state that, despite the availability of professional testing standards sponsored jointly by APA, AERA, and NCME for the past 30 years, such standards have had little impact on educational testing at the state and local levels. They suggest three general guidelines to help prevent the abuse of standardized testing in education (p. 9-10):

> First, standardized testing should not be used in isolation, but instead should be coupled with other assessment techniques, such as teachers' observations and academic records that relate directly to instruction. Second, before new testing programs are implemented on a broad scale, to help inform student promotion, graduation, or placement decisions, or to help evaluate teachers, schools, or educational programs, educational models for what will result from such classifications and the resources to implement those models should be available. Third, state and local testing programs should be monitored regularly for their impact on

> minority, culturally different, and special education students and on curriculum and teaching practices. This last step would help insure that the consequences of new testing programs do not fall most heavily, by default, on relatively powerless students.

Despite decades of social welfare and affirmative action programs, minorities, particularly blacks, do not enter the "race of life" today on an equal basis with the white majority. They are disadvantaged due to lower standards of living, health, and education; and these disparities have increased, rather than declined, since the early 1970s (Jaynes & Williams, 1989). Nonetheless, majority resentment over social welfare programs, especially affirmative action programs, has become a powerful undercurrent in race relations. Many whites consider discrimination a thing of the past, and some even believe that blacks now have an unfair advantage in competing for societal goods which, they feel, should be awarded solely on the basis of merit (Kluegel & Smith, 1986). However, a report by the National Research Council concludes that, aside from a few well-publicized anecdotal examples, virtually all the evidence contradicts this popular misconception. The report states that there remains a considerable amount of overt discrimination against blacks, and that whites are much more likely to support societal integration in theory than specific governmental steps to achieve it (Jaynes & Williams, 1989). As Shapiro (1989, p. 15) states, "the implicit message ... is that white America, left to its own devices, will never complete the unfinished task of creating racial equality".

Such is the contentious policy context of the Golden Rule strategy, and the legislative efforts to extend it to other arenas. In this context, standardized tests are held by many to be powerful and objective tools for detecting merit and imposing accountability. Others consider them to be mechanisms for

perpetuation of an inequitable status quo, and a means by which society sets standards to which the individual, with only minimal assistance, is held responsible (Cohen & Haney, 1980; First & Cardenas, 1986; NCTPP, 1990). These divergent points of view influence concepts of what constitutes discrimination and bias in testing, and thus impact the scientific context of the Golden Rule procedures.

Scientific Context of the Golden Rule Procedures: Bias and Adverse Impact

Three types of equality were distinguished in the discussion of equality above: equality of treatment, equality of opportunity, and equality of condition. In the sphere of education, one can distinguish three parallel types of equality: equality of educational provision, equality of educational opportunity, and equality of educational outcomes or results (Karmel, 1985). The argument for guaranteeing equality of educational provision alone is based on the belief that, if children are provided similar school resources, their success will be determined by their individual effort and innate ability. This view of educational equality is held to be the proper one by some, typically by members of the majority. The argument for the provision of equality of educational opportunity is derived from the recognition that some groups of children are subject to educative disadvantages which are not experienced by others, and which are beyond the control of the school. The concept of equality of educational opportunity justifies a greater investment of time and resources in the education of disadvantaged children. This notion is espoused by some and rejected by others as "unfair" to majority, non-disadvantaged children. If we were able to provide majority and minority students with perfect equality of educational provision and educational opportunity from birth, it should logically result in equality of educational

outcomes between groups (unless, as discussed by Shepard (1982), one believes in biological determinism: the innate superiority of one group). However, the current social order is far from providing equality of educational provision or educational opportunity, thus equality of educational outcomes is unlikely (Karmel, 1985).

There are some who take the position, often implicitly, that equality of educational outcomes is a right, and should be achieved artificially if it does not arise in the natural course of events. Advocates of this position do not find affirmative action plans, which attempt only to compensate for accrued disadvantagement, to be adequate. This position holds that it is not the opportunities or the life chances of groups which should be equalized, but their actual outcomes in the various competitions of life. (This position, for example, would endorse manipulating test content and item characteristics so as to guarantee that minority and majority examinees achieved the same mean score on a test, and, if the test is a competency test or selection test, that equal proportions of majority and minority examinees were certified as competent or selected.)

Karmel (1985) and others, view equality of educational outcomes as a goal for which to strive, and as an indicator of the effectiveness of efforts to provide greater equality of life chances. This position supports affirmative action and preferential treatment practices, which are directed at provision of equality of educational opportunity, but would not support any efforts to attain equality of educational outcomes artificially (i.e., by manipulating the content or item characteristics of tests solely to obtain equal mean scores), since such efforts would obscure the true state of inequality of educational provision or opportunity which led to disparate outcomes. As Shepard (1987, p. 7) states:

"Once one agrees to prespecify group differences, the test can no longer be considered an honest measure of those differences (or of anything correlated with them)."

The concept of test bias is closely linked to the definitions of educational equality discussed above. In his discourse on the meaning of the concept, Flaugher (1978, p. 671) wrote:

> The definition of test bias -- the inventory of the ways in which the term is used -- has many widely disparate aspects frequently stemming from entirely different universes of discourse ... It is essential to keep all of these aspects in mind, for we continually run the risk of losing perspective when we settle on one operational definition of test bias and then proceed to forget that it is only that. No matter what definition we use, because the concept is a public one we are never going to encompass all that it contains.

The "different universes of discourse" to which Flaugher refers are in fact different conceptualizations about the nature of equality and its relationship to justice. In the late 1960s and early 1970s, psychometricians responded to this general lack of clarity about the meaning of bias by adopting precise, often narrow, operational definitions of bias (Berk, 1982; Cole & Moss, 1989). This led to a proliferation of alternative, often competing, methods to detect test bias which has not abated to this day. (In 1976, an entire issue of the *Journal of Educational Measurement* was devoted to issues of bias (Jaeger, 1976).) One consequence of this abundance of definitions of bias and attendant detection methods is that it is easy to find some basis for directing a charge of bias against any measurement instrument, but it is nearly impossible to obtain incontrovertible evidence to substantiate or refute the charge (Berk, 1982).

Like Flaugher (1978), Shepard (1982, p. 9) warns against the adoption of rigid operational definitions of test bias, stating that Scriven "debunked

operational definitions because they trade simplicity and clarity for accuracy, by trying to represent too simply and concretely the original more complex and more important concepts." Because operational definitions are simplifications, they capture some, but not all, of the meaning of the concept; their very simplicity and clarity are sources of inadequacy (Scriven, 1988). Consistent with this view, several aspects of test bias, and a number of the meanings ascribed to the term, will be examined in this chapter. A unified concept of test bias, provided by Cole and Moss (1989) will then be used to relate and synthesize these alternative definitions of test bias.

As Bond (1981), Shepard (1982), Angoff (1989), and others have noted, both the professional literature and public discourse have tended to approach issues related to test bias from one of two perspectives. From one perspective, entire tests are considered potentially to be biased (cf. Bond, 1981); from the other perspective, individual test items are considered potentially to be biased (cf. Berk, 1982). Although writers and researchers have tended to address issues related to test bias from one or the other of these perspectives, the two positions are not unrelated or mutually exclusive. For example, an entire test would generally be considered biased if a sufficiently large proportion of the items comprising it were considered to be biased.

The professional literature and public discourse also reflect a variety of positions on the continuum of whether bias resides in a test (or test item), or in a specific use of the test (Shepard, 1982). Some writers have sought to clarify this issue by distinguishing between bias in testing, which they view as residing in the test; and unfairness in testing, which they view as consequent to a specific use of a test (cf. Jensen, 1980; Reynolds, 1982b). Some writers have even

suggested that it is possible to use a biased test either fairly or unfairly. Shepard (1982, p. 10) judges this distinction to be problematic for two reasons:

> First, everyday understandings of the two key terms is do not unequivocally convey the intended difference. To be biased is to be unfair, unjust, prejudiced. Calling your test "biased" conveys very nearly the same message as a placard calling an employer "unfair". ... The distinction between bias that is somehow *in* the test and unfairness that is in the *use* of the test is also an awkward distinction for psychometricians to have made. Most authors define bias as a type of invalidity (Green, 1975; Reynolds, 1980). Bias, however, is now being taken as an inherent feature of a test, while its opposite, validity, has always been considered to be a property of test use, not of the test itself.

The distinction between bias and fairness discussed above is also problematic in that it uses the term *fairness* in a manner different from its use in the broader philosophical literature (cf. Rawls, 1971; Wolff, 1977). Shepard (1982, p. 10) suggests that the muddle which results from attempts to separately define and distinguish bias and unfairness can be avoided while remaining "faithful to the rule that validity must always pertain to the particular inferences made from a test." She suggests that we "admit different degrees of externality through which bias may be more or less closely associated with the use of a test, rather than its internal characteristics" (Shepard, 1982, p. 10.) Shepard suggests that we conceptualize a validity continuum for tests, anchored at one end by unbiased tests that measure what they are intended to measure, and do so equally well for all groups. Near the other extreme are tests with internal characteristics or psychometric properties such that any interpretation of scores on the test is always of suspect validity for some group(s). In between are tests that provide valid prediction or description for all groups under some circumstances, but may

be of questionable validity under other circumstances. At the extreme end of this continuum are tests whose unfairness can only be resolved by resolving issues of social justice and values. Using this schema, it is always the test use (or interpretation of the test results), rather than the test itself, which is actually judged biased or unbiased. In some cases, however, it is clear that biased interpretation is closely associated with characteristics of the test, (and hence could potentially be corrected through alteration of the test itself). In other cases, the biased interpretation is associated with the conditions or context of testing, and cannot be addressed through alteration of the test itself.

Keeping these considerations in mind, the various ways in which the term bias has been used by the public or in the professional literature will be reviewed in the next section of this chapter.

Definitions of Bias

Writing in 1978, Flaugher identified eight distinct concepts of bias promulgated by the public or in the professional literature. These are: 1) test bias as differences in distributions of scores; 2) test bias as overinterpretation; 3) test bias as sexism; 4) test bias as content; 5) test bias as atmosphere; 6) test bias as the selection model; 7) test bias as the wrong criterion; and 8) test bias as differences in test-criterion correlations. Each of these definitions has inherent assumptions about the nature of equality. Each is also still in use in some form today, although the terminology has changed in some cases, and further concepts of and definitions of bias have been added. These concepts of bias are discussed individually in the next section of this chapter, followed by a discussion that relates them to a unified concept of test bias provided by Cole and Moss (1989).

Differences in Distributions of Scores

According to Flaugher (1978), test bias as differences between the distributions of scores of groups is not a legitimate standard for identifying bias, since the current inequities of educational provision and opportunity almost guarantee discrepancies in educational achievement. (Discussions of this concept typically focus on the mean scores for each group, since the mean is a generally recognized measure of typical performance, but the arguments and considerations apply equally to median scores, or to overall distributions of scores.) Many others in the measurement profession agree that a difference between the distributions of scores of groups is not synonymous with bias (cf. Bond, 1987; Jaeger, 1987; Karmel, 1985; Reynolds, 1982a; Reynolds & Brown, 1984; Shepard, 1981, 1982). However, Flaugher states that this definition of bias cannot be dismissed or ignored, because it is the premise from which many members of the public start. According to Flaugher, if it were possible to construct an achievement-free aptitude test, the standard of equality of group means could legitimately be applied, but achievement-free aptitude testing is a myth. Flaugher, like Karmel (1985), states that discrepancies between the distributions of scores of minority and majority examinees is important evidence that the legitimate goals of achieving equality of educational provision and opportunity have not been attained.

In some cases, this definition of bias is espoused out of naivete; in others it reflects the belief that society should guarantee equality of educational outcomes, as well as equality of educational provision and educational opportunity. Although not accepted as a definition of bias by the scientific community, this focus on the difference between distributions of scores for groups of examinees

lies at the heart of the Golden Rule stipulation which specifies a maximum acceptable difference between the proportions of correct answers for black and white examinees for each item.

Bias as Overinterpretation

Flaugher (1978) next discusses seven other definitions of test bias, which encompass many of the ways in which the term is used by professionals. Overinterpretation is discussed as a type of test bias by both Flaugher (1978) and Bond (1981). Flaugher (1978) noted that the worthwhile attributes of human behavior comprise a wide spectrum, and that a great discrepancy exists between the portion of that spectrum which can be accurately measured, and the portion which the public thinks can be measured. Indeed, a discrepancy exists between what can be measured and what the profession implies *is* measured. Tests are often designated by the constructs we wish them to measure, rather than by their content. Thus, for example, a test of commonly encountered problems is called a test of "practical judgment." If the examinee is required to offer suggestions or generate solutions, the word "creativity" may appear in the title of the test (Flaugher, 1978). Similarly, tests composed of academic problems are called tests of "general intelligence" or "mental ability" (Bond, 1981). Yet, as Flaugher (1978) notes, it is a great leap from being unable to work a few pencil and paper problems to being declared lacking in practical judgement, creativity, or general intelligence. It is in fact a leap to assume that the tests measure the same constructs for all groups.

When bias through overinterpretation is operant, it is not the test itself, its content, or its administration, which are considered biased; it is the interpretation and use made of the test results. Bias through overinterpretation cannot be

detected through scrutiny of the test itself, perhaps beyond its title. Nor can bias through overinterpretation be directly addressed through manipulation of the test content, format, or administration. For these reasons, test bias through overinterpretation, although important and pervasive, is not amenable to investigation in this study.

Bias as Sexism

Flaugher (1978) also discusses test bias as sexism, noting that in most respects the question of fairness to women can be treated identically with that of fairness for ethnic minorities, but also noting that our very language is pervaded with a distinct masculine bias. He particularly remarks on the generic use of male nouns and pronouns when content refers to both sexes. As Flaugher (1978, p. 674) states: "Tests may not be any more guilty of such bias than other users of the language, but they are seen as an appropriate medium through which a desirable social change may be effected, a change not confined to the function of the tests themselves." Bias due to sexism could function either through test item content or through testing atmosphere, both of which are discussed below. In some cases, test items contain language which is offensive to some group, but which apparently does not affect performance of the group on the item. Cole and Nitko suggest that the term "facial bias" be applied to items that contain words or content that appears to disfavor one group, but which do not empirically appear to disfavor one group (Cole & Moss, 1989). Tittle (1982) relates such content issues to the test takers' self-respect, and argues that test content should include balanced representation of the perspectives of less advantaged groups, including women.

Although sexism as a form of test bias is an interesting and important issue, it is not of central interest to the topic of this study, which is specifically focused

on racial bias. As Cole and Moss (1989, p. 207) state, the issue of facial bias is also "outside the traditional validity arena," and hence not addressed in this dissertation.

Bias as Item Content

Test bias is often seen as a consequence of the content of the items which compose the test. A biased test is often conceived as one made up largely of items which are in some sense unfair to some group (Bond, 1981; Flaugher, 1978). Often, inter-group differences in cultural experience, language usage, and vocabulary are responsible for such unfairness, when such vocabulary or experience do not ostensibly relate to the construct the test is intended to measure. As Bond (1981) points out, an item which requires an examinee to recognize the similarity of the relationship of a pig to a sty to that of a chicken to a coop may be a valid measure of verbal analogical reasoning for rural children, but it probably is not for urban children. If the test is used with a population composed of both urban and rural children, it is unlikely to measure the same construct in both groups, and could be considered biased against one group.

Two types of methods have been used to attempt to detect items with content that is unfair to some group. The first involves having test items scrutinized by panels of "experts" chosen for their supposed ability to recognize test items that are unfair to specific subgroups of examinees. Items judged as unfair to a subgroup, usually a minority group, are eliminated from the test. Although this approach imparts a degree of face validity to a test, it has not been shown to consistently improve the performance of minorities compared to their performance on the original "biased" version of the test. The degree of agreement among members of the panel of experts is often low, and the items identified as

biased are sometimes those upon which the minority group in question performs relatively well (Flaugher, 1978; Hackett et al., 1987; Plake, 1980; Schmeiser, & Ferguson, 1978). However, the approach is still in widespread use (cf. *Allen v. Alabama State Board of Education*, 1985; McAllister, 1987; Texas House Bill 1377; Tittle, 1982; Werner, 1988; Zoreff & Williams, 1980). Berk (1982) suggests that, rather than considering judgmental methods to be less adequate methods for detecting item bias, it would be more perceptive to regard them as identifying different kinds of bias than empirically-based methods. As Berk (1982, p. 4) states, judgmental methods play an important role, as "a study should begin with judgment and end with judgment." As discussed above, Cole and Nitko suggest that the term "facial bias" be applied to items that contain words or content that appears to disfavor one group, but which do not empirically appear to disfavor one group or show evidence of bias (such as differential functioning) (Cole & Moss, 1989).

The second method used to attempt to detect items with content unfair to some group has empirical as well as judgmental components. In this method, data on the ethnic (or gender) group membership of each examinee is obtained, together with test scores. Item performance statistics are then calculated for each group. Items which are particularly difficult for a minority group, compared to the group's performance on the other items of the test, are considered to be potentially biased. Such items are generally scrutinized for common format, content or vocabulary, in the hope of avoiding such items in future tests. Very often, little similarity is found among such items, and the etiology of their disparate difficulty remains therefore obscure (Flaugher, 1978). This approach

forms the basis for the definition of bias as differential item functioning, discussed below.

Bias as Test Atmosphere/Situational Bias

If examinees are to demonstrate their abilities or aptitudes on a test, they must feel sufficiently secure and comfortable that they can devote their attention to the test. Persons who feel out of place, unwelcome, or foredoomed to failure will probably not perform as well as persons of equal ability who feel secure, comfortable, and optimistic. Flaugher (1978) refers to this as "atmosphere bias," and states that gender and ethnic differences can often create such situations. Bias resulting from discomfiting factors in the external testing situation, rather than the test itself, are termed "situational bias" by Berk (1982), Jensen (1980), and others.

Even if the testing atmosphere is not overtly hostile, some persons may be inhibited by the very fact of being confronted with a test, and it seems likely that minorities may be disproportionately represented among such persons. Unfortunately, empirical scrutiny for this type of bias if difficult, as it requires observation of the situation of test administration and/or study of the nature of the examinees. Flaugher (1978) notes, however, that the issue of atmosphere bias goes beyond the individual examinee and the individual test administration. Testing is pervasive in American public education, and for some subpopulations the very process of testing itself may create discouragement and despair. In particular, the application of nationally normed tests to groups who enter the school system with enormous educative handicaps due to poverty and malnutrition, such as children in some inner-city schools, and the subsequent

detailed documentation of their consequent poor performance, may provide nothing but ill effects for these students and their teachers.

When test bias is due to atmosphere, it is not the test itself, its content, or procedures for administration, which are considered biased. Bias due to atmosphere cannot be detected through scrutiny of the test itself, nor can it be addressed through manipulation of the test content. For these reasons, bias due to atmosphere is not amenable to investigation in this study.

Bias as the Selection Model

Standardized tests are often used as a basis for decisions regarding employment, admission to schools or programs, licensure, or other selection decisions. Bias in the selection model, also called *selection bias*, has received considerable attention in the last two decades (cf. Flaugher, 1978; McNemar, 1975; Peterson & Novick, 1976; Thorndike, 1971a). In selection bias, the fact that the predictor variable in use is a standardized test is incidental. In fact, considerations of selection bias would be equally applicable were selection based upon performance on other instruments, such as a psychomotor test, a verbally administered test, a structured interview, past academic performance, past job experience, or the results of structured observation. The bias allegedly resides in the rules concerning how decisions based upon the predicted performance are made: i.e., the selection model. Traditionally, common decision points or cut scores have been used for all ethnic and gender groups. This has been considered fair, in that all persons achieving the same score were treated alike. However, a variety of alternative models have been suggested to be fair on the basis of other values, such as the assurance that equal proportions of applicants likely to succeed are admitted from all groups (cf. Jaeger, 1976). These various

models have been discussed extensively by Peterson and Novick (1976). Flaugher (1978, p. 676) states that the various selection models used or advocated are in fact explications of various value systems, and that the search for a truly fair selection model is "a search for a set of values upon which everyone can agree." As discussed at length earlier in this chapter, values regarding preferential treatment and the meaning of equality of opportunity are widely divergent in this country, thus universal consensus on a common selection model is unlikely.

Since selection bias relates to the use to which a test is put, rather than the test itself, such bias cannot be detected through scrutiny of the test itself. For this reason, it is not amenable to investigation in this study.

Bias as the Wrong Criterion

Flaugher (1978) and Gulliksen (1976) discuss the relationship of the "criterion problem" to test bias. As Flaugher (1978, p. 676) states, criterion problems exemplify the complex nature of bias, cut across many of the other aspects and definitions of bias, and impose themselves on test bias in a variety of ways.

Bias in society is not limited to standardized tests. Many of the criteria that tests are used to predict are susceptible to contamination with bias. For example, supervisors, who are typically majority-group members, may tend to rate minority employees more harshly than equivalently performing majority employees. Minority students may be graded more harshly than majority students, or confront educative hardships not faced by majority-group students. If the criterion is contaminated by bias, maximizing the test-criterion correlation merely entrenches the preexisting bias.

Another way in which a problem with the criterion can lead to bias is through differences in reliabilities between the predictor and the criterion it is used to predict (Flaugher, 1978).. Tests are often used to predict such non-standardized-test criteria as subsequent grades, grade point averages, or supervisory ratings of job performance. These criteria are often less reliable than the predictor tests, leading to greater differences between the means of majority and minority groups on the predictor test than on the criterion (Flaugher, 1978; Goldman & Widawski, 1976; Thorndike, 1971). This can lead to inaccurate prediction and to bias if no attempt is made to determine whether the mean differences on the criterion would in fact be equivalent under conditions of equivalent reliability (Flaugher, 1978).

Gulliksen (1976) and Flaugher (1978) stress that there is no simple solution to criterion bias, but that careful scrutiny of criteria must be a systematic and ongoing process.

Bias as Differences in Test-Criterion Correlations

Flaugher (1978) also discusses test bias as single-group or differential validity. He defines "single-group validity" as the finding of a predictive validity coefficient significantly different from zero for one group, but not for another. He uses the term "differential validity" to refer to the finding of a significant difference between a test's predictive validity coefficients for two groups. This finding is also referred to as "differential correlation" in the literature, and that term is used for the phenomenon henceforth in this study; the term "differential validity" is given a different, broader meaning below (as per Cole & Moss, 1989). Although differences in test-criterion correlations between groups appears to provide compelling evidence of bias, especially in cases where the test is

intended to predict or substitute for the criterion, Flaugher (1978), L. Bond (personal communication, November 1, 1989), and others have noted that such differences are rarely found. The large proportion of unexplained variance in any criterion score may in part account for the elusiveness of this phenomenon, but Flaugher (1978, p. 674) concludes that "they [differences in test-criterion correlation across groups] are not very potent phenomena relative to all the other possible sources of problems in the interaction of minorities and testing."

Bias as Differential Prediction

Since publication of Flaugher's inventory of conceptualizations of bias in 1978, other definitions have emerged in the professional literature. Flaugher cited differential correlation (differences between the majority and minority in test-criterion correlations) as a type of bias. Differential prediction, or differences between the regression equations predicting the criterion from the test for the majority and minority groups, is also considered evidence of bias (Cole & Moss, 1989; Goldman & Hewett, 1975; Goldman & Richards, 1974; Goldman & Widawski, 1976; Hartigan & Wigdor, 1989; Reynolds, 1982a). Differential correlation addresses whether the test predicts equally well for the majority and the minority; differential prediction addresses whether it predicts in the same way for the majority and minority. Like differential correlation, differential prediction appears to provide a straightforward and appealing definition of bias. However, like differential correlation, it is of limited utility since such differences in regression equations are in fact rarely found (L. Bond, personal communication, November 1,1989; Cleary, 1968; Cole & Moss, 1989; Goldman & Hewett, 1975; Goldman & Widawski, 1976). In other cases, where differences in regression equations between groups are found, the equations derived from the

majority group are found to overpredict, rather than underpredict, for the minority group (cf., Goldman & Richards, 1974),

Bias as Differential Item Performance for Groups of Equal Ability

The advent of latent trait theory, item response theory (IRT), and related methodologies has led to another definition of bias in a test item and in a test. In IRT, an item is considered to be biased if two groups of equal ability exhibit different performance, since the differences in performance must be attributed to factors other than ability. This definition seems both intuitively appealing and unachievable, since demonstrating that two groups are of equal ability seems inherently impossible. IRT methodology, however, purports to distinguish between ability and performance, and creates "item characteristic curves" for items that are independent of the distribution of ability levels of the subjects used to construct them (Angoff, 1989; Cole & Moss, 1989; Hambleton, 1989; Ironson, 1982; Ironson, 1983; Lord, 1980). As Marasculio and Slaughter (1981, p. 229) explain:

> In IRT, an item is said to be unbiased if the characteristic curves for the item measured on two groups are identical. If they are identical, they have the same intercept (or probability of a correct response when ability is very low), the same slope (or discrimination power), and the same inflection point (or item difficulty).

IRT methodology, and its attendant definition of bias, has gained great acceptance in the measurement community, and is widely used by large testing organizations and in large-scale studies (cf. Flaugher & Schrader, 1978; Hills, 1989; Holland & Thayer, 1985; Kok, Mellenbergh, & Van Der Flier, 1985; Peterson & Flesher, 1982; Rogers, Dorans & Schmitt, 1986; Zwick, 1990).

However, the procedures for estimating ITR parameters and testing the identity of item characteristic curves are computationally complex, and not readily understood by lay persons. They also require extremely large sample sizes for practical differences to exhibit statistical significance (Marasculio & Slaughter, 1981). For these reasons, a variety of simplified methods which offer approximations to IRT methods and share the same underlying definition of bias have emerged (cf. Linn & Harnisch, 1981; Ironson, 1982; Scheuneman, 1979; Zwick, 1990). Most of these methods have received considerable criticism, largely on purely statistical grounds (cf. Baker, 1981; Marasculio & Slaughter, 1981).

Other methods for detecting differential item functioning have focused on examination of item difficulties in an attempt to identify items that were relatively more difficult for one group than for another. Item difficulty values were computed for each group, and each difficulty level was converted to a standard scale, called delta. The item deltas were plotted against each other, and outliers were considered suspect (Angoff, 1982; 1989; Cole & Moss, 1989). However, it has been argued that these methods both tend to yield false positives, and fail to detect legitimately biased items (Angoff, 1989; Cole & Moss, 1989).

Bias as Differences in Factor Structure

If the factor structure of a test for one group is different from that for another group, the test might measure different constructs in the two groups. The factor analysis approach has been used as a method for investigating bias in a test. However, most empirical studies have yielded similar factor structures across groups (Cole & Moss, 1989; Reynolds, 1982a). Evaluating and judging the degree of similarity of factor structures for two groups has also proved problematic. Cole and Moss (1989) state that factor analytic approaches to the

investigation of bias are not widely used, both because of the complexity of the theory and the methodological difficulties in performing such analyses.

A Unified Definition of Bias

Although Flaugher (1978) does not make the observation himself, all of the definitions of bias he discusses, except for the definition of bias as differences in distributions of scores, reflect various ways in which the results of testing, or interpretations based upon these results, are less valid for one group than for another. Subsequently developed concepts of bias are also derived from various models of impairment of validity. This is consistent with the widely held value that all members of society have the right to equality of educational opportunity and opportunity to compete: if measurement instruments are more accurate or valid for some groups than for others, decisions and rules of competition based upon these instruments would not support equal competition.

Measurement specialists once viewed validity as being composed of several distinct types, (i.e., content validity, criterion related validity, predictive validity, etc.), each of which was considered applicable to different situations (cf. APA, et al., 1974; Gay, 1976). As Berk (1982, p. 2) notes, previous efforts to provide a framework for sorting out the various definitions of test bias and their attendant methodologies have typically been based on the traditional validity triumvirate: content validity, criterion-related validity, and construct validity (cf. Jensen, 1980; Reynolds, 1982b). However, these frameworks have generally proved inadequate, since particular bias issues often do not relate distinctly to one of the three categories (Berk, 1982).

It is now recognized that construct validity subsumes all other forms of validity (Cronbach, 1989: Messick, 1989). Other forms of validity evidence, such as content representativeness, support claims to construct validity, but provide only partial evidence. As Messick (1989) states, the only source of evidence not incorporated in construct validity is appraisal of the social consequences of test use. Appraisal of social consequences includes adverse impact, a critical social policy issue (Cronbach, 1976; Messick, 1989).

Cole and Moss (1989, p. 205) provide a definition of bias which derives from the unified definition of validity: "Bias is differential validity of a given interpretation of a test score for any definable, relevant subgroup of test takers." Since all types of validity evidence are now unified under construct validity, and bias is defined broadly as differential validity, bias can also be viewed as a unified construct. Rather than view different definitions of bias, and the attendant methods for detecting bias, as competing, they can be viewed as complementary approaches for investigating the unitary construct validity of the inferences derived from test results for all groups of interest. The question "Is this test (or test item) biased?" can not receive a simple yes-or-no answer. Rather, a series of connected concerns relating to the fairness of the test (or item) to all examinees must be examined when answering the question (Cole and Moss, 1989; Hackett, et al., 1987). Because bias is defined as differential validity, validation theory becomes the conceptual basis for investigations concerning test or item bias (Cole & Moss, 1989, p. 205).

Applying Cole and Moss's (1989) unified definition of bias to the inventory of the ways in which the term "bias" has previously been used, and to Messick's (1989) discussion of sources of evidence of validity, allows one to identify two

distinct questions regarding the effect of using standardized tests on minorities: 1) Is there any reason to believe that the results produced by this test, or the interpretation of these results, are inaccurate or scientifically suspect for some group? and 2) Will use of this test result in undesirable or unacceptable social or economic consequences for some group? The first of these questions addresses whether the test is biased, and is a scientific issue amenable to investigation and possible correction through the scientific method. The second question, which is often synonymous with differences in distributions of scores between groups, addresses whether use of the test results in unacceptable adverse impact on some group. This is a social policy issue, and not amenable to correction through the scientific method (Cronbach, 1976). In fact, pseudo-scientific attempts at reversing adverse impact may obscure and confuse important bias issues.

Recognizing that all investigation of bias relates to the search for evidence of differential construct validity, Cole and Moss (1989) suggest that evidence pertinent to bias be grouped into five categories: 1) internal test structure; 2) external test relationships; 3) content and format; 4) test administration and scoring; and 5) constructs in context. These categories are complementary, and are not mutually exclusive.

Evidence in the category of internal test structure concerns the interrelationships among a test's parts. Since reliability is a necessary, but not sufficient, precondition for validity, evidence of internal consistency reliability falls into this category. Evidence related to item difficulties and item discrimination values, which have long been examined as a part of test construction and test analysis, also fall into this category. Evidence related to item interrelationships and differential item functioning, including approaches

focused on comparative item difficulties (cf. Angoff, 1982; Holland & Thayer, 1985) and latent trait methods (cf. Dorans, 1984; Rogers, Dorans, & Schmitt, 1986) also fall within this category. Comparisons of the factor structure of a test for various groups also serve as evidence in this category, according to Cole and Moss, (1989).

The category of external test relationships includes the relationships of test scores to a number of variables external to the tests. The unifying concern is that these relationships might be different for various groups (Cole & Moss, 1989). Evidence pertinent to both convergent and discriminant validity issues, as well as issues related to predictive and concurrent validity, is in this category. Flaugher's (1978) definition of "bias as differential correlation" is within this category, as is the more recently developed definition of bias as differential prediction.

The category of content and format includes evidence which addresses whether the content and format of the test are appropriate for the measurement of the intended construct. Judgmental and empirical evidence as to whether the test's wording, content, or format of presentation invalidly disadvantages any group fall within this category. Flaugher's (1978) definition of "bias as content" is within this category.

The category of test administration and scoring includes evidence related to whether the way in which the test is administered elicits maximum performance from all groups. The category also includes evidence related to the existence of group-related influences irrelevant to the construct the test is intended to measure on the ratings of responses. Flaugher's (1978) definition of

"bias as atmosphere" is within this category, as well as aspects of his definition of "bias as sexism."

The category of constructs in context concerns whether the construct explicated in the testing purpose includes skills or knowledge that are irrelevant to the intended interpretation of the test. This is a broad category, including consideration of such seemingly disparate factors as the effect of coaching on test results, as well as concerns over the selection of the groups for which bias is to be investigated. Flaugher's (1978) definition of "bias as overinterpretation" falls into this category, as do his definitions of "bias as the selection model" and "bias as the wrong criterion." As Cole and Moss (1989) state, careful consideration should be given to the constructs a test is intended to measure, and factors affecting this, when a test is initially developed or selected.

These five categories proposed by Cole and Moss are not seen as different types of bias, but as different areas in which evidence about bias can be examined, as part of the unitary process of construct validation. For the empirical component of this study, evidence within each of these categories will be examined to the extent that it is feasible and appropriate, given the nature of the tests and the data at hand.

Empirical Research Questions to be Addressed in this Study

Any consideration given to issues of adverse impact, preferential treatment, and bias in test use necessarily proceeds from a position that man has some natural rights that are antecedent to the laws and rules under which he lives. However, as noted earlier, a variety of positions can be taken as to what types of equality men should be guaranteed. At times, the Golden Rule strategy has been described as a mechanism to reduce test bias (cf. Weiss, 1987), while at

other times, it has been described as a mechanism to reduce adverse or disparate impact (cf. Haney & Reidy, 1987). Many members of the public see the two issues as synonymous, as noted by Flaugher (1978). As discussed earlier, members of the measurement profession do not equate the issues of bias and disparate impact. For the purposes of this study, bias will be defined as the scientific issue of differential construct validity (described earlier), and adverse or disparate impact will be defined as differences between the distributions of scores of black and white examinees.

The stipulations of the Golden Rule strategy are clearly intended to reduce disparate impact, since they explicitly require avoidance of the use of test items on which black and white examinees perform differently. The Golden Rule procedures were not confined to this stipulation, however, and some believe that use of the procedures will reduce test bias. However, measurement professionals have expressed some doubt as to whether application of the Golden Rule procedure will accomplish either the reduction of disparate impact or the reduction of bias. In addition, questions exist as to whether it is appropriate or desirable to manipulate test characteristics, as the Golden Rule procedures do, in order to reduce disparities between the score distributions of black and white examinees. (Among those who accept that all members of society deserve equal opportunity to compete, there does not seem to be a question about the appropriateness or desirability of striving to make tests equally valid for all examinees.)

Two broad question concerning the effect of standardized test use on minorities were distinguished above. These were: 1) Is there any reason to believe that the results produced by this test, or the interpretation of these

results, are inaccurate or scientifically suspect for some group? and 2) Will use of this test result in undesirable or unacceptable social or economic consequences for some group? Empirically, this study investigates two major research questions which parallel these broad questions. The first major research question examined in this study is: Is application of the Golden Rule procedures effective in reducing the disparate impact of test use? The second major research question is: Is application of the Golden Rule procedures effective in reducing test bias?

Both empirical research questions were investigated using data on the performances of students on the Georgia Eighth-Grade Criterion Referenced Tests in Mathematics and Reading in the Spring of 1986, and subsequently, on the Georgia Basic Skills Tests in the Fall of 1987. The Georgia Basic Skills Test, administered at the beginning of the tenth grade, also has Mathematics and Reading components.

Investigation of the Effect of the Golden Rule Procedures on Disparate Impact

To investigate the first research question, the effect of applying the Golden Rule procedures on disparate impact, the average performances of black and white examinees on each item of the Eighth-Grade Reading and Mathematics Tests and the Tenth-Grade Basic Skills Tests in Mathematics and Reading was determined. "Subtests" of each of the four tests, composed solely of items that conform to the Golden Rule stipulations, without protection of content representativeness, were then constructed. Differences between distributions of scores of black and white examinees on the original four tests were then compared to corresponding differences between distributions of scores for black and white examinees on the four synthetic tests which conform to the Golden Rule specifications. Details of the methodological procedures for this

investigation are provided in the next chapter of this dissertation. The results of these investigations are discussed in light of the philosophical and policy contexts of adverse impact and preferential treatment in the final chapter of this dissertation.

Investigation of the Effect of the Golden Rule Procedures on Bias

Investigation of the second major research question, the effect of applying the Golden Rule procedures on test bias, also involved use of "synthetic" tests that are constructed to conform to the Golden Rule stipulations, without specific protection of content representativeness. As Flaugher (1978, p. 671) and others (Hackett, et al., 1987) have stated, the question of whether a test is biased is not a "yes/no" question, with a single unequivocal answer. Rather, a series of connected concerns relating to the fairness of the test to all examinees must be examined. To address the research question dealing with the effects of the Golden Rule procedures on bias, various methods of detecting test bias described in the earlier review of professional literature have been organized according to the five categories proposed by Cole and Moss (1989). Given available resources and the nature of the data at hand, sources of evidence appropriate to each category proposed by Cole and Moss (1989) were examined to determine the effects of applying the Golden Rule specifications.

Under the category of internal test structure, as proposed by Cole and Moss (1989), several subsidiary research questions were examined. These were:

1. What is the effect of applying the Golden Rule procedures on the internal-consistency reliability of tests?
2. Does application of the Golden Rule procedures increase the similarity of the reliability of tests for black and white subgroups?

3. What is the effect of applying the Golden Rule procedures (without specific protection of content representativeness) on the content representativeness of tests?
4. What is the effect of applying the Golden Rule procedures on the average item difficulty (as defined in classical true-score theory) of tests?
5. What is the effect of applying the Golden Rule procedures on the average item by total-score correlation of tests?

Each of these subsidiary research questions was investigated using each of the four tests available, the Eighth-Grade Reading and Mathematics Tests and the Tenth-Grade Basic Skills Tests in Mathematics and Reading, allowing investigation of the generalizability of findings across two subject areas and two grade levels.

Examination of the effect, if any, of applying of the Golden Rule procedures on differences between the factor structures of the tests for black and white examinees is beyond the scope of this study, although such an investigation would be pertinent to Cole and Moss's (1989) category of internal test structure. Similarly, investigation of whether items that fail to conform to the Golden Rule specifications also exhibit differential item functioning would also be germane to this category of evidence, but such investigation is beyond the scope of this study.

Under the category of external test relationships, three subsidiary research questions were investigated. These were:

1. What is the effect of applying the Golden Rule procedures on the overall predictive validity of tests?
2. Does application of the Golden Rule procedures increase the similarity of test-criterion correlations for black and white examinees?
3. Does application of the Golden Rule procedures increase the similarity of regression equations for black and white examinees (i.e., does it reduce differential prediction)?

These three research questions were investigated by using the unaltered Grade 10 Basic Skills Tests as criteria. Application of the Golden Rule procedures

to the Eighth-Grade Reading and Mathematics Tests will be examined, as it effects the relationship of these tests to the subsequently administered Grade 10 Basic Skills Tests. Data were not available to investigate divergent validity, which is also a source of evidence under this category of bias inquiry.

Only limited evidence in the category of test content and format was examined in this study. The items that were excluded from the tests by applying the Golden Rule procedures were reviewed by the researcher for content similarities, but it was beyond the scope of this study to compose panels of experts to review these items for discriminatory content. Similarly, it was beyond the scope of this study to examine whether items which violate the specifications of the Golden Rule procedures also exhibit differential functioning, as defined by IRT and related methods. The format of the tests is not affected by applying the Golden Rule procedures, thus this area was not examined. The effects of applying the Golden Rule stipulations on the content representativeness of the tests was investigated by comparing the proportions of items addressing each test objective in the original tests with the proportions addressing each objective in the tests composed of items which conform to the Golden Rule specifications.

The administration and scoring of the tests is not affected by applying the Golden Rule procedures, so no evidence in this area was examined in this study. Aspects of situational bias also fall into this category, but would not be affected by the Golden Rule procedures, and hence were not investigated in this study.

As stated earlier, the category of constructs in context is broad, and includes several types of apparently dissimilar evidence. Flaugher's (1978) definition of bias as the wrong criterion falls into the category of constructs in context. Concerns that bias pervades both a test and the criterion it is to predict

are widespread. In this study, it was possible to apply the Golden Rule procedures to both a test and its criterion, and examine the resulting effect. If the Golden Rule procedures reduce bias, one would anticipate that both test-criterion correlations and associated regression equations would be more similar for blacks and whites when the Golden Rule procedures were applied to both the test and the criterion, than when they were applied to neither. If disparate impact were reduced and predictive validity were increased, selection bias would be reduced. These relationships were examined empirically in this study.

Thus, one subsidiary research question was addressed within the category of constructs in context:

1. Does application of the Golden Rule procedure to both a test and the criterion it is intended to predict reduce the dissimilarity of correlations between the test and the criterion, and reduce the dissimilarity regression equations, for black and white examinees?

For this question, the Grade 10 Basic Skills Tests was again used as criteria, as they were in the investigation of predictive validity. In this case, however, the Basic Skills Tests was also modified to conform to the Golden Rule specifications. The effects was investigated by using the Eighth-Grade Reading and Mathematics tests as predictors.

The details of the methodological procedures used to investigate these research questions are provided in the next chapter of this dissertation. The results of investigations of the major research questions and each of the subsidiary research questions are discussed, and conclusions are drawn in light of the philosophical, policy, and scientific contexts of adverse impact and preferential treatment, in the final chapter of this dissertation.

CHAPTER III

METHODOLOGY

In the previous chapter, a set of hierarchical research questions to be investigated in this study were derived from the professional literature. This chapter describes the procedures followed in completing empirical data analyses used to address these research questions. The research questions of this study were investigated using data on students' performances on the Georgia Eighth-Grade Criterion Referenced Tests and their subsequent performance on the Georgia Basic Skills Tests (BSTs). This chapter begins with a description of the sample of Georgia students used in the study, followed by a discussion of the instruments used to collect data from the sample: the Georgia Eighth-Grade Criterion Referenced Tests and the Georgia Basic Skills Tests.

Data for this study were provided to this researcher on a computer readable tape, thus this researcher did not participate in data collection. However, a brief description is given of the conditions under which data were purportedly collected. Data reduction procedures performed by this researcher are also described.

This study required the creation of synthetic versions of the Georgia standardized tests which were composed only of those items from the original tests which conform to the Golden Rule stipulations for Type I items. These synthetic versions of the tests were created under two sets of assumptions: in one, examinees' correct answer rates were regarded as sample statistics, in the other they were regarded as population parameters. The procedures by which

the synthetic tests were constructed under each set of assumptions are described later in this chapter.

Investigation of the research questions of this study involved a series of examinations of the characteristics of, and relationships between and among, examinees' performances on the synthetic tests which conform to the Golden Rule procedures, and on the original standardized tests. Procedures for investigation of each research question are described in a latter section of this chapter.

Sample

The sample for this study consisted of students in Georgia's public schools who completed the Georgia Eighth-Grade Criterion Referenced Tests in Mathematics and Reading in the spring of 1986, and who subsequently completed the Georgia Basic Skills Tests, administered at the beginning of the tenth grade, in the fall of 1987.

All eighth-grade students in public schools in Georgia were administered the Eighth-Grade Criterion Referenced Tests in the Spring of 1986. Twenty different test forms were used in the Spring, 1986, administration of the Eighth-Grade Criterion Referenced Tests. Test booklets for these 20 forms were spiraled, and all students taking a single form were candidates for inclusion in the sample used in this study. This selection method approximated linear systematic selection of a five-percent sample of Georgia eighth-grade students, and resulted in roughly proportional representation of students from each school and school district in the state.

The final sample used in this study consisted of a subset of the Georgia students for whom scores were available on both the designated form of the

Eighth-Grade Criterion Referenced Tests from the spring 1986 administration, and on any form of the Basic Skills Tests from the fall 1987 administration. Data from the two years were matched on the basis of name, gender, race, and school system. Students who were in the eighth-grade in Georgia in the spring of 1986, but who had not matriculated to the tenth grade by the fall of 1987, or who changed school systems or left the state before the fall of 1987, were excluded from the final sample, as were students for whom data from the two years could not be matched for other reasons, such as change of name or the listing of a different ethnic group membership or gender in the two years (S. Gabrielson, personal communication, February 13, 1990; D. Davis, personal communication, February 14, 1990).

The original sample providing data for this study consisted of 1812 students, but information on ethnic group membership was not provided for five of these subjects: three females and two males. These five subjects were omitted from the study, giving a final sample size of 1807 students. These 1807 students were enrolled in all 194 Georgia public school systems in existence at that time. Table 1, below, presents the composition of the sample by race and gender.

Representativeness of the Sample

The composition of this sample is loosely representative of the demographic composition of the student population of public schools in Georgia. According to the *Digest of Educational Statistics* (National Center for Educational Statistics, 1989) the total student population enrolled in public schools in Georgia in the fall of 1986 was approximately 61% white, 38% black,

Table 1

Composition of Sample by Race and Gender

Race	Females	Males	Total	%
White	689	580	1269	70.23
Black	285	232	517	28.61
American Indian	1	0	1	0.06
Asian	3	5	8	0.44
Hispanic	3	2	5	0.28
Other	4	3	7	0.39
Total	985	822	1807	

<1% Hispanic, <1% Asian/Pacific Islander, and <1% American Indian. The exact ethnic composition of eighth-graders in Georgia's public schools in 1986 is not known, however, it is possible that white students are disproportionately overrepresented, and blacks students underrepresented in the sample, since they compose 70% and 29%, respectively, of the sample, compared to approximately 61% and 38% of the population for which racial composition data are available.

As discussed above, the final sample used in this study included 1807 public school students for whom both scores on the Georgia Eighth-Grade Criterion Referenced Tests from the spring of 1986 and scores on the Georgia Basic Skills Tests from the fall of 1987 were available. According to the *Digest of Educational Statistics* (National Center for Educational Statistics, 1988) there were 83,280 students enrolled in the eighth-grade in public schools in Georgia in 1986. This sample represents approximately 2.2% of those students. Approximately five percent of all eighth-graders, or 4,164 students,

were originally candidates for inclusion in this study. The actual sample was thus approximately 43% of the potential sample. It is clear that there were substantial numbers of students for whom it was not possible to match scores for the two years under consideration.

The subsamples of both black students and white students are large enough to estimate proportions for corresponding populations within plus or minus 5% with 95% confidence, as determined by the formula (Cochran, 1977):

$$n = \frac{(t/e)^2 [P (1-P)]}{1 + (1/N) [(t/e)^2 P (1-P) -1]}$$

where n = sample size
t = standard normal deviate
e = allowable estimation error
P = population proportion
N = population size

coupled with assumed population proportions of .50 (which gives maximum variance, resulting in conservative estimates.) However, the reader must bear in mind that the sample does not represent a random sample of the population of eighth-graders or tenth-graders. The sample was selected in such a way as to exclude students who moved to a different school district, were retained in grade, dropped out of school, changed their names, or recorded different ethnic group memberships, names, or genders on the two testing dates. It is likely that these students have lower academic achievement, on average, than the students in the final sample. It is also possible that black students were disproportionately excluded from the

sample by these factors. Had a simple random sample of students from either year been used, it is possible that even more items would have failed to meet the criteria for Golden Rule Type I items (as discussed subsequently in this chapter). The size of this effect is not estimable.

Data Collection Instruments: The Georgia Criterion Referenced Tests

The empirical research questions of this study were investigated using data on students' performances on the Georgia Eighth-Grade Criterion Referenced Tests in Mathematics and Reading in the spring of 1986, and subsequently, on the Reading and Mathematics tests of the Georgia Basic Skills Tests in the fall of 1987. The Georgia Basic Skills Tests are criterion-referenced tests, and are first administered to students at the beginning of the tenth grade. All four tests are produced for the Georgia Department of Education by the Georgia Assessment Project of Georgia State University.

The Eighth-Grade Criterion Referenced Tests in Mathematics and Reading and the Georgia Basic Skills Tests are administered as part of the Georgia statewide testing program. The primary aim of the statewide testing program is "providing information to teachers, students, parents, concerned citizens, and educational policy and decision makers " (Georgia Department of Education, 1988b, p. I-1). This information is collected "to aid teachers and administrators in instructional planning, to aid students and their parents in personal decision-making, and to aid educators and the general public in evaluating the effectiveness of educational programs" (Georgia Department of Education, 1988b, p. I-1). According to the Georgia Student Assessment Handbook, the criterion-referenced tests are "primarily used to customize each student's learning program with his or her specific needs" (Georgia Department of Education, 1988b, p. I-1). The Eighth-Grade Criterion

Referenced Tests in Mathematics and Reading are also specifically intended to "identify students who may need additional learning experiences in the basic skills before taking the High School Basic Skills Tests ...in grade 10" (Georgia Department of Education, 1987a, p.1) The Basic Skills Tests, which must be passed in order to graduate from high school, are specifically intended to protect the integrity of the high school diploma, and are used so that "educators, parents and students can be assured that a student who attains a Georgia high school diploma possesses at least minimal levels of many important basic tools of lifelong learning" (Georgia Department of Education, 1982b. p. 1).

The Georgia Eighth-Grade Criterion-Referenced Tests

Edition 3 of the Georgia Eighth-Grade Criterion-Referenced Tests was used in this study. This edition was copyrighted in 1983 by the Georgia Department of Education, and includes separately numbered tests in mathematics and reading. Reading and mathematics results from the Georgia Criterion-Referenced Tests are always treated and reported separately; a total score is never calculated.

The Reading Test of Edition 3 of the Georgia Eighth-Grade Criterion-Referenced Tests is composed of 122 items which reflect three skill areas and 11 objectives. Reading Objectives 1 through 3 address the skill area of Literal Comprehension; Reading Objectives 4 through 7 address the skill area of Inferential Comprehension; and Reading Objectives 8 through 11 address the skill area of Problem Solving (Georgia Department of Education, 1987b). These Reading objectives are listed in Appendix B.

The Mathematics Test of Edition 3 of the Georgia Eighth-Grade Criterion-Referenced Tests is composed of 127 items which reflect three skill

areas and 12 objectives. Mathematics Objectives 1 through 4 address the skill area of Concept Identification; Mathematics Objectives 5 through 7 address the skill area of Component Operations; and Mathematics Objectives 8 through 12 address the skill area of Problem Solving (Georgia Department of Education, 1987b). These Mathematics objectives are listed in Appendix C.

Each item of Edition 3 of the Georgia Eighth-Grade Criterion-Referenced Tests corresponds to one objective and one skill area. (Information is not available on the objectives assessed by Items 52 & 65 of the Reading Test, or Items 19, 22, 72, and 75 of the Mathematics Test.) Table 2, below, shows the item numbers corresponding to each reading objective and skill area. Table 3, below, shows the item numbers corresponding to each mathematics objective and skill area. (The actual item numbers corresponding to each objective, as well as the numbers of items relating to each objective, are listed, since they will be important to subsequent sections of this study which address the effects of applying of the Golden Rule procedures on the content representativeness of tests.)

Table 2

Correspondence of Items of Georgia Eighth-Grade Criterion-Referenced Test (Edition 3) in Reading to Test Skill Areas and Objectives*

Skill Area/Objective	Corresponding Items	Total # of Items
Literal Comprehension		
Objective 1	35, 36, 47, 57, 68, 90, 115	7
Objective 2	1, 5, 7, 11, 32, 41, 42, 43, 44, 45, 48, 84, 86, 113, 116, 122	16
Objective 3	30, 31, 33, 34, 53 ,56, 63, 93, 94, 118, 119, 120	12
Inferential Comprehension		
Objective 4	6, 10, 14, 22, 23, 49, 54, 58, 67, 71, 72, 73, 83, 87, 89, 98, 99, 100, 110, 111, 112, 117	22
Objective 5	15, 46, 51, 55, 62, 85, 91, 121	8
Objective 6	19, 25, 28,37, 61, 96, 105	7
Objective 7	59, 60, 69, 77, 88, 92	6
Problem Solving		
Objective 8	12, 13, 18, 38, 39, 40, 70, 74, 75, 76, 82, 108	12
Objective 9	4, 8, 16, 24, 29, 64, 66, 95, 97, 107, 114	11
Objective 10	2, 9, 20, 21, 26, 79, 81, 101, 102, 103, 104, 106	12
Objective 11	3, 17, 27, 50, 78, 80, 109	7

* Information is not available on which objectives Items 52 & 65 are intended to address.

Table 3

Correspondence of Items of Georgia Eighth-Grade Criterion-Referenced Test (Edition 3) in Mathematics to Test Skill Areas and Objectives*

Skill Area/Objective	Corresponding Items	Total # of Items
Concept Identification		
Objective 1	4, 15, 23, 46, 58, 73, 77, 78, 93, 98, 117	11
Objective 2	7, 29, 33, 39, 51, 53, 63, 70, 71, 76, 83, 99, 108, 118, 126	15
Objective 3	1, 26, 31, 43, 50, 67, 87, 95, 107	9
Objective 4	5, 10, 14, 28, 41, 47, 49, 57, 60, 74, 97, 100, 109, 115	14
Component Operations		
Objective 5	35, 66, 88, 110	4
Objective 6	6, 11, 17, 27, 37, 42, 45, 54, 65, 84, 89, 91, 106, 111, 114, 122	16
Objective 7	2, 12, 16, 18, 24, 30, 34, 36, 48, 55, 61, 62, 69, 81, 86, 94, 101, 103, 104, 112, 124, 125	22
Problem Solving		
Objective 8	25, 32, 52, 68, 85, 90, 92, 123	8
Objective 9	6, 8, 21, 64, 80, 113, 116, 121	8
Objective 10	13, 44, 79, 96, 127	5
Objective 11	20, 38, 56, 59, 102, 105, 120	7
Objective 12	3, 40, 82, 119	4

* Information is not available on which objectives Items 19, 22, 72, & 75 are intended to address.

The Georgia Basic Skills Tests

The Georgia Basic Skills Tests consist of criterion-referenced tests in mathematics and reading, with passing scores established for each test. There is also a Georgia Basic Skills Test in writing which was not included in this study. Students in Georgia must pass each of the Basic Skills Tests in order to graduate from high school with a regular diploma. Students are first

administered the tests in the fall of the tenth grade, but may repeat any test which they do not pass on first attempt. Scores on various editions of the Basic Skills Tests are converted to a common scale score, and passing scores are set in terms of this scale score. Different minimum passing scores are used for students who entered the ninth grade in different years, for students repeating tests, and for transfer students (Georgia Department of Education, 1982a; 1988a). The formulae for conversion of raw scores to scaled scores for the form of the tests used in this study were not available to this researcher, thus only raw scores were used in this study.

Form 01, Edition 13 of the Georgia Basic Skills Tests was used in this study. This edition was copyrighted in 1987 by the Georgia Department of Education, and includes separately numbered tests in mathematics and reading. Reading and mathematics results from the Georgia Basic Skills Tests are always treated and reported separately; a total score is never calculated.

The Mathematics Test of Form 01, Edition 13 of the Basic Skills Tests is composed of 112 items which reflect three skill areas and 14 mathematics objectives . Mathematics Objectives 1 through 5 address the skill area of Concept Identification; Mathematics Objectives 6 through 9 address the skill area of Component Operations; and Mathematics Objectives 10 through 14 address the skill area of Problem Solving (Georgia Department of Education, 1988a). These mathematics objectives are listed in Appendix D.

The Reading Test of Form 01, Edition 13 of the Georgia Basic Skills Tests is composed of 105 items which reflect three skill areas and 12 reading objectives . Reading Objectives 1 through 3 address the skill area of Literal Comprehension; Reading Objectives 4 through 7 address the skill area of

Inferential Comprehension; and Reading Objectives 8 through 12 address the skill area of Problem Solving (Basic Skills Tests Fall 1987- Spring 1988 Interpretive Guide). These reading objectives are listed in Appendix E.

Each item of Form 01, Edition 13 of the Georgia Basic Skills Tests corresponds to one objective and one skill area. (Information is not available on which objective Item 14 of the Mathematics Test is intended to address.) Table 4, below, shows the item numbers corresponding to each mathematics objective and skill area. Table 5, below, shows the item numbers corresponding to each reading objective and skill area. (The actual item numbers corresponding to each objective, as well as the numbers of items relating to each objective, are listed, since they will be important to subsequent sections of this study which address the effects of applying of the Golden Rule procedures on the content representativeness of tests.)

Table 4

Correspondence of Items of Georgia Basic Skills Test (Form 01, Edition 13) in Mathematics to Test Skill Areas and Objectives*

Skill Area/Objective	Corresponding Items	Total # of Items
Concept Identification		
Objective 1	4, 13, 15, 20, 41, 49, 58, 60, 66, 75, 76, 82, 100	13
Objective 2	48, 68, 73, 92, 102	5
Objective 3	10, 16, 35, 40, 47, 50, 56, 78, 91, 109	10
Objective 4	18, 28, 84	3
Objective 5	7, 45, 55, 64, 103	
Component Operations		
Objective 6	6, 17, 37, 51, 59, 69	6
Objective 7	1, 5, 9, 11, 19, 21, 23, 27, 38, 52, 62, 72, 77, 79, 89, 93, 101, 107, 110	19
Objective 8	22, 26, 34, 61, 105	5
Objective 9	3, 36, 42, 43, 74, 81, 95, 96, 104	9
Problem Solving		
Objective 10	44, 90, 97, 108	4
Objective 11	8, 31, 33, 46, 83, 98	6
Objective 12	12, 24, 29, 54, 65, 70, 85, 88	8
Objective 13	2, 57, 86, 99, 106, 112	6
Objective 14	25, 30, 32, 39, 53, 63, 67, 71, 80, 87, 94, 111	12

* Information is not available on which objective Item 14 is intended to address.

Table 5

Correspondence of Items of Georgia Basic Skills Test (Form 01, Edition 13) in Reading to Test Skill Areas and Objectives

Skill Area/Objective	Corresponding Items	Total # of Items
Literal Comprehension		
Objective 1	31, 54, 81, 92, 101	5
Objective 2	9, 12, 13, 25, 35, 48, 52, 53, 55, 57, 65, 66, 69, 94, 95	15
Objective 3	3, 4, 32, 42, 43, 61, 90, 102, 103	9
Inferential Comprehension		
Objective 4	1, 2, 17, 18, 19, 24, 34, 56, 63, 68, 77, 79, 82, 97, 99,	15
Objective 5	14, 20, 44, 45, 58, 70, 75, 93	8
Objective 6	5, 21, 30, 36, 50, 96	6
Objective 7	8, 28, 39, 40, 86, 104, 105	7
Problem Solving		
Objective 8	6, 7, 27, 67, 85, 89	6
Objective 9	10, 22, 26, 37, 39, 71, 74, 83, 87, 91, 98	11
Objective 10	23, 46, 49, 59, 62, 72, 73, 78, 84	9
Objective 11	11, 16, 29, 47, 64, 80, 88	7
Objective 12	15, 33, 41, 51, 60, 76, 100	7

There is more than one form of Edition 13 of the Georgia Basic Skills Tests exists, and the number of items addressing each objective varies somewhat among forms. Table 6, below, shows the numbers of items typically associated with each objective on the Basic Skills Tests, as reported in the *Basic Skills Tests Fall 1987- Spring 1988 Interpretive Guide* (Georgia Department of Education, 1988a).

The reader will note that the skill areas assessed by the Georgia Eighth-Grade Mathematics Test are identical to those assessed by the Georgia Basic Skills Test in Mathematics, and that many of the mathematics objectives

assessed by the two tests are closely parallel. Similarly, the skill areas assessed by the Georgia Eighth-Grade Reading Test are identical to those assessed by the Georgia Basic Skills Test in Reading, and many of the reading objectives assessed by the two tests are parallel.

Table 6

Typical Numbers of Items Assessing Each Objective Across Forms of Georgia Basic Skills Tests

Mathematics Objectives	Number of Items	Reading Objectives	Number of Items
Objective 1	12-14	Objective 1	5-6
Objective 2	4-5	Objective 2	15-20
Objective 3	9-11	Objective 3	8-10
Objective 4	3-4	Objective 4	15-16
Objective 5	4-6	Objective 5	7-8
Objective 6	6-7	Objective 6	5-6
Objective 7	18-20	Objective 7	6-7
Objective 8	4-5	Objective 8	6-8
Objective 9	8-10	Objective 9	9-11
Objective 10	3-4	Objective 10	8-10
Objective 11	5-6	Objective 11	5-7
Objective 12	7-9	Objective 12	6-7
Objective 13	7-8		
Objective 14	12-14		

Data Collection and Reduction

Data for this study were provided by the Georgia Assessment Project of Georgia State University on computer readable magnetic tape. The data provided to this researcher consisted of records of each subject's gender, school system code, ethnic group membership, and selected option for each item on each of the four criterion-referenced tests.

The *Student Assessment Handbook* (State Assessment Programs, 1988b), which serves as the policy manual for student assessment practices in Georgia, provides some information on the purported data collection procedures. The *Handbook* indicates that both the Eighth-Grade Criterion-Referenced Tests and the Basic Skills tests are secure tests, with distribution controlled by test coordinators at the system and school level. Testing is optional for handicapped students and students with limited English proficiency; all other students are administered the Eighth-Grade Criterion-Referenced Tests in the spring of their Eighth-Grade year, and the Basic Skills Tests in the fall of their tenth grade year. Make-up sessions are held for absentees and students who are temporarily disabled at the regular testing time. Some assistive devices are allowed for students with visual or auditory handicaps, however, tests may not be read aloud to any students. If a student does not pass any portion of the Basic Skills Tests, the student must repeat and pass that portion at a regularly scheduled retesting time in order to become eligible to graduate with a regular high school diploma. All students used in this study were attempting the Basic Skills Tests for the first time.

Teachers serve as test administrators for both the Eighth-Grade Criterion-Referenced Tests and the Basic Skills Tests. Computer scannable answer sheets are used with the Mathematics and Reading Tests of both the Eighth-Grade Criterion-Referenced Tests and the Basic Skills Tests. Test answer sheets are returned to the Georgia Assessment Project of Georgia State University for computer scanning and scoring.

This study used data from students' performances on both the Eighth-Grade Criterion-Referenced Tests and the Basic Skills Tests. The data sets

from the two years were matched by personnel in the Georgia Assessment Project of Georgia State University. Student name, gender, race and school system were used as bases for matching. (Students in Georgia are not assigned unique identification numbers).

As stated above, data were provided to this researcher in the form of students' selected option for each test item. An answer key was also provided. Students' responses were scored by the researcher using the Data Step procedures of the SAS (SAS Institute, 1985a) computer program. In scoring students' responses, students who failed to provide an answer to a question were scored as having answered the question incorrectly. Students who marked more than one answer to a question were also scored as having answered incorrectly.

Creation of Synthetic Tests Conforming to the Golden Rule Stipulations

This study required the creation of a synthetic version of each of the four original standardized tests (Eighth-Grade Mathematics, Eighth-Grade Reading, Basic Skills Test in Mathematics, and Basic Skills Test in Reading) composed only of those items from the original tests which conformed to the Golden Rule stipulations for Type I items. The content validity of the synthetic tests was not specifically protected through inclusion of Type II items as necessary to ensure content representativeness.

As discussed in the previous chapter of this dissertation, Type I items must conform to two criteria. First, the correct answer rate for blacks, whites, and all examinees must not be less than 40% at the .05 level of statistical significance. Second, the correct answer rates of blacks and whites must differ by no more than 15% at the .05 level of statistical significance. Previous

researchers investigating the effects of applying the Golden Rule procedures have treated the correct answer rates observed in their sample as though they were population parameters (and thus disregarded the parts of the stipulations reading "at the .05 level of statistical significance") (G. Marco, personal communication, January 18, 1990; R. Linn, personal communication, January 19, 1990). For the purposes of this study, it was decided to treat the correct answer rates as sample statistics, and, in separate analyses, as population parameters. Thus, *two* synthetic versions of each test were created: one composed of items which conformed to the stipulations of Type I items based upon procedures using sample statistics, and another composed of items which conformed to the stipulations of Type I items when the observed correct answer rates for the sample were treated as population parameters.

For the sake of convenience, these synthetic tests composed of items which conform to the Golden Rule stipulations for Type I items will hence be designated by the prefix "GR-." Whether the synthetic test is based upon procedures for sample statistics or population parameters will also be designated in the prefix by an "S" for sample, or a "P" for population. For example, the synthetic version of the Eighth-Grade Mathematics Test composed only of Type I items, as determined by treating the observed correct answer rates as sample statistics, is designated as "GRS-Eighth-Grade Mathematics." The synthetic version of the Basic Skills Test in Reading composed only of Type I items, as determined when the observed correct answer rates are treated as population parameters, is designated as "GRP-BST Reading."

To create the synthetic tests in which the sample's correct answer rates were treated as population parameters, each subject's responses to each test item were scored, and the proportions of black examinees, white examinees, and of all examinees who answered each item correctly were determined, using the FREQUENCIES procedure of the SAS computer program (SAS Institute, 1985b). Items for which the correct answer rates of black examinees or of whites examinees were 40% or less were excluded from the synthetic tests. (Since black examinees and white examinees together comprise almost 99% of the sample, there were no items which over 40% of both black and white examinees answered correctly, but which less than 40% of the total sample answered correctly. However, the correct answer rates for the entire sample were calculated). Each item was also screened using the second criterion for Type I items. The correct answer rate for black examinees was subtracted from the correct answer rate for white examinees for each item. If the absolute value of the difference between the two proportions was 0.15 or greater, the item was excluded from the synthetic test, as not conforming to the second criterion for Type I items. The characteristics of the resultant synthetic tests are described in the next chapter of this dissertation.

To create the synthetic tests in which the sample statistics were treated as such, a one tailed t-test was performed using the proportion of correct answers for each group, testing the null hypothesis: H_O: P=0.40 against the alternative hypothesis: H_A: P<0.40, using a Type-I error rate of 0.05. Items for which the correct answer rates of black examinees or of whites examinees was less than 40% at the 0.05 level of statistical significance were excluded from these synthetic tests. (Since there were 1269 white examinees and 517 black

examinees in the sample, the critical region for rejection of the null hypothesis was $t < -1.645$). Each item was also compared to the second criterion for Type I items. The correct answer rate for black examinees was subtracted from the correct answer rate for white examinees for each item. The standard error of estimate of each of these differences was computed using the formula $SE_d = \sqrt{(SE_w^2 + SE_b^2)}$, where SE_w^2 is the variance error of estimate of the proportion of correct answers for white examinees, and SE_b^2 is the variance error of estimate of the proportion of correct answers for black examinees. A one tailed t-test was performed using the absolute value of each of these differences, testing the null hypothesis: $H_O: \Delta = 0.15$ against the alternative hypothesis: $H_A: \Delta > 0.15$, (where Δ is the absolute value of the difference between the proportions correct for black and white examinees), using a Type-I error rate of 0.05. If the absolute value of the difference between the two proportions was 0.15 or greater, at the .05 level of statistical significance, the item was excluded from these synthetic tests, as not conforming to the second criterion for Type I items. (Since there were 1269 white examinees and 517 black examinees in the sample, the critical region for rejection of the null hypothesis was $t > 1.645$). The characteristics of the resultant synthetic tests are described in the next chapter of this dissertation.

Application of the Golden Rule "winnowing" procedures described above to all items in each of the four tests led to some items being classified as Type I items. These items were retained in the synthetic versions of the tests. Other items were omitted from the synthetic versions of the tests because of failure to conform to either the first or second criterion for Type I items, or failure to conform to both criteria. The original standardized tests are secure,

thus the text of the items failing to conform to Type I specifications cannot be reproduced here. Appendix F contains the correct answer rates for black examinees and white examinees on the items of the Eighth-Grade Reading Test, and the t-statistics for testing H_O: $P = 0.40$ against H_A: $P < 0.40$. It also shows the difference between the proportion correct for black and white examinees on the items of the Eighth-Grade Reading Test, and the t-statistics for testing H_O: $\Delta = 0.15$ against the alternative hypothesis: H_A: $\Delta > 0.15$. Appendix G contains the correct answer rates for the entire sample on the items of the Eighth-Grade Reading Test, and the t-statistic for testing H_O: $P=0.40$ against H_A: $P<0.40$. Appendices H and I contain this information for the items of the Eighth-Grade Mathematics Test. Appendices J and K contain the information for the items of the Basic Skills Test in Reading, and Appendices L and M contain the information for the Basic Skills Test in Mathematics.

Methodology for Addressing the Research Questions

This study addressed two major research questions. The first major question was: Is application of the Golden Rule procedures effective in reducing the adverse impact of test use? The second major question was: Is application of the Golden Rule procedures effective in reducing test bias? A set of hierarchical research questions which define these two major research questions was presented in Chapter II of this dissertation. Investigation of these subsidiary research questions involved a series of examinations of the characteristics of, and relationships between and among, the synthetic tests which conformed to the Golden Rule stipulations and the original

standardized tests. The procedures for investigating each subsidiary research question are described in the remainder of this chapter.

Methodology for Research Question 1

The first major research question of this study, which concerns whether application of the Golden Rule procedures is effective in reducing the adverse impact of test use, was examined empirically through one subsidiary research question, Question 1A:

1A. Does application of the Golden Rule procedures make the mean of scores of blacks and whites more similar?

This question was investigated by comparing the difference between the average total score for black examinees and the average total score for white examinees on each original standardized test with the difference between the average total scores for black examinees and white examinees on the corresponding synthetic tests which conform to the stipulations for Golden Rule Type I items. Separate comparisons were made using the synthetic tests in which the correct answer rates were treated as sample statistics, and those in which they were treated as population parameters. These comparisons are described more fully in the following paragraphs.

To examine whether the average total scores for black and white examinees were more similar on the GRS-Eighth-Grade Reading Test than on the original Eighth-Grade Reading Test, the average total scores for both black and white examinees were determined for both the original test and for the GRS-Eighth-Grade Reading Test, using the MEANS procedure of the SAS

computer program (SAS Institute, 1985b). The difference between the average total scores for black and white examinees was then determined.

Since the items selected for inclusion in the synthetic tests actually represent a sample of possible items, it was desired to use a measure of the variance of the total score on the synthetic tests which would reflect the variation of the total score across samples of items of size equal to the length of the synthetic tests. The subject sample was considered sufficiently large that persons were ignored as a source of variance in this analysis. The variance of the differences between black and white examinees' total scores on random samples of n items selected from N possible items is given by the formula:

$$\sigma^2_{dTot} = n\,\sigma^2_{dp}\,(1 - n/N)$$

where σ^2_{dp} is the variance of the differences between the proportions of correct answers for black and white examinees across all N items on the original test, calculated using the formula:

$$\sigma^2_{dp} = \frac{\sum_{i=1}^{N} (dp_i - \mu_{dp})^2}{(N-1)}$$

[Note the use of (N-1), rather than N, in the denominator of the formula (Jaeger, 1984, p. 42; Cochran, 1977, p. 23).]

The above formula gives the variance of the difference between black and white examinees' total scores across random samples of items. Of course, the items included in the synthetic test forms created to conform to the Golden Rule stipulations were not randomly selected. However, the above formula may be used to estimate how unlikely it would be to observe a

difference of the magnitude observed on a synthetic test, were the items selected for inclusion in the synthetic test not systematically different from randomly selected samples of items. The expected difference between the average total scores for black and white examinees on a test of randomly selected items of length equal to the synthetic test was determined by multiplying the difference observed on the original standardized test by the ratio of the number of items in the synthetic test to the number of items in the original test. Since the above formula yields a population variance, a non-directional z-test was performed, testing the null hypothesis H_O: $\Delta_{8R'} = d_{GRS\text{-}8R}$ against the alternative hypothesis: H_A: $\Delta_{8R'} \neq d_{GRS\text{-}8R}$, where $\Delta_{8R'}$ is the difference between black and white examinees' average scores on the population of randomly equivalent original tests, adjusted for test length (by multiplying the observed difference by the ratio of the number of items on the synthetic and original tests); and $d_{GRS\text{-}8R}$ is the difference between black and white examinees' average scores on the GRS-Eighth-Grade Reading Tests. A Type-I error rate of five percent was used for the z-test. (A directional alternative hypothesis, H_A: $\Delta_{R'} > \Delta_{GRS\text{-}8R}$, might seem indicated here, since the Golden Rule procedures are intended to reduce the disparity between the mean scores for black and white examinees. However, Linn and Drasgow (1987) and Marco (1988) have suggested that application of the Golden Rule procedures might actually increase the disparity between the mean scores of black and white examinees. For this reason, a non-directional alternative hypothesis was used with this research question).

The procedures described above for comparing the similarity of black and white examinees' total scores on the original Eighth-Grade Reading Test

and on the GRS-Eighth-Grade Reading Test were repeated for each of the other logical test comparisons: the original Eighth-Grade Reading Test and the GRP-Eighth-Grade Reading Test, the original Eighth-Grade Mathematics Test and the GRS-Eighth-Grade Mathematics Test, the original Eighth-Grade Mathematics Test and the GRP-Eighth-Grade Mathematics Test, the original BST Mathematics Test and the GRS-BST Mathematics Test, the original BST Mathematics Test and the GRP-BST Mathematics Test, the original BST Reading Test and the GRS-BST Reading Test, and the original BST Reading Test and the GRP-BST Reading Test. Table 7, below, lists each corresponding original and synthetic test for which the differences in average scores for black and white examinees were tested, as described above. The results of investigation of Research Question 1A are presented in the next chapter of this dissertation.

Methodologies for Addressing Research Question 2

The second major research question of this study, which concerns whether application of the Golden Rule procedures is effective in reducing test bias, was examined empirically through investigation of ten subsidiary research questions, Questions 2A through 2J. Questions 2A through 2E concern examination of the effects of applying the Golden Rule procedures on a test's internal structure. Questions 2F through 2H concern examination of the effects of applying the Golden Rule procedures on the relationships of a test to various external factors. Question 2I concerns examination of the effects of applying the Golden Rule procedures on a test's content and format. Question 2J addresses the effects of applying the Golden Rule procedures on what Cole and Moss (1989) refer to as "constructs in context." The

methodology for examination of each of these subsidiary research questions is described in the remainder of this chapter.

Table 7
Original and Synthetic Tests Compared in Research Question 1A*

Original Standardized Test	Corresponding Synthetic Test
Eighth-Grade Reading Test	GRS-Eighth-Grade Reading Test
Eighth-Grade Reading Test	GRP-Eighth-Grade Reading Test
Eighth-Grade Mathematics Test	GRS-Eighth-Grade Mathematics Test
Eighth-Grade Mathematics Test	GRP-Eighth-Grade Mathematics Test
BST Mathematics Test	GRS-BST Mathematics Test
BST Mathematics Test	GRP-BST Mathematics Test
BST Reading Test	GRS-BST Reading Test
BST Reading Test	GRP-BST Reading Test

*Comparisons made in terms of differences between average total scores for black and white examinees.

The reader will note that the magnitude of a number of the effects explored in research Questions 2A through 2J were not tested for statistical significance. There were two reasons why the results of investigating some questions were not tested for statistical significance. First, statistical tests cannot readily be applied to many of the effects examined in this section, due to the nature of the original and the synthetic tests. Traditional statistical hypothesis tests either assume that observations (or effects) are independent, or they assume that observations are related and paired. Since the synthetic

tests are composed of subsets of the items of the original tests, subjects' scores on the two sets of tests are certainly not independent. When test items, rather than subjects, are the unit of analysis, the picture is more clouded: some items are identical in the two types of tests, while other items exist only on the original test, and cannot be paired with any item on the synthetic test. Testing many of the effects examined in this section, (for example, the effect of applying of the Golden Rule procedures on the average item-total correlation or the average item difficulty for a test), involves consideration of the independence of both tests and subjects, and would result in violation of some of the assumptions of traditional hypothesis testing methods. Many of the effects are simply not amenable to traditional hypothesis testing. Second, statistical significance is not synonymous with practical importance, and, when issues of bias are at question, practical importance is the issue of paramount concern. Tests of statistical significance are designed to detect whether the effects found are likely to exist in the population from which the sample has been selected, or are likely due to chance variations due to sampling or measurement error. The sample size in this study is sufficiently large that, to the extent that the sample can be taken as representative of some population, any effects found for the sample which were of sufficient size to be of practical importance would, in all likelihood, be found to be statistically significant, if an appropriate hypothesis test could be applied. The number of items in each test is also sufficiently large that, for effects that depend on the number of items rather than the number of subjects, any effects found which

were of sufficient size to be of practical importance would, in all likelihood, be found to be statistically significant, could an appropriate hypothesis test be applied.

Methodology for Research Question 2A

Research Question 2A addresses the effect of applying the Golden Rule procedures on the overall internal-consistency reliability of tests. This question was examined by computing an index of internal consistency reliability, Cronbach's coefficient alpha (Cronbach, 1951; Thorndike, 1982), for each of the four original tests, and also for each of the eight synthetic tests. The formula used for computing these coefficients was:

$$\alpha = \frac{n}{(n-1)} \left(1 - \frac{\sum s_i^2}{s_t^2}\right)$$

where n = number of items in the test
s_i^2 = variance of item i
s_t^2 = variance of the total test.

Shortening a test reduces its reliability. The Spearman-Brown Prophecy Formula (c.f., Thorndike, 1982; Allen & Yen, 1979) can be used to estimate the effect of changing a test's length on the test's reliability. Since the original tests were all longer than their corresponding synthetic tests, the Spearman-Brown Prophecy Formula was used to correct for any differences in reliability which were due solely to differences in test length. After coefficient alpha was determined for each of the synthetic tests, the Spearman-Brown Prophecy Formula was applied to estimate what the reliability of the synthetic test

would be, were it the same length as the original test. The formula used for the Spearman-Brown corrections was (c.f., Allen & Yen, 1979, p. 85):

$$r_{mx} = (m\alpha) / (1 + (m-1)\,\alpha)$$

where r_{mx} = the estimated reliability of the lengthened test
m = the proportion by which the test is lengthened
α = coefficient alpha, the estimated internal consistency reliability of the shortened synthetic test

The values obtained for coefficient alpha (adjusted for length for the synthetic tests) are presented and discussed in the next chapter of this dissertation, for each logical comparison of original to synthetic test (i.e., Eighth-Grade Reading to GRS-Eighth-Grade Reading and to GRP- Eighth-Grade Reading; Eighth-Grade Mathematics to GRS-Eighth-Grade Mathematics and to GRP- Eighth-Grade Mathematics; etc.) The differences between the coefficients for the original and synthetic tests were not tested for statistical significance.

Methodology for Research Question 2B

Research Question 2B addresses the effect of applying the Golden Rule procedures on the difference between the reliability of a test for black and white examinees. This question was examined by calculating Cronbach's coefficient alpha separately for black and white examinees for each original and synthetic test. The procedure and formula for computing coefficient alpha were given in the previous section of this chapter. The Spearman-Brown Prophecy Formula was used to estimate the reliability of the synthetic tests for black and white examinees, were the synthetic tests of the same

length as the tests from which they were derived. (Use of the Spearman-Brown Prophecy Formula is described in the previous section of this chapter.) The difference between coefficient alpha for black examinees and for white examinees was then determined for each original test. The difference between coefficient alpha, adjusted for test length, for black examinees and for white examinees, was also determined for each synthetic test.

In the next chapter of this study, the differences between reliabilities for black and white examinees for the original and synthetic tests (adjusted for test length) are presented and discussed for each logical comparison of original to synthetic test (i.e. Eighth-Grade Reading to GRS-Eighth-Grade Reading and to GRP-Eighth-Grade Reading; Eighth-Grade Mathematics to GRS-Eighth-Grade Mathematics and to GRP- Eighth-Grade Mathematics; etc.) The differences between these differences in Coefficient Alpha were not tested for statistical significance.

Methodology for Research Question 2C

Research Question 2C addresses the effect of applying the Golden Rule procedures, (without specific protection of content representativeness through inclusion of Type II items), on the content representativeness of a test. The numbers of items corresponding to each test objective were given for each of the four original tests in Tables 2, 3, 4, and 5 above. To examine this research question, the numbers and proportions of items addressing each objective of the each of the original tests was determined. The numbers and proportions of items addressing each objective were also determined for the synthetic tests.

The proportions of items addressing each objective of the original tests, and the corresponding proportions for the synthetic tests, are presented and discussed in the next chapter of this dissertation, for each logical comparison of original to synthetic test (i.e. Eighth-Grade Reading to GRS-Eighth-Grade Reading and to GRP-Eighth-Grade Reading; Eighth-Grade Mathematics to GRS-Eighth-Grade Mathematics and to GRP-Eighth-Grade Mathematics; etc.) To test the effect of applying of the Golden Rule procedures on the overall content of a test, a Chi-square Goodness-of-Fit test was performed for each synthetic test (Siegel & Castellan, 1988). The numbers of items addressing each objective on the synthetic test were treated as the "observed" values in the Chi-square Goodness-of-Fit test. The "expected" values were determined by multiplying the number of items in the synthetic test by the proportion of items addressing each objective in the original test. For each logical comparison of synthetic to original test, the null hypothesis that the observed proportions could have been sampled from a population with the given expected values was tested (Siegel & Castellan, 1988, p. 45). The degrees of freedom for each test was equal to k-1, where k is the number of objectives on the test. A Type-I error rate of 0.05 was used for each test. The results of these Chi-square Goodness of Fit tests are presented in the subsequent chapter of this dissertation.

The question of whether the proportion of items addressing each objective on a synthetic test was different from the proportion that would be expected to address that objective, had a number of items equal to the length of the synthetic test been drawn at random from the items of the original test was also investigated, by determining the proportion of items addressing each

objective in each original and corresponding synthetic test. The proportions of items addressing each objective in the original tests were considered population parameters. A 95% confidence interval was determined for the proportion of items addressing each objective for the synthetic tests (Rovell, 1962). If this confidence interval contained the population proportion (i.e., the proportion addressing the objective on the original test) then it was considered tenable that the proportion of items addressing that objective on the synthetic test was not significantly different (at the 0.05 level) from what would be expected from a random sample of items from the original test. Since the original tests had 11 to 14 objectives each, by the Bonferonni inequality (Glass and Hopkins, 1984, p. 381) the experiment wise error rate across objectives for each original test of approximately 55% to 70%.

Several forms of the edition of the Basic Skills Test used in this study exist, and the number of items addressing each objective varies from form to form. Table 6 above shows the range of items typically associated with each objective on the Basic Skills Tests. To determine whether the number of items addressing each objective on the synthetic BST tests would fall within this typical range, were the synthetic tests of length equal to the original tests, the proportion of items addressing each objective in the synthetic tests was multiplied by the number of items in the original test. This resulted in projected numbers of items per objective for the synthetic tests, were they of length equal to the original tests. These projected numbers of items per objective for the synthetic tests, compared to the typical range of items for each objective, are also presented and discussed in the subsequent chapter of this dissertation.

Methodology for Research Question 2D

Research Question 2D addresses the effect of applying the Golden Rule procedures on the average item difficulty of a test. This research question was investigated by determining the correct answer rate (across ethnic groups) for each item on each original and synthetic test. These correct answer rates (item difficulties) were then averaged for each test, to yield the average proportion of correct responses to items for each test.

There is one source of error variance in the average item difficulties of the original tests: subjects. Since the items selected for inclusion in the synthetic tests actually represent a sample of possible items, there are two sources of error variance in the average item difficulties of the synthetic tests: subjects and items. It was desired to use a measure of the variance of the average item difficulties on the synthetic tests which would reflect the variation of the average item difficulties across samples of items of size equal to the length of the synthetic tests. The subject sample was considered sufficiently large that subjects were ignored as a source of variance for this analysis. The variance of the average item difficulty on random samples of n items selected from N possible items is given by the formula:

$$\sigma^2_{\overline{id}} = (\sigma^2_{id} / n)\,(1 - n/N)$$

where σ_{id}^2 is the variance of the proportion of correct answers across all N items on the original test calculated using the formula:

$$\sigma^2_{id} = \frac{\sum_{i=1}^{N} (id_i - \mu_{id})^2}{(N-1)}$$

[Note the use of (N-1), rather than N, in the denominator of the formula (Jaeger, 1984, p. 42; Cochran, 1977, p. 23).]

The above formula gives the variance of tests' average item difficulty across random samples of items. The items included in the synthetic test forms created to conform to the Golden Rule stipulations were not randomly selected. However, the above formula may be used to estimate how unlikely it would be to observe an average item difficulty of the magnitude observed on a synthetic test, were the items selected for inclusion in the synthetic test not systematically different from randomly selected samples of items. Since the above formula yields a population variance, a non-directional z-test was performed, testing the null hypothesis H_O: $Diff_{8R} = Diff_{GRS\text{-}8R}$ against the alternative hypothesis: H_A: $Diff_{8R} \neq Diff_{GRS\text{-}8R}$, where $Diff_{8R}$ is the average item difficulty of a population of tests that are randomly equivalent to the original Eighth-Grade Reading Test, but of length equal to the GRS-Eighth-Grade Reading Test, and $Diff_{GRS\text{-}8R}$ is the average item difficulty of the GRS-Eighth-Grade Reading Test. A Type-I error rate of five percent was used for the z-test. (A non-directional, rather than a directional, alternative hypothesis was chosen here, because Linn and Drasgow (1987) and Marco (1988) have suggested that the direction of effects of applying the Golden Rule procedures might be different from what one would anticipate on the basis of the wording of the procedures).

The procedures described above for comparing the average item difficulties of the original Eighth-Grade Reading Test and the GRS-Eighth-Grade Reading Test were repeated for each of the other logical test comparisons, as listed in Table 7, above. The results of investigation of Research Question 2D are presented in the next chapter of this dissertation.

Methodology for Research Question 2E

Research Question 2E addresses the effect of applying the Golden Rule procedures on the average item-total correlation of a test. To address this research question, the scored responses to each item were correlated with total test score. Fisher's Z-transformation, as described by Glass and Hopkins (1984, p. 304-307), was then performed on each item-total correlation. These Fisher Z's were then averaged for each test. The average Fisher Z was then converted to a correlation coefficient, which represented the average item-total correlation of the test.

There is one source of error variance in the average item-total correlations of the original tests: subjects. Since the items selected for inclusion in the synthetic tests actually represent a sample of possible items, there are two sources of error variance in the average item-total correlations of the synthetic tests: subjects and items. To compare the average item-total correlation of synthetic and corresponding original tests, a measure of the variance of the average Fisher's Z's (corresponding to the average item-total correlations) on the synthetic tests which reflected the variation of the average Fisher Z's (and average item-total correlations) across samples of items of size equal to the length of the synthetic tests was used. Subjects were ignored as a source of variance in these analyses. The variance of the average Fisher's Z's for item-total correlations on random samples of n items selected from N possible items is given by the formula:

$$\sigma_{\bar{z}}^2 = (\sigma_z^2/n)(1 - n/N)$$

where σ_z^2 is the variance of the Fisher's Z's across all N items on the original test, as described in Chapter III (Jaeger, 1984, p. 42; Cochran, 1977, p. 23).

The above formula yields the variance of tests' average Fisher Z's (corresponding to the average item-total correlations) across random samples of items. The items included in the synthetic test forms created to conform to the Golden Rule stipulations were not randomly selected. However, the above formula may be used to estimate how unlikely it would be to observe an average Fisher's Z of the magnitude observed on a synthetic test, were the items selected for inclusion in the synthetic test not systematically different from randomly selected samples of items. Since the above formula yields a population variance, a non-directional z-test was performed, testing the null hypothesis H_O: $IT_{8R} = IT_{GRS\text{-}8R}$ against the alternative hypothesis: H_A: $IT_{8R} \neq IT_{GRS\text{-}8R}$, where IT_{8R} is the Fisher's Z corresponding to the average item-total correlation of a population of tests that are randomly equivalent to the original Eighth-Grade Reading Test, but of length equal to the GRS-Eighth-Grade Reading Test, and $IT_{GRS\text{-}8R}$ is the Fisher's Z corresponding to the average item-total correlation of the GRS-Eighth-Grade Reading Test. A Type-I error rate of five percent was used for the z-test.

The procedures described above for comparing the average item-total correlations of the original Eighth-Grade Reading Test and the GRS-Eighth-Grade Reading Test were repeated for each of the other logical test comparisons, as listed in Table 7, above. The results of investigation of Research Question 2E are presented in the next chapter of this dissertation.

Methodology for Research Question 2F

Research Question 2F addresses the effect of applying the Golden Rule procedures on the predictive validity of a test. In the examination of this research question, Eighth-Grade test results were used to predict total scores

on the unaltered Basic Skills Tests. The correlation between the Eighth-Grade Reading Test and the BST Reading Test was determined, as was the correlation between the GRS-Eighth-Grade Reading Test and the BST Reading Test, and the correlation between the GRP-Eighth-Grade Reading Test and the BST Reading Test. The correlations between the Eighth-Grade Mathematics Test and the BST Mathematics Test, the GRS-Eighth-Grade Mathematics Test and the BST Mathematics Test, and the GRP-Eighth-Grade Mathematics Test and the BST Mathematics Test were also determined.

Shortening a test lowers its reliability, which attenuates its predictive validity. The synthetic tests are shorter than the corresponding original tests, and, as described above, the Spearman-Brown Prophecy Formula was used to project what the reliabilities for the synthetic tests would have been, had the synthetic test been of length equal to the original tests. Allen and Yen (1979, p. 98) provide a formula for correcting a coefficient of prediction for the unreliability of a predictor. This formula projects what the coefficient of prediction would be, were the predictor measured with perfect reliability. A variant of this formula was used estimate what the coefficient of prediction for each synthetic test would have been, had the synthetic test had a reliability coefficient equal to that estimated by the Spearman-Brown Prophecy Formula (which projected what the reliability for the synthetic test would have been, had the synthetic test been of the same length as the original test). The formula used was: $r_{12}' = r_{12} (\sqrt{r_{11}}' / \sqrt{r_{11}})$ where r_{11} is the observed reliability of the synthetic test, r_{11}' is the projected reliability for the synthetic tests , had it been of the same length as the original test, r_{12} is the observed correlation between the synthetic test and the criterion, and r_{12}' is the projection of what

the coefficient of prediction for the synthetic test would have been, had the synthetic test had a reliability coefficient equal to that estimated by the Spearman-Brown Prophecy Formula.

A 95% confidence interval was determined for each coefficient of predictive validity (adjusted for unreliability due to length for the synthetic tests), using Fisher's Z-transformation, as described by Glass and Hopkins (1984, p. 304-307).

Since both the original and synthetic Eighth-Grade tests were used to predict the original BST tests, this situation is amenable to testing for statistical significance, using the procedure for testing dependent correlation coefficients, as described by Glass and Hopkins (1984, p. 310-311). The difference between the coefficients of prediction for the Eighth-Grade Reading Test and the GRS-Eighth-Grade Reading Test was tested for statistical significance, as were the differences for the Eighth-Grade Reading Test and the GRP-Eighth-Grade Reading Test, the Eighth-Grade Mathematics Test and the GRS-Eighth-Grade Mathematics Test, and for the Eighth-Grade Mathematics Test and the GRP-Eighth-Grade Mathematics Test. These tests were performed by calculating the t-statistic for the null hypothesis: H_O: $\rho_{8,BST} = \rho_{GR8,BST}$ against the alternative hypothesis: H_A: $\rho_{8,BST} \neq \rho_{GR8,BST}$, where $\rho_{8,BST}$ is the correlation between an original eighth-grade test and a corresponding BST test in the population of tests that are randomly equivalent to the original tests used in this dissertation, but of length equal to the synthetic tests; and $\rho_{GR8,BST}$ is the correlation between a synthetic eighth-grade test and a corresponding BST test. A Type-I error rate of 0.01 was used for each of the four hypothesis tests. By the Bonferonni inequality (Glass and

Hopkins, 1984, p. 381) this results in an experiment wise error rate for this research question of not more than five percent.

The coefficients of prediction, with confidence intervals, for each original test and its corresponding synthetic tests, are presented, compared, and discussed in the subsequent chapter. The results of the t-tests for statistically significant differences between the predictive validity of the original Eighth-Grade tests and the corresponding synthetic tests are also reported.

Methodology for Research Question 2G

Research Question 2G addresses the effect of applying the Golden Rule procedures on the difference between a test's coefficients of prediction for black examinees and for white examinees. In the examination of this research question, eighth-grade test results were again used to predict total scores on the unaltered Basic Skills Tests. The correlation between the Eighth-Grade Reading Test and the BST Reading Test was determined separately for black examinees and for white examinees, as were the correlations between the GRS-Eighth-Grade Reading Test and the BST Reading Test, and the GRP-Eighth-Grade Reading Test and the BST Reading Test. The correlations between the Eighth-Grade Mathematics Test and the BST Mathematics Test, the GRS-Eighth-Grade Mathematics Test and the BST Mathematics Test, and the GRP-Eighth-Grade Mathematics Test and the BST Mathematics Test were also determined separately for black and white examinees.

Shortening a test lowers its reliability, which attenuates its predictive validity. The synthetic tests are shorter than the corresponding original tests, and, as described above, the Spearman-Brown Prophecy Formula was used to

project what the reliabilities for the synthetic tests would have been, had the synthetic test been of length equal to the original tests. Allen and Yen (1979, p. 98) provide a formula for correcting a coefficient of prediction for the unreliability of a predictor. This formula projects what the coefficient of prediction would be, were the predictor measured with perfect reliability. A variant of this formula was used estimate what the coefficient of prediction for each synthetic test would have been, had the synthetic test had a reliability coefficient equal to that estimated by the Spearman-Brown Prophecy Formula (which projected what the reliability for the synthetic test would have been, had the synthetic test been of the same length as the original test). The formula used was: $r_{12}' = r_{12} (\sqrt{r_{11}}' / \sqrt{r_{11}})$ where r_{11} is the observed reliability of the synthetic test, r_{11}' is the projected reliability for the synthetic test , had it been of the same length as the original test, r_{12} is the observed correlation between the synthetic test and the criterion, and r_{12}' is the projection of what the coefficient of prediction for the synthetic test would have been, had the synthetic test had a reliability coefficient equal to that estimated by the Spearman-Brown Prophecy Formula.

The difference between the coefficient of prediction for black examinees and for white examinees was then determined for the Eighth-Grade Tests in Reading and Mathematics. The differences between the coefficients of prediction for black examinees and for white examinees, adjusted for differences in reliabilities attributable to differences in test length, were also determined for the GRS-Eighth-Grade Tests in Reading and Mathematics and the GRP-Eighth-Grade Tests in Reading and Mathematics. Since black and white examinees comprise separate subsamples, this situation is amenable to

testing for statistical significance, using the procedure for testing independent correlation coefficients, as described by Glass and Hopkins (1984, p. 307-309). The null hypothesis that the predictive validity is the same for black and white examinees ($H_O: \rho_W = \rho_B$) was tested for each original and synthetic eighth-grade test. The projected values were used for synthetic tests, as described above. (Subjects were the source of variance in these tests, rather than items, as in some other hypothesis tests in this dissertation.)

These differences between the coefficients of prediction for black and white examinees are presented, compared, and discussed for each original test, and its corresponding synthetic tests, in the next chapter of this dissertation.

Methodology for Research Question 2H

Research Question 2H addresses the effect of applying the Golden Rule procedures on the difference between the regression equations relating a test to its criterion for black examinees and for white examinees. Again, eighth-grade test results were used to predict total scores on the unaltered Basic Skills Tests. In the investigation of this research question, the regression equations predicting BST Reading total scores from Eighth-Grade Reading total scores were determined separately for black examinees and for white examinees. The regression equations for predicting BST Reading total scores from GRS-Eighth-Grade Reading total scores were also determined for black examinees and for white examinees, as were regression equations predicting BST Reading total scores from GRP-Eighth-Grade Reading total scores. Likewise, the regression equations for black examinees and for white examinees were determined for predicting BST Mathematics total scores from the Eighth-

Grade Mathematics total scores, from total scores on GRS-Eighth-Grade Mathematics test and from total scores on the GRP-Eighth-Grade Mathematics test. A 95% confidence interval was determined for the slope and intercept in each regression equation. The GLM procedure of the SAS computer program was used for these calculations (SAS Institute, 1985b).

The degree of differential regression, if any, for the original and synthetic eighth-grade tests was also evaluated by fitting a regression model in which ethnic group membership was included as a predictor. (Only black and white ethnic groups were included in these analyses). This allowed evaluation of the interaction of ethnic group membership with eighth-grade test performance in the prediction of BST test performance.

The next chapter presents the regression equations relating each original and synthetic Eighth-Grade test to the corresponding original Basic Skills Test. The slopes and intercepts of the regression equations for black examinees and for white examinees are compared for each test, and the degree of similarity of regression equations for black and white examinees when the original Eighth-Grade tests are used as predictors is compared to that when the synthetic Eighth-Grade tests are used as predictors. The regression equations in which ethnic group membership is used as a predictor are also presented, and the coefficients associated with the interactions of ethnic group membership and eighth-grade test results are discussed.

Methodology for Research Question 2I

Research Question 2I addresses the content and format of the items of the original tests that did not meet the criteria to be classified as Type I items. (The examination of the content representativeness of the tests composed

only of Type I items was discussed above, in the section presenting the methodology used in addressing Research Question 2C). To address this research question, the researcher read the two original reading tests and determined that there were five major formats used in item presentation in the tests, and six major types of content presented in the items. These categories were used to form a six-by-five matrix of item content by format for the reading tests. An identical treatment led to the construction of a seven-by-four matrix of content by format for the mathematics tests. Each item of the original tests was then assigned to a cell of the matrix. (The researcher was not aware of an item's classification as Type I or Type II when completing this sorting). The numbers of items in each cell, the numbers of items excluded from both the GRS- and GRP- synthetic tests, and the number excluded from the GRP- synthetic tests only were then determined for each original test. Performance of a Chi-square test on the matrices was not feasible, since there were a large number of empty cells in each matrix. The results of this sorting are presented in the next chapter. The reader should note that no procedures were followed to establish the construct validity of the categories of item formats or content, nor were any procedures used to ensure the reliability of the assignment of items to categories. This investigation must be regarded as preliminary.

Large numbers of items were classified as Type II items, and it was beyond the scope of this dissertation study to compose panels of experts to review the items for discriminatory language or content. Since the tests themselves are secured, and the numbers of items classified as Type II are

very large, it was infeasible to reproduce each item classified as Type II in this report.

Methodology for Research Question 2J

Research Question 2J addresses the effect of applying the Golden Rule procedures to both a test and the criterion it is intended to predict (where the criterion is also a test) on the differences between the coefficients of prediction and the regression equations for black examinees and for white examinees. In Research Questions 2G and 2H, discussed above, original and synthetic eighth-grade test scores were used to predict original BST test scores. In the investigation of this research question, original eighth-grade tests scores were used to predict original BST scores, and synthetic eighth-grade test scores were used to predict scores on the synthetic BST tests. The similarity of coefficients of prediction and regression equations for black and white examinees were then compared between the unaltered original tests and the synthetic tests which conform to the Golden Rule stipulations.

To investigate this research question, the correlation coefficients between the total scores on the original Eighth-Grade Test in Reading and on the BST Reading Test were determined for black examinees and for white examinees. The regression equations predicting BST Reading Test total scores from total scores on the original Eighth-Grade Test in Reading were also determined for black examinees and for white examinees. The correlations of total scores on the original Eighth-Grade Test in Mathematics and on the BST Mathematics Test were also determined for black examinees and for white examinees, as were regression equations predicting the BST Mathematics Test

total scores from total scores on the original Eighth-Grade Test in Mathematics.

The correlation coefficients between the total scores on the GRS-Eighth-Grade Test in Reading and on the GRS-BST Reading Test were determined for black examinees and for white examinees, as were regression equations predicting GRS-BST Reading Test total scores from total scores on the GRS-Eighth-Grade Test in Reading for black and for white examinees. These relationships were also determined for the GRP-Eighth-Grade Test in Reading and the GRP-BST Reading Test, for the GRS-Eighth-Grade Test in Mathematics and the GRS-BST Mathematics Test, and for the GRP-Eighth-Grade Test in Mathematics and the GRP-BST Mathematics Test.

Shortening a test reduces its reliability, which attenuates its correlation with another test. The synthetic tests are shorter than the corresponding original tests, and, as described above, the Spearman-Brown Prophecy Formula was used to project what the reliabilities for the synthetic tests would have been, had the synthetic test been of length equal to the original tests. Allen and Yen (1979, p. 98) provide a formula for correcting a correlation coefficient for attenuation due to unreliability of a test. This formula projects what the correlation coefficient would be, were the test measured with perfect reliability. A variant of this formula was used to estimate what the correlation between the synthetic eighth-grade and BST tests would have been, had the synthetic tests had reliability coefficients equal to those estimated by the Spearman-Brown Prophecy Formula (which projected what the reliability for each synthetic test would have been, had the synthetic test been of the same length as the corresponding original test). The

formula used was: $r_{12}' = r_{12}\,[(\sqrt{r_{11}'}\ \sqrt{r_{22}'})/\ (\sqrt{r_{11}}\ \sqrt{r_{22}})]$ where r_{11} is the observed reliability of the synthetic eighth-grade test, r_{22} is the observed reliability of the synthetic BST test, r_{11}' is the projected reliability for the synthetic eighth-grade test , had it been of the same length as the corresponding original test, r_{22}' is the projected reliability for the synthetic BST test , had it been of the same length as the corresponding original test, r_{12} is the observed correlation between the synthetic eighth-grade and BST tests, and r_{12}' is the projection of what the correlation between the synthetic eighth-grade and BST tests would have been, had the synthetic tests had a reliability coefficients equal to those estimated by the Spearman-Brown Prophecy Formula.

To further clarify the effect of applying the Golden Rule procedures to both a test and the criterion it is to predict, regression models were fit in which ethnic group membership was included as a predictor. (Only black and white ethnic groups were included in these analyses).

In the next chapter of this dissertation, the correlations and regression equations for blacks and whites for the original tests and for the synthetic tests are presented and discussed, and the similarity between the values for black examinees and white examinees on the original tests is compared to that on the synthetic tests.

CHAPTER IV

RESULTS

In the previous chapter, the procedures followed in investigating a set of hierarchical research questions were presented. This chapter presents the results of empirical analyses described in the previous chapter. The chapter is divided into three sections. The first section presents the results of creating the synthetic tests conforming to the Golden Rule stipulations for Type I items. This dissertation study addressed two major research questions, and the second section of this chapter presents the results of investigating Research Question 1, which addressed whether application of the Golden Rule procedures is effective in reducing the adverse impact of test use. The third section of this chapter presents the results of investigating Research Question 2, which addressed whether application of the Golden Rule procedures is effective in reducing test bias. Research Question 2 was addressed through investigation of a series of subsidiary research questions; the results of each are discussed in turn. The final section of this chapter presents a summary of the results of empirical investigation of the research questions.

Creation of Synthetic Tests Conforming to the Golden Rule Stipulations

This study required the creation of two synthetic versions of each of the four original standardized tests (Eighth-Grade Mathematics, Eighth-Grade Reading, Basic Skills Test in Mathematics, and Basic Skills Test in Reading) composed only of those items from the original tests which conformed to the Golden Rule stipulations for Type I items. The content validity of the synthetic tests was not

specifically protected through inclusion of Type II items as necessary to ensure content representativeness.

As discussed in the previous chapter of this dissertation, Type I items must conform to two criteria. First, the correct answer rate for blacks, whites, and all examinees must not be less than 40% at the .05 level of statistical significance (Stipulation a). Second, the correct answer rates of blacks and whites must differ by no more than 15% at the .05 level of statistical significance (Stipulation b). Appendix F contains the correct answer rates for black examinees and white examinees on the items of the Eighth-Grade Reading Test, and t-statistics for testing H_O: P=0.40 against H_A: P<0.40. It also contains the differences between proportions correct for black and white examinees on the items of the Eighth-Grade Reading Test, and t-statistics for testing H_O: $\Delta = 0.15$ against the alternative hypothesis: H_A: $\Delta > 0.15$. Appendix G contains correct answer rates for all examinees on items of the Eighth-Grade Reading Test, and t-statistics for testing H_O: P=0.40 against H_A: P<0.40. Appendices H and I contain this information for items of the Eighth-Grade Mathematics Test. Appendices J and K contain the information for items of the Basic Skills Test in Reading, and Appendices L and M contain the information for the Basic Skills Test in Mathematics.

As discussed in the previous chapter, it was decided to treat the correct answer rates as sample statistics, and, in separate analyses, as population parameters. Thus, *two* synthetic versions of each test were created: one composed of items which conformed to the stipulations of Type I items based upon procedures using sample statistics, and another composed of items which conformed to the stipulations of Type I items when the observed correct answer rates for the

sample were treated as population parameters. Synthetic tests composed of items which conform to the Golden Rule stipulations for Type I items, as determined by treating the observed correct answer rates as sample statistics, are designated with a prefix "GRS-". Synthetic tests composed of items which conform to the Golden Rule stipulations for Type I items, as determined when the observed correct answer rates are treated as population parameters, have been designated with a prefix "GRP-". A test item could be classified as Type II for one of three reasons: 1) because it failed to meet Stipulation a for Type I items (i.e., the correct answer rate was less than 0.40 for black examinees, or for white examinees, or for all examinees); 2) because it failed to meet Stipulation b for Type I items (i.e., the difference between the correct answer rates for black and white examinees was more than 0.15); or 3) it could fail to meet both Stipulation a and Stipulation b. For each of the eight synthetic tests created in this study, Table 8, below, contains the number of items in the original test, the number classified as Type I items, the number of items in the original test that failed to meet Stipulation a of the Golden Rule procedures, the number that failed to meet Stipulation b of the Golden Rule procedures, and the number that failed to meet both Stipulation a and Stipulation b.

As inspection of Table 8 and Appendices F through K indicates, substantial numbers of items from the original tests failed to be classified as Type I items. This was particularly true for the mathematics tests. The most common reason for items failing to be classified as Type I was differences larger than 0.15 between the correct answer rates of black and white examinees. Consistent with the findings of Marco (1987; 1988), application of this stipulation led to the rejection (i.e., classification as Type II) of many items which were of mid-level

difficulty for black examinees. For some items, correct answer rates for black examinees were less than 0.40; often, the differences between the correct answer rates of black and white examinees was greater than 0.15 for these items.

Table 8
Classification of Items of Original Standardized Tests

Synthetic Test	# items on original test	# classified as Type I	# failing Stipulation a	# failing Stipulation b	# failing both a & b
GRP- 8th-Gr. Reading	122	93 (76%)	0	28	1
GRP-8th-Gr. Mathematics	127	61 (48%)	5	48	13
GRP-BST Reading	105	73 (70%)	0	30	2
GRP-BST Mathematics	112	53 (47%)	4	40	1
GRS-8th-Gr. Reading	122	108 (89%)	0	13	1
GRS-8th-Gr. Mathematics	127	86 (68%)	5	28	8
GRS-BST Reading	105	90 (86%)	0	13	2
GRS-BST Mathematics	112	71 (63%)	2	29	9

Neither reading test had items with observed correct answer rates for white examinees or for all examinees that were less than 0.40. The correct

answer rate was less than 0.40 for white examinees for one item on the Eighth-Grade Mathematics Test, and for no items on the BST Mathematics test. The correct answer rates for all examinees were less than 0.40 for four items on the Eighth-Grade Mathematics Test, and for one item on the BST Mathematics test. (The reader will note that black and white examinees, taken together, compose approximately 99% of the sample, and the items for which the correct answer rates for all examinees was less than 0.40 also had correct answer rates of less than 0.40 for either black or white subgroups, or for both subgroups.)

Results of Investigation of Research Question 1: Adverse Impact

Results of Investigation of Research Question 1A

Major Research Question 1 addressed the effectiveness of applying the Golden Rule procedures in reducing the adverse impact of test use. Research Question 1A was investigated by comparing the difference between the average total score for black examinees and the average total score for white examinees on each original standardized test with the difference between the average total scores for black examinees and white examinees on corresponding synthetic tests which conformed to the stipulations for Golden Rule Type I items. Separate comparisons were made using synthetic tests in which the correct answer rates were treated as sample statistics, and those in which they were treated as population parameters. Table 9, below, contains the average total scores, by race, for the Eighth-Grade Reading Test and the GRP-Eighth-Grade Reading Test. Tables 10 through 16 contain this information for each remaining comparison of original to standardized test described in Chapter III of this dissertation.

Table 9

Average Total Scores, Percent Correct ( ), Standard Error of the Mean (S.E.M.), Difference Between Average Total Scores and Difference between Average Percents Correct ( ), for Black and White Examinees on the Eighth-Grade Reading Test and the GRP-Eighth-Grade Reading Test

Test	# items on test	White Examinees Total Score T_W	White Examinees S.E.M.	Black Examinees Total Score T_b	Black Examinees S.E.M.	difference $d = (T_W - T_b)$
Eighth-Grade Reading	122	108.322 (88.8%)	0.351	95.468 (78.3%)	0.721	12.854 (10.5%)
GRP-Eighth-Grade Reading	93	84.947 (91.3%)	0.240	77.822 (83.7%)	0.515	7.125 (7.7%)

Table 10

Average Total Scores, Percent Correct ( ), Standard Error of the Mean (S.E.M.), Difference Between Average Total Scores and Difference between Average Percents Correct ( ), for Black and White Examinees on the Eighth-Grade Mathematics Test and the GRP-Eighth-Grade Mathematics Test

Test	# items on test	White Examinees Total Score T_W	White Examinees S.E.M.	Black Examinees Total Score T_b	Black Examinees S.E.M.	difference $d = (T_W - T_b)$
Eighth-Grade Mathematics	127	97.626 (76.9%)	0.471	78.967 (62.2%)	0.803	18.659 (14.7%)
GRP-Eighth-Grade Mathematics	61	50.658 (83.0%)	0.178	45.986 (75.4%)	0.333	4.672 (7.7%)

Table 11

Average Total Scores, Percent Correct ( ), Standard Error of the Mean (S.E.M.), Difference Between Average Total Scores and Difference between Average Percents Correct ( ), for Black and White Examinees on the BST Test in Reading and the GRP-BST Test in Reading

Test	# items on test	White Examinees Total Score T_W	White Examinees S.E.M.	Black Examinees Total Score T_b	Black Examinees S.E.M.	difference $d = (T_W - T_b)$
BST Reading	105	89.998 (85.7%)	0.329	77.495 (73.8%)	0.635	12.503 (11.9%)
GRP-BST Reading	73	64.168 (87.9%)	0.210	57.946 (79.4%)	0.414	6.222 (8.5%)

Table 12

Average Total Scores, Percent Correct ( ), Standard Error of the Mean (S.E.M.), Difference Between Average Total Scores and Difference between Average Percents Correct ( ), for Black and White Examinees on the BST Test in Mathematics and the GRP-BST Test in Mathematics

Test	# items on test	White Examinees Total Score T_W	White Examinees S.E.M.	Black Examinees Total Score T_b	Black Examinees S.E.M.	difference $d = (T_W - T_b)$
BST Mathematics	112	86.574 (77.3%)	0.449	68.876 (61.5%)	0.747	17.689 (15.79%)
GRP-BST Mathematics	53	44.797 (84.0%)	0.240	39.849 (75.2%)	0.323	4.948 (9.3%)

Table 13

Average Total Scores, Percent Correct ( ), Standard Error of the Mean (S.E.M.), Difference Between Average Total Scores and Difference between Average Percents Correct ( ), for Black and White Examinees on the Eighth-Grade Reading Test and the GRS-Eighth-Grade Reading Test

Test	# items on test	White Examinees Total Score T_W	White Examinees S.E.M.	Black Examinees Total Score T_b	Black Examinees S.E.M.	difference $d = (T_W - T_b)$
Eighth-Grade Reading	122	108.322 (88.8%)	0.351	95.468 (78.3%)	0.721	12.854 (10.5%)
GRS-Eighth-Grade Reading	108	96.940 (89.8%)	0.294	87.330 (80.9%)	0.616	9.61 (8.9%)

Table 14

Average Total Scores, Percent Correct ( ), Standard Error of the Mean (S.E.M.), Difference Between Average Total Scores and Difference between Average Percents Correct ( ), for Black and White Examinees on the Eighth-Grade Mathematics Test and the GRS-Eighth-Grade Mathematics Test

Test	# items on test	White Examinees Total Score T_W	White Examinees S.E.M.	Black Examinees Total Score T_b	Black Examinees S.E.M.	difference $d = (T_W - T_b)$
Eighth-Grade Mathematics	127	97.626 (76.9%)	0.471	78.967 (62.2%)	0.803	18.659 (14.7%)
GRS-Eighth-Grade Mathematics	86	69.380 (80.7%)	0.278	60.625 (70.5%)	0.507	8.755 (10.2%)

Table 15

Average Total Scores, Percent Correct ( ), Standard Error of the Mean (S.E.M.), Difference Between Average Total Scores and Difference between Average Percents Correct ( ), for Black and White Examinees on the BST Test in Reading and the GRS-BST Test in Reading

Test	# items on test	White Examinees Total Score T_w	White Examinees S.E.M.	Black Examinees Total Score T_b	Black Examinees S.E.M.	difference $d = (T_w - T_b)$
BST Reading	105	89.998 (85.7%)	0.329	77.495 (73.8%)	0.635	12.503 (11.9%)
GRS-BST Reading	90	78.210 (86.9%)	0.271	69.159 (76.8%)	0.526	9.051 (10.1%)

Table 16

Average Total Scores, Percent Correct ( ), Standard Error of the Mean (S.E.M.), Difference Between Average Total Scores and Difference between Average Percents Correct ( ), for Black and White Examinees on the BST Test in Mathematics and the GRS-BST Test in Mathematics

Test	# items on test	White Examinees Total Score T_w	White Examinees S.E.M.	Black Examinees Total Score T_b	Black Examinees S.E.M.	difference $d = (T_w - T_b)$
BST Mathematics	112	86.574 (77.3%)	0.449	68.876 (61.5%)	0.747	17.689 (15.79%)
GRS-BST Mathematics	71	57.443 (80.9%)	0.258	49.627 (69.9%)	0.447	7.816 (11.0%)

As the results in Tables 9 through 16 illustrate, application of the Golden Rule procedures increased the average proportion of correct answers for both black and white examinees. This was true for both the synthetic tests created by treating the observed item-correct-answer rates as population parameters (the GRP- tests), and the synthetic tests created by treating the observed item-correct-answer rates as sample statistics (the GRS- tests).

To estimate how unlikely it would be to observe a difference between the average total scores for black and white examinees of the magnitude observed on the synthetic GRP-Eighth-Grade Reading Test, were the items included in the synthetic test not systematically different from randomly selected samples of items of the Eighth-Grade Reading Test, the null hypothesis H_O: $\Delta_{8R'} = d_{GRP\text{-}8R}$ was tested against the alternative hypothesis: H_A: $\Delta_{8R'} \neq d_{GRP\text{-}8R}$, where $\Delta_{R'}$ is the difference between black and white examinees' average scores population of tests randomly equivalent to the Eighth-Grade Reading Test, adjusted for test length (by multiplying the observed difference by the ratio of number of items on the synthetic and original tests); and $d_{GRP\text{-}8R}$ is the difference between black and white examinees' average scores on the GRP-Eighth-Grade Reading Test. To test this null hypothesis, the population standard deviation (σ_{dTot}) of the differences between black and white examinees' total scores on random samples of 93 items (the length of the synthetic test) selected from 122 possible items (the length of the original test) was determined*, and a non-directional z test was conducted. (As discussed in Chapter III of this dissertation, subjects were discounted as a source of variance for this analysis. This decision is supported by the small standard errors of the mean total scores on the original and synthetic tests.)

---

* Calculation of σ_{dTot} for each hypothesis test is presented in Appendix N.

Table 17 below presents $\Delta_{8R'}$, d_{GRP-8R}, σ_{dTot}, and the value of the z test for the null hypothesis H_O: $\Delta_{R'} = d_{GRP-8R}$. For each of the other comparisons of original to synthetic test listed in Chapter III, Table 17 also shows the difference between average total scores for black and white examinees on the original test, adjusted for test length; the difference between average total scores for black and white examinees on the synthetic test; the population standard deviation of the difference σ_{dTot}; and the results of the hypothesis test.

As shown in Table 17, the null hypothesis that the difference between the average total scores of black and white examinees observed on the synthetic test was not significantly different from the expected difference between the average total scores of black and white examinees on a randomly equivalent to the original test but of length equal to the synthetic test was rejected for every comparison of synthetic to original test. Using an experiment-wise Type-I error rate of 0.05 for each set of comparisons, one must conclude that, for these tests, application of the Golden Rule procedures does reduce the difference between the average total scores of black and white examinees, both in cases where the observed item-correct-answer rates are treated as population parameters, and when they are treated as sample statistics. This is true despite the fact that application of the Golden Rule procedures raised the average proportion of correct answers for both black and white examinees.

Table 17

Results of Hypotheses Tests for Research Question 1A: Comparing the Differences Between Average Total Scores of Black and White Examinees on Tests Composed of Golden Rule Type I Items and on Tests of Equal Length Composed of Randomly Selected Items from the Original Standardized Tests

Tests Compared (Null Hypothesis)	Original Test: Adjusted Difference	Synthetic Test: Difference	σ_{dTot}	z
Eighth-Grade Reading/ GRP-Eighth-Grade Reading (H_O: $\Delta_{8R'} = d_{GRP-8R}$)	9.799	7.125	0.30293	8.827*
Eighth-Grade Mathematics/ GRP-Eighth-Grade Mathematics (H_O: $\Delta_{8M'} = d_{GRP-8M}$)	8.962	4.672	0.49451	8.676*
BST Reading/ GRP-BST Reading (H_O: $\Delta_{BSTR'} = d_{GRP-BSTR}$)	8.693	6.222	0.30031	8.227*
BST Mathematics/ GRP-BST Mathematics (H_O: $\Delta_{BSTM'} = d_{GRP-BSTM}$)	8.371	4.948	0.40945	8.360*
Eighth-Grade Reading/ GRS-Eighth-Grade Reading (H_O: $\Delta_{8R'} = d_{GRS-8R}$)	11.379	9.61	0.22682	7.799*
Eighth-Grade Mathematics/ GRS-Eighth-Grade Mathematics (H_O: $\Delta_{8M'} = d_{GRS-8M}$)	12.635	8.755	0.46279	8.384*
BST Reading/ GRS-BST Reading (H_O: $\Delta_{BSTR'} = d_{GRS-BSTR}$)	10.717	9.051	0.22830	7.303*
BST Mathematics/ GRS-BST Mathematics (H_O: $\Delta_{BSTM'} = d_{GRS-BSTM}$)	11.214	7.816	0.39506	8.601*

* probability < 0.01

Results of Investigation of Research Question 2: Test Bias

Results of Investigation of Research Question 2A

Research Question 2A addressed the effect of applying the Golden Rule procedures on the overall internal-consistency reliability of tests. As described in the previous chapter of this dissertation, Cronbach's Coefficient Alpha (Cronbach, 1951) was calculated for each of the four original tests, and also for each of the eight synthetic tests composed only of Golden Rule Type I items. Table 18, below, contains the value of Cronbach's Coefficient Alpha for each original and synthetic test.

Shortening a test lowers its reliability, and the synthetic tests are shorter than the original tests. To be able to judge the reliabilities of the synthetic tests, corrected for reduction in length, the Spearman-Brown Prophecy Formula (c.f., Allen & Yen, 1979) was used to project what the reliability of the synthetic tests would be, were they of length equal to the corresponding original tests. These projected reliabilities are shown in parentheses beside the corresponding observed values in Table 18. As inspection of Table 18 shows, the observed values of the Alpha coefficient were lower for the synthetic tests than for the corresponding original tests in every case, both for synthetic tests created by treating observed item-correct-answer rates as sample statistics, and for synthetic tests created by treating observed item-correct-answer rates as population parameters. When the Alpha coefficients for the original tests are compared to the projections of what the Alpha coefficients for the synthetic tests would be, were the synthetic tests as long as the original tests, the Alpha coefficients for the original tests are higher in every case but one. The estimated Alpha coefficient of the GRP-Eighth-Grade Reading Test is slightly larger than that of the original

Eighth-Grade Reading Test: 0.9476 compared to 0.9463. In every other comparison of observed original test to projected synthetic test, values of the Alpha coefficient for the original test are higher than corresponding projections. However, corresponding values are very similar. All the observed values of the Alpha coefficient for the original tests were over 0.93; all the projected values for the synthetic tests were over 0.92.

Table 18

Values of Coefficient Alpha for Original and Synthetic Tests

Test	Coefficient Alpha:		
	Original Test:	GRP-Synthetic Test: Observed (Projected)	GRS-Synthetic Test: Observed (Projected)
Eighth-Grade Reading 0.946297		0.932407 (0.947633)	0.935993 (0.942919)
Eighth-Grade Mathematics	0.947800	0.847716 (0.920570)	0.904717 (0.933430)
BST Reading	0.932924	0.894581 (0.924276)	0.917606 (0.928535)
BST Mathematics	0.948618	0.868040 (0.932889)	0.906174 (0.938405)

Results of Investigation of Research Question 2B

Research Question 2B addresses the effect of applying the Golden Rule procedures on the difference between the reliability of a test for black examinees and for white examinees. As described in the previous chapter of this dissertation, this question was examined by calculating Cronbach's Coefficient Alpha (Cronbach, 1951) separately for black examinees and white examinees, for each original and synthetic test. The Spearman-Brown Prophecy Formula was

used to estimate what the reliabilities of the synthetic tests would be, were they of the same length as the original tests. The difference between the values of the Alpha coefficients for black examinees and white examinees was then determined for each original test. The difference between coefficient Alpha, adjusted for test length, for black examinees and white examinees was determined for each synthetic test. Table 19, below, contains coefficient Alpha for black examinees and white examinees for the Eighth-Grade Reading Test, and the difference between the coefficients for black examinees and white examinees. It also shows coefficient Alpha for black examinees and white examinees for the GRS-Eighth-Grade Reading Test, the projected coefficients for black examinees and white examinees (estimating the coefficients had the synthetic test been of the same length as the original test), and the difference between these projected coefficients for black examinees and white examinees. Tables 20 through 26 contains this information for the remaining comparisons of original to synthetic tests.

Table 19

Values of Coefficient Alpha, by Race, for the Eighth-Grade Reading Test and the GRS-Eighth-Grade Reading Test

	Coefficient Alpha:		
	Eighth-Grade Reading: Observed	GRS-Eighth-Grade Reading: Observed	Projected
White Examinees	0.935667	0.924198	0.932308
Black Examinees	0.938116	0.930068	0.937592
Difference ($\alpha_{white} - \alpha_{black}$)	-0.002449		-0.005284

Table 20

Values of Coefficient Alpha, by Race, for the Eighth-Grade Mathematics Test and the GRS-Eighth-Grade Mathematics Test

	Coefficient Alpha:		
	Eighth-Grade Math.: Observed	GRS-Eighth-Grade Math.: Observed	Projected
White Examinees	0.936763	0.888784	0.921884
Black Examinees	0.932320	0.892786	0.924795
Difference ($\alpha_{white} - \alpha_{black}$)	0.004443		-0.002912

Table 21

Values of Coefficient Alpha, by Race, for the BST Reading Test and the GRS-BST Reading Test

	Coefficient Alpha:		
	BST Reading: Observed	GRS-BST Reading: Observed	Projected
White Examinees	0.920123	0.905454	0.917851
Black Examinees	0.919749	0.905454	0.917851
Difference ($\alpha_{white} - \alpha_{black}$)	0.000374		0.000000

Table 22

Values of Coefficient Alpha, by Race, for the BST Mathematics Test and the GRS-BST Mathematics Test

	Coefficient Alpha:		
	BST Mathematics: Observed	GRS-BST Mathematics: Observed	Projected
White Examinees	0.939466	0.896516	0.931816
Black Examinees	0.929096	0.884534	0.923572
Difference ($\alpha_{white} - \alpha_{black}$)	0.010370		0.008244

Table 23

Values of Coefficient Alpha, by Race, for the Eighth-Grade Reading Test and the GRP-Eighth-Grade Reading Test

	Coefficient Alpha:		
	Eighth-Grade Reading: Observed	GRP-Eighth-Grade Reading: Observed	Projected
White Examinees	0.935667	0.920762	0.938438
Black Examinees	0.938116	0.927545	0.943800
Difference ($\alpha_{white} - \alpha_{black}$)	-0.002449		-0.005361

Table 24

Values of Coefficient Alpha, by Race, for the Eighth-Grade Mathematics Test and the GRP-Eighth-Grade Mathematics Test

	Coefficient Alpha:		
	Eighth-Grade Math.: Observed	GRP-Eighth-Grade Math.: Observed	Projected
White Examinees	0.936763	0.825145	0.907620
Black Examinees	0.932320	0.841951	0.917294
Difference ($\alpha_{white} - \alpha_{black}$)	0.004443		-0.009674

Table 25

Values of Coefficient Alpha, by Race, for the BST Mathematics Test and the GRP-BST Mathematics Test

	Coefficient Alpha:		
	BST Mathematics: Observed	GRP-BST Mathematics: Observed	Projected
White Examinees	0.939466	0.853305	0.924768
Black Examinees	0.929096	0.849189	0.922475
Difference ($\alpha_{white} - \alpha_{black}$)	0.010370		0.002293

Table 26

Values of Coefficient Alpha, by Race, for the BST Reading Test and the GRP-BST Reading Test

	Coefficient Alpha:		
	BST Reading: Observed	GRP-BST Reading: Observed	Projected
White Examinees	0.920123	0.880651	0.913892
Black Examinees	0.919749	0.884719	0.916934
Difference ($\alpha_{white} - \alpha_{black}$)	0.000374		-0.003042

As shown in Tables 19 through 26, the differences between the Alpha coefficients for black examinees and white examinees on the original tests were very small. The Eighth-Grade Reading Test had a slightly higher reliability for black examinees than for white examinees; the remaining three original tests had slightly higher reliabilities for white examinees than for black examinees. The projected Alpha coefficients for the synthetic tests are also very similar for black and white examinees. In three comparisons of original to synthetic test, the difference between the coefficients was greater for the synthetic test; in the other five comparisons, the difference was greater for the original test.

The differences between the projected coefficients for black examinees and white examinees for both the GRS- and GRP- Eighth-Grade Reading Tests were slightly larger than were corresponding differences on the original tests. Both synthetic tests were slightly more reliable for black examinees than for white examinees, as was the original test. The Alpha coefficient for white examinees on the original Eighth-Grade Mathematics Test was slightly higher than for black examinees. For both the GRS- and GRP- Eighth-Grade Mathematics Tests, the

coefficients were slightly higher for black examinees than for white examinees. The difference between the values for black and white examinees was smaller on the GRS-Eighth-Grade Mathematics Test than on the original Eighth-Grade Mathematics Test. The difference between the coefficients for black examinees and white examinees on the GRP-Eighth-Grade Mathematics Test was larger than on the original Eighth-Grade Mathematics Test. The difference between the coefficients for black examinees and white examinees was smaller on both the GRS- and GRP- BST Mathematics Tests than on the original BST Mathematics Test. The difference between the coefficients for black examinees and white examinees was also smaller on both the GRS- and GRP- BST Reading Tests than on the original BST Reading Test. (In fact, the coefficients for black examinees and white examinees were identical on the GRS-BST Reading Test).

Results of Investigation of Research Question 2C

Research Question 2C addressed the effect of applying the Golden Rule procedures, (without specific protection of content representativeness through inclusion of Type II items), on the content representativeness of a test. Appendix O contains the item numbers of items addressing each objective of the original tests; the appendix also contains lists of the items that were excluded from the corresponding synthetic tests. To test the effect of applying the Golden Rule procedures on the overall content of a test, a Chi-square Goodness-of-Fit test was conducted for each synthetic test. The numbers of items that addressed each objective of the synthetic test were treated as the "observed" values in the Chi-square Goodness-of-Fit test. The "expected" values were determined by multiplying the number of items in the synthetic test by the proportion of items that addressed each objective of the original test. The results of the Chi-square

Goodness-of-Fit test for each synthetic test are presented in Appendix P. The null hypothesis that the distribution of items across the test objectives was not different from what would have been expected in a population composed of equal-sized random samples of items drawn from the original test was retained for each of the synthetic tests, with probability ≥ 0.80.

The question of whether the proportion of items that addressed each individual objective of a synthetic test was different from the proportion which would have been expected to address that objective, had samples of items of size equal to the length of the synthetic test been drawn at random from the original test, was investigated by determining the proportion of items that addressed each objective of each original and corresponding synthetic test. The proportions of items that addressed each objective of the original tests were considered to be population parameters. A 95% confidence interval was determined for the proportion of items that addressed each objective of the synthetic tests. If this confidence interval contained the population proportion (i.e., the proportion of items that addressed the objective of the original test) then it was considered tenable that the proportion of items that addressed the same objective of the synthetic test was not significantly different (at the 0.05 level) from what would have been expected among random samples of items drawn from the original test. Appendix Q contains the number and proportion of items that addressed each objective of each of the synthetic tests, and of the corresponding original test. A 95% confidence interval is also shown for the proportion of items on the synthetic tests that addressed each objective. As shown in the appendix, in every case a 95% confidence interval on the proportion of items in the synthetic tests that addressed each objective included the corresponding proportion of

items that addressed that objective of the original test. Thus, one must conclude that the proportion of items that addressed each test objective of the synthetic tests did not differ significantly from the expected proportion of items that addressed each objective of the original tests.

There are several forms of the edition of the Basic Skills Test used in this study, and the numbers of items that address each objective vary from form to form. The range of items typically associated with each objective on the Basic Skills Tests in Mathematics is shown in Table 27, below. To determine whether the number of items that addressed each objective of the synthetic BST tests would fall within the typical range, were the synthetic tests of length equal to the original tests, the proportion of items that addressed each objective of the synthetic tests was multiplied by the number of items in the original test. This resulted in projected numbers of items per objective for the synthetic tests, were they of length equal to the original tests. These projected numbers of items per objective for the synthetic tests, compared to the typical range of items for each objective of the various forms of the BST Mathematics tests, are also presented in Table 27 below. Table 28 contains this information for the Basic Skills Tests in Reading, and the corresponding synthetic tests. As shown in Tables 27 and 28, the numbers of items projected to assess each objective of tests composed solely of Type I items, but of length equal to the original tests, was frequently out of the range of items that addressed the objective across forms of the Basic Skills Tests.

Table 27

Typical Numbers of Items that Assessed Each Objective Across Forms of Georgia Basic Skills Test in Mathematics, and Projected Numbers of Items that Assessed Each Objective of GRS- and GRP- BST Mathematics Tests

Mathematics Objectives	BST Tests Items per Objective	GRS-BST Mathematics Projected Items per Objective	GRP-BST Mathematics Projected Items per Objective
Objective 1	12-14	18.93*	21.17*
Objective 2	4-5	4.70	2.13*
Objective 3	9-11	6.27*	8.40
Objective 4	3-4	3.14	4.26*
Objective 5	4-6	4.70	6.38*
Objective 6	6-7	3.14*	4.26*
Objective 7	18-20	25.20*	23.30*
Objective 8	4-5	6.27*	4.26
Objective 9	8-10	6.27*	4.26*
Objective 10	3-4	4.70*	2.13*
Objective 11	5-6	6.27*	4.26*
Objective 12	7-9	3.14*	4.26*
Objective 13	7-8	7.84	8.40*
Objective 14	12-14	11.09*	14.78*

* Outside range of number of items that assessed objective, across forms of the BST.

Table 28

Typical Numbers of Items that Assessed Each Objective Across Forms of Georgia Basic Skills Test in Reading, and Projected Numbers of Items that Assessed Each Objective of GRS- and GRP- BST Reading Tests

Reading Objectives	BST Tests Items per Objective	GRS-BST Reading Projected Items per Objective	GRP-BST Reading Projected Items per Objective
Objective 1	5-6	3.47*	4.31*
Objective 2	15-20	16.38	18.69
Objective 3	8-10	7.04*	8.61
Objective 4	15-16	16.38*	15.86
Objective 5	7-8	9.36*	8.61*
Objective 6	5-6	5.88	7.14*
Objective 7	6-7	8.19*	8.61*
Objective 8	6-8	7.04	7.14
Objective 9	9-11	11.66*	11.55*
Objective 10	8-10	8.19	7.14*
Objective 11	5-7	7.04*	5.78
Objective 12	6-7	4.62*	1.47*

* Outside range of number of items assessed objective, across forms of the BST.

Results of Investigation of Research Question 2D

Research Question 2D addressed the effect of applying the Golden Rule procedures on the average item difficulty of a test. The difficulty levels (i.e., the correct answer rates across ethnic groups) of each item of the original tests are shown in Appendices G, I, K, and M. Table 29, below, contains the average item difficulty for each synthetic test, and for the corresponding original test. The variance of item difficulties across items on the original test (σ_{id}^2) is also given. As illustrated in Table 29, the average item "difficulty" is higher on each synthetic test than on the corresponding original test. A higher difficulty level (p-value) corresponds to a larger proportion of examinees answering the item correctly.

Table 29

Average Item Difficulty Levels for Original and Synthetic Tests

Test	Average Item Difficulties			
	Original Test: Difficulty	($\sigma_{id}2$)	GRP-Synthetic Test:	GRS-Synthetic Test:
Eighth-Grade Reading	0.85745	0.00931	0.89124	0.87191
Eighth-Grade Mathematics	0.72663	0.02477	0.80843	0.77754
BST Reading	0.82284	0.00892	0.85436	0.83998
BST Mathematics	0.72770	0.02117	0.81834	0.7775

In comparing the average item difficulties of the original and synthetic tests, it was decided to use a measure of the variance of average item difficulties across random samples of items of size equal to the length of the original tests. As discussed in Chapter III of this dissertation, subjects were discounted as a source of variance in these analyses, and the average item difficulties of the original tests were regarded as population parameters. The variance of the average item difficulty on random samples of n items selected from N possible items is given by the formula:

$$\sigma_{\bar{id}^2} = (\sigma^2_{id}/n)\,(1- n/N)$$

where $\sigma_{id}2$ is the variance of the proportion of correct answers across all N items on the original test, as described in Chapter III (Jaeger, 1984, p. 42; Cochran, 1977, p. 23). The items included in the synthetic test forms created to conform to the Golden Rule stipulations were not randomly selected. The above formula was used to estimate how unlikely it would be to observe an average item

difficulty of the magnitude observed on a synthetic test, were the items selected for inclusion in the synthetic test not systematically different from randomly selected samples of items. Since the above formula yields a population variance, a non-directional z-test was performed, testing the null hypothesis H_O: $Diff_{8R} = Diff_{GRP\text{-}8R}$ against the alternative hypothesis: H_A: $Diff_{8R} \neq Diff_{GRP\text{-}8R}$, where $Diff_{8R}$ is the average item difficulty of a population of tests that are randomly equivalent to the original Eighth-Grade Reading Test, but of length equal to the GRS-Eighth-Grade Reading Test, and $Diff_{GRP\text{-}8R}$ is the average item difficulty of the GRS-Eighth-Grade Reading Test. A Type-I error rate of five percent was used for the z-test. Table 30, below, contains the results of this hypothesis test, and parallel hypothesis tests for each comparison of original to synthetic test.

The differences between the average item difficulties of the synthetic tests and the corresponding original tests were relatively small, but were statistically significant (at $\alpha = 0.05$) for every comparison of original to synthetic test, as indicated in Table 30. For the standardized tests and subject sample used in this dissertation, applying the Golden Rule procedures significantly increased the average item difficulty of a test (i.e., made the average proportion of correct answers higher proportion of correct answers higher).

Table 30

Results of Hypotheses Tests for Research Question 2D: Comparing the Average Item Difficulties of Original Tests and Synthetic Tests Composed of Golden Rule Type I Items

Tests Compared (Null Hypothesis)	Average Item Difficulty: Original Test	Synthetic Test	$\sigma_{\bar{id}}$	z
Eighth-Grade Reading/ GRP-Eighth-Grade Reading (H_O: $Diff_{8R} = Diff_{GRP\text{-}8R}$)	0.85745	0.89124	0.004878	-6.927*
Eighth-Grade Mathematics/ GRP-Eighth-Grade Mathematics (H_O: $Diff_{8M} = Diff_{GRP\text{-}8M}$)	0.72663	0.80843	0.014527	-5.631*
BST Reading/ GRP-BST Reading (H_O: $Diff_{BSTR'} = Diff_{GRP\text{-}BSTR}$)	0.82284	0.85436	0.006102	-5.166*
BST Mathematics/ GRP-BST Mathematics (H_O: $Diff_{BSTM'} = Diff_{GRP\text{-}BSTM}$)	0.72770	0.81834	0.014506	-6.248*
Eighth-Grade Reading/ GRS-Eighth-Grade Reading (H_O: $Diff_{8R'} = Diff_{GRS\text{-}8R}$)	0.85745	0.87191	0.003145	-4.598*
Eighth-Grade Mathematics/ GRS-Eighth-Grade Mathematics (H_O: $Diff_{8M'} = Diff_{GRS\text{-}8M}$)	0.72663	0.77754	0.009643	-5.279*
BST Reading/ GRS-BST Reading (H_O: $Diff_{BSTR'} = Diff_{GRS\text{-}BSTR}$)	0.82284	0.83998	0.003763	-4.555*
BST Mathematics/ GRS-BST Mathematics (H_O: $Diff_{BSTM'} = Diff_{GRS\text{-}BSTM}$)	0.72770	0.7775	0.010448	-4.766*

* probability < 0.01

Results of Investigation of Research Question 2E

Research Question 2E addressed the effect of applying the Golden Rule procedures on the average item-total correlation of a test. To address this research question, the correlation of each item with the total test score was determined for items on the original and synthetic tests. A Fisher's Z-transformation, as described by Glass and Hopkins (1984, p. 304-307), was calculated for each item-total correlation. These values of Fisher's Z were then averaged for each test. Each test's average Fisher's Z was converted to a correlation coefficient, which represented the average item-total correlation of the test. Appendix R lists the item-total correlation and corresponding values of Fisher's Z for the items of each original and synthetic test. Table 31, below, contains the average Fisher's Z, and corresponding average item-total correlation for each original and synthetic test. The variance of the Fisher's Z statistics across test items (σ_z^2) is also listed for each original test.

Table 31

Average Item-Total Correlations, and Corresponding Values of Fisher's Z, for Original and Synthetic Tests

Test	Average Item-Total Correlation and Fisher's Z statistic						
	Original Test:			GRP-Synthetic Test:		GRS-Synthetic Test:	
	$\overline{\text{Correlation}}$	$\overline{Z}$	(σ_z^2)	$\overline{\text{Correlation}}$	$\overline{Z}$	$\overline{\text{Correlation}}$	$\overline{Z}$
Eighth-Grade Reading	0.37758	0.39724	0.01113	0.37498	0.39420	0.37196	0.39069
Eighth-Grade Mathematics	0.36367	0.38110	0.01605	0.32614	0.33850	0.33487	0.34830
BST Reading	0.35862	0.37530	0.00529	0.34733	0.36240	0.35109	0.36669
BST Mathematics	0.38362	0.40430	0.01164	0.35408	0.37010	0.36323	0.38060

As shown in Table 31 and Appendix R, the four original tests were composed of items with point-biserial correlations ranging from 0.10 to 0.56. The point-biserial correlations of the items composing the original tests were fairly heterogeneous, and the average point-biserial correlation ranged from 0.35 to 0.39 for the four original tests. The average point biserial correlations of the synthetic tests were somewhat smaller than the values for corresponding original tests in every case.

To compare the average point-biserial correlation of synthetic and corresponding original tests, a measure of the variance of the average Fisher's Z statistic (corresponding to the average item-total correlations) on the synthetic tests was used. This parameter reflected the variation of the average Fisher's Z statistics (and average item-total correlations) across samples of items of size equal to the length of the synthetic tests. As discussed in Chapter III, persons were ignored as a source of variance in these analyses. The variance of the average Fisher's Z statistic for item-total correlations on random samples of n items selected from N possible items is given by the formula:

$$\sigma_{\bar{z}}^2 = (\sigma_z^2/n)\,(1 - n/N)$$

where σ_z^2 is the variance of the Fisher's Z statistics across all N items on the original test, as described in Chapter III (Jaeger, 1984, p. 42; Cochran, 1977, p. 23).

The above formula yields the variance of tests' average Fisher's Z statistics (corresponding to the average item-total correlations) across random samples of items. The items included in the synthetic test forms created to conform to the Golden Rule stipulations were not randomly selected. However, the above formula may be used to estimate how unlikely it would be to observe an average

Fisher's Z statistic of the magnitude observed on a synthetic test, were the items selected for inclusion in the synthetic test not systematically different from randomly selected samples of items. Since the above formula yields a population variance, a non-directional z-test was performed, testing the null hypothesis H_O: $IT_{8R} = IT_{GRS\text{-}8R}$ against the alternative hypothesis: H_A: $IT_{8R} \neq IT_{GRS\text{-}8R}$, where IT_{8R} is the Fisher's Z statistic corresponding to the average item-total correlation of a population of tests that are randomly equivalent to the original Eighth-Grade Reading Test, but of length equal to the GRS-Eighth-Grade Reading Test; and $IT_{GRS\text{-}8R}$ is the Fisher's Z statistic corresponding to the average item-total correlation of the GRS-Eighth-Grade Reading Test. A Type-I error rate of five percent was used for the z-test.

The procedures described above for comparing the average item-total correlations of the original Eighth-Grade Reading Test and the GRS-Eighth-Grade Reading Test were repeated for each of the other comparisons of original to synthetic test. The results of these hypotheses tests are shown in Table 32, below.

Table 32
Results of Hypotheses Tests for Research Question 2E: Comparing the Average Point-Biserial Correlation of Original Tests and Synthetic Tests Composed of Golden Rule Type I Items

Tests Compared (Null Hypothesis)	Fisher Z Corresponding to Average Point-Biserial Correlation: Original Test	Synthetic Test	$\sigma_{\bar{z}}$	z
Eighth-Grade Reading/ GRP-Eighth-Grade Reading (H_O: $IT_{8R} = IT_{GRP\text{-}8R}$)	0.39724	0.39420	0.00533	0.570
Eighth-Grade Mathematics/ GRP-Eighth-Grade Mathematics (H_O: $IT_{8M} = IT_{GRP\text{-}8M}$)	0.38110	0.33850	0.01169	3.643*
BST Reading/ GRP-BST Reading (H_O: $IT_{BSTR'} = IT_{GRP\text{-}BSTR}$)	0.37530	0.36240	0.00470	2.745*
BST Mathematics/ GRP-BST Mathematics (H_O: $IT_{BSTM'} = IT_{GRP\text{-}BSTM}$)	0.40430	0.37010	0.01076	3.180*
Eighth-Grade Reading/ GRS-Eighth-Grade Reading (H_O: $IT_{8R'} = IT_{GRS\text{-}8R}$)	0.39724	0.39069	0.00344	1.905
Eighth-Grade Mathematics/ GRS-Eighth-Grade Mathematics (H_O: $IT_{8M'} = IT_{GRS\text{-}8M}$)	0.38110	0.34830	0.00790	4.226*
BST Reading/ GRS-BST Reading (H_O: $IT_{BSTR'} = IT_{GRS\text{-}BSTR}$)	0.37530	0.36669	0.00290	2.983*
BST Mathematics/ GRS-BST Mathematics (H_O: $IT_{BSTM'} = IT_{GRS\text{-}BSTM}$)	0.40430	0.38060	0.00775	3.059*

* probability < 0.01

Although the differences between the average point-biserial correlations of the original tests and corresponding synthetic tests were relatively small, they were statistically significant for all but one original test. As shown in Table 32, the average item-total correlation was significantly higher for both original mathematics tests than for the corresponding synthetic tests. The average item-total correlation was higher for the original BST Reading test than for the corresponding synthetic tests. These findings were consistent for synthetic tests formed by treating observed correct answer rates as sample statistics, and for those formed by treating observed correct answer rates as population parameters. The average item-total correlation of the Eighth-Grade Reading Test was not significantly different from that of either the GRS- or the GRP-Eighth-Grade Reading Test. (The reader will recall that, of the original tests, the Eighth-Grade Reading Test had the smallest proportion of Type II items, both when the observed correct answer rates were treated as sample statistics, and when they were treated as population parameters. Applying the Golden Rule procedures had less impact on the item composition of this test than on that of the other three original tests).

Results of Investigation of Research Question 2F

Research Question 2F addressed the effect of applying the Golden Rule procedures on the overall predictive validity of a test. In examining this research question, results of the original and synthetic Eighth-Grade tests were used to predict total scores on the unaltered Basic Skills Tests. Shortening a test lowers its reliability, which attenuates its predictive validity. The synthetic tests are shorter than the corresponding original tests, and, as described in Chapter III, the Spearman-Brown Prophecy Formula was used to project what the reliabilities of

the synthetic tests would have been, had the synthetic tests been of length equal to corresponding original tests. These projected reliabilities were listed in Table 18, above. A variant of the formula for correction for attenuation (Allen & Yen, 1979) was used to estimate what the coefficient of prediction for each synthetic test would have been, had the synthetic test had a reliability coefficient equal to that estimated by the Spearman-Brown Prophecy Formula. Table 33, below, contains the correlations of the original and synthetic Eighth-Grade Reading Tests with the original BST in Reading. Table 34 contains the correlations of the original and synthetic Eighth-Grade Mathematics Tests with the original BST in Mathematics. There tables also contain projected coefficients of prediction, corrected for differences in reliability due to length of test. A 95% confidence interval is also shown for each correlation coefficient involving an original Eighth-Grade test, and for each projected correlation coefficient involving a synthetic test.

Table 33

Correlations between Original and Synthetic Eighth-Grade Reading Tests and the BST in Reading

	Correlation with BST Reading:			
	Observed	(95% CI)	Projected	(95% CI)
Eighth-Grade Reading	0.782872	(0.764, 0.800)	N/A	
GRP-Eighth-Grade Reading	0.730668		0.736610	(0.715, 0.757)
GRS-Eighth-Grade Reading	0.758489		0.761290	(0.741, 0.780)

Table 34

Correlations between Original and Synthetic Eighth-Grade Mathematics Tests and the BST in Mathematics

	Correlation with BST Mathematics:			
	Observed	(95% CI)	Projected	(95% CI)
Eighth-Grade Mathematics	0.865071	(0.853, 0.876)	N/A	
GRP-Eighth-Grade Mathematics	0.749244		0.780776	(0.762, 0.798)
GRS-Eighth-Grade Mathematics	0.808032		0.820754	(0.805, 0.835)

As shown in Tables 33 and 34, coefficients of prediction are higher for both the original Eighth-Grade Reading and Mathematics Tests than for the corresponding synthetic tests, even when the coefficients of prediction for the synthetic tests have been adjusted to account for differences in reliability due to differences in test length between the original and synthetic tests.

Since both the original and synthetic Eighth-Grade test results were used to predict original BST test results, the difference between the coefficients of prediction for the Eighth-Grade Reading Test and the GRS-Eighth-Grade Reading Test was tested for statistical significance, using procedures for testing dependent correlations coefficients. These procedures are described in Chapter III of this dissertation. Corresponding differences between the predictive validities of the Eighth-Grade Reading Test and the GRP-Eighth-Grade Reading Test, the Eighth-Grade Mathematics Test and the GRS-Eighth-Grade Mathematics Test, and the Eighth-Grade Mathematics Test and the GRP-Eighth-Grade Mathematics Test were also tested. These tests were performed by

calculating t-statistics for null hypotheses of the form: H_O: $\rho_{8,BST} = \rho_{GR8',BST}$ against alternative hypotheses of the form: H_A: $\rho_{8,BST} \neq \rho_{GR8,BST}$, where $\rho_{8,BST}$ is the correlation between the population of tests that are randomly equivalent to the original eighth-grade test and the corresponding BST test, and $\rho_{GR8,'BST}$ is the correlation between the synthetic eighth-grade test and the corresponding BST test, adjusted for differences in reliability due to lengths of tests. Table 35, below, contains the results of these hypotheses tests. (Testing the difference between dependent correlation coefficients also requires knowledge of the intercorrelation between the all the variables. Correlations between corresponding original and synthetic Eighth-Grade test results are also shown in Table 35, in the column labeled: 'Correlation: Original & Synthetic.' These correlation coefficients have also been adjusted to correct for differences in reliability due to differences in the lengths of corresponding synthetic and original tests.)

As inspection of Table 35 shows, the predictive validity of the original Eighth-Grade Reading Test was significantly higher than the adjusted predictive validities of the GRS- or GRP-Eighth-Grade Reading Tests. The predictive validity of the original Eighth-Grade Mathematics Test was also significantly higher than the adjusted predictive validities of the GRS- or GRP-Eighth-Grade Mathematics Tests. (The coefficients of prediction of the synthetic tests were adjusted upward to account for differences in predictive validity attributable to differences in reliability due to differences in the lengths of corresponding original and synthetic tests). (The reader should bear in mind that, in these analyses, subjects were the source of variance used in the hypotheses tests, rather than items, as in several earlier hypotheses tests in this

dissertation study). The differences between the predictive validities of the original and synthetic tests are large enough to be of potential practical, as well as statistical, significance.

Table 35

Results of Hypotheses Tests for Research Question 2F: Comparing the Correlations between Original Eighth-Grade Test Results and Synthetic Eighth-Grade Test Results with Original BST Test Results

Predictor Tests Compared (Null Hypothesis)	Correlation with BST Results: Original Test Observed (*O*)	Synthetic Test Projected (*S*)	Diff. (*O* - *S*)	Correlation: Original & Synthetic	t
Eighth-Grade Reading/ GRP-Eighth-Grade Reading (H_O: $\rho_{8R,BST} = \rho_{GRP\text{-}8R',BST}$)	0.782872	0.736610	0.046262	0.984986	19.245*
Eighth-Grade Mathematics/ GRP-Eighth-Grade Mathematics (H_O: $\rho_{8M,BST} = \rho_{GRP\text{-}8M',BST}$)	0.865071	0.780776	0.084295	0.954210	24.705*
Eighth-Grade Reading/ GRS-Eighth-Grade Reading (H_O: $\rho_{8R,BST} = \rho_{GRP\text{-}8R',BST}$)	0.782872	0.761290	0.021582	0.995544	16.406*
Eighth-Grade Mathematics/ GRS-Eighth-Grade Mathematics (H_O: $\rho_{8M,BST} = \rho_{GRS\text{-}8M',BST}$)	0.865071	0.820754	0.044317	0.981028	20.108*

* probability < 0.01

Results of Investigation of Research Question 2G

Research Question 2G addressed the effect of applying the Golden Rule procedures on the difference between a test's coefficients of prediction for black examinees and for white examinees. In examining this research question, eighth-grade test results were again used to predict total scores on the unaltered Basic Skills Tests. To investigate this research question, the correlations between original and synthetic eighth-grade test results and the BST test results were determined separately for black examinees and white examinees. Shortening a test lowers its reliability, which attenuates its predictive validity. The synthetic tests are shorter than the corresponding original tests, and, as described in Chapter III, the Spearman-Brown Prophecy Formula was used to project what the reliabilities of the synthetic tests would have been for black examinees and for white examinees, had the synthetic tests been of length equal to the original tests. These projected reliabilities were listed in Tables 19 through 26, above. A variant of the formula for correction for attenuation (Allen & Yen, 1979) was used to estimate what the coefficients of prediction for each synthetic test would have been, had the synthetic test had reliability coefficients (for each ethnic group) equal to those estimated by the Spearman-Brown Prophecy Formula. The differences between the coefficient of prediction for black examinees and the coefficient for white examinees was then determined for the original eighth-grade tests. The differences between the coefficients of prediction for black examinees and for white examinees, adjusted for differences in reliability due to differences in test lengths, were also determined for the synthetic eighth-grade tests. Table 36, below, contains coefficients of prediction for black examinees and white examinees for each original test, and adjusted coefficients of prediction for

black examinees and white examinees for each synthetic test. Differences between corresponding coefficients for black examinees and white examinees are also shown.

As shown in Table 36, application of the Golden Rule procedures to the eighth-grade tests examined in this study lowered the predictive validities of the tests for white examinees in all cases, even after the predictive validities of the synthetic tests were adjusted upward to compensate for differences in validity attributable to differences in reliability due to differences in test lengths. Application of the Golden Rule procedures to the eighth-grade tests examined in this study also lowered the predictive validities of the tests for black examinees in all cases, even after the predictive validities of the synthetic tests were adjusted upward to compensate for differences in validity attributable to differences in reliability due to differences in test lengths. However, application of the Golden Rule procedures had a smaller effect on coefficients of prediction for black examinees than it did on coefficients for white examinees.

The difference between the predictive validities for black examinees and white examinees was relatively small for both the original and synthetic eighth-grade tests, as shown in Table 36. The coefficient of prediction for black examinees was somewhat higher than that for white examinees on the original Eighth-Grade Reading Test. The coefficients of prediction for the GRP- and GRS-Eighth-Grade Reading Tests were also higher for black examinees than for white examinees. The difference between the coefficients of prediction for black examinees and white examinees was greater on both of the synthetic tests than on the original Eighth-Grade Reading Test. The coefficient of prediction for white examinees was somewhat higher than for black examinees on the original

Table 36

Correlations of Original Eighth-Grade Test Results and Synthetic Eighth-Grade Test Results with Original BST Test Results, by Race

Predictor:	Correlation with BST Results: White Examinees Observed r_W (95%CI)	White Examinees Projected r_W (95%CI)	Black Examinees Observed r_b (95%CI)	Black Examinees Projected ρ_b (95%CI)	difference $d = (r_W - r_b)$
Eighth-Grade Reading	0.721162 (0.693, 0.747)	N/A	0.772777 (0.736, 0.805)	N/A	-0.051615
GRP-Eighth-Grade Reading	0.663115	0.669450 (0.638, 0.699)	0.739390	0.745840 (0.705, 0.782)	-0.076390
GRS-Eighth-Grade Reading	0.696088	0.708931 (0.680, 0.735)	0.754426	0.767831 (0.730, 0.801)	-0.05890
Eighth-Grade Mathematics	0.844322 (0.828, 0.859)	N/A	0.808025 (0.776, 0.836)	N/A	0.036297
GRP-Eighth-Grade Mathematics	0.724059	0.759149 (0.735, 0.782)	0.710814	0.741937 (0.701, 0.778)	0.017212
GRS-Eighth-Grade Mathematics	0.786869	0.790314 (0.769, 0.810)	0.754813	0.757860 (0.719, 0.792)	0.032454

Eighth-Grade Mathematics Test. The coefficients of prediction for both the GRP- and GRS- Eighth-Grade Mathematics Tests were also higher for white examinees than for black examinees. The difference between the coefficients of prediction for black examinees and white examinees was smaller for both of the synthetic tests than for the original Eighth-Grade Mathematics Test.

The null hypothesis that the predictive validities are the same for black examinees and white examinees (H_O: $\rho_W = \rho_B$) was tested for each original and synthetic eighth-grade test. The projected values were used for synthetic tests, as described above. (Subjects were the source of variance in these tests, rather than

items, as in some other hypotheses tests conducted in this dissertation.) For each eighth-grade test, Table 37, below, contains the null hypothesis to be tested, the correlation for white examinees and the corresponding Fisher's Z, the correlation for black examinees and the corresponding Fisher's Z, and the value of z obtained for the hypothesis test. Since there were 1269 white examinees and 517 black examinees in this study, the standard error of the difference between Fisher's Z statistics, $\sigma_{Z1-Z2} = 0.0523$ for each of the hypotheses tests summarized in Table 37.

As shown in Table 37, the difference between the coefficients of prediction for black examinees and white examinees was statistically significant (at $\alpha = 0.05$) for the original and synthetic eighth-grade reading tests; coefficients were consistently larger for black examinees. The difference between the coefficients of prediction for black examinees and white examinees was statistically significant for the original Eighth-Grade Mathematics Test, with white examinees having the larger value. The differences between corresponding coefficients of prediction for black examinees and white examinees were not statistically significant for the synthetic eighth-grade mathematics tests. Application of the Golden Rule procedures produced inconsistent results, in that the Golden Rule procedures did not eliminate a significant difference between coefficients of prediction for black examinees and white examinees on the Eighth-Grade Reading test, (in fact, the procedures exacerbated the difference); but the procedures did eliminate a significant difference between the coefficients of prediction for black examinees and white examinees on the Eighth-Grade Mathematics test.

Table 37

Results of Hypotheses Tests for Research Question 2G: Comparison of Correlations of Eighth-Grade Test Results and BST Test Results for Black Examinees and White Examinees

Predictor: (Null Hypothesis)	White Examinees Correlation r_W	White Examinees Fisher's Z	Black Examinees Correlation r_b	Black Examinees Fisher's Z	z
Eighth-Grade Reading (H_O: $\rho_W = \rho_B$)	0.721162	0.910061	0.772777	1.027185	-2.239*
GRP-Eighth-Grade Reading (H_O: $\rho_W = \rho_B$)	0.669450	0.809746	0.745840	0.963423	-2.938**
GRS-Eighth-Grade Reading (H_O: $\rho_W = \rho_B$)	0.708931	0.88503	0.767831	1.01502	-2.485*
Eighth-Grade Mathematics (H_O: $\rho_W = \rho_B$)	0.844322	1.23604	0.808025	1.121312	2.194*
GRP-Eighth-Grade Mathematics (H_O: $\rho_W = \rho_B$)	0.759149	0.994203	0.741937	9.954775	0.754
GRS-Eighth-Grade Mathematics (H_O: $\rho_W = \rho_B$)	0.786869	1.06312	0.757860	0.991168	1.376

* probability < 0.05
** probability < 0.01

Results of Investigation of Research Question 2H

Research Question 2H addressed the effect of applying the Golden Rule procedures on the difference between the regression equations relating a test to its criterion for black examinees and white examinees. Again, eighth-grade test results were used to predict total scores on the unaltered Basic Skills Tests. Regression equations predicting BST scores from original and synthetic eighth-grade test scores were estimated separately for black examinees and for white examinees. A 95% confidence interval was calculated for the slope and intercept of each regression equation. Table 38, below, contains these estimates, by race, for regressions, of original and synthetic eighth-grade reading test results on results of the original BST Reading Test. Table 39 contains corresponding estimates for the mathematics tests.

The regressions of the BST test results on the results of each original and synthetic eight-grade test are shown graphically below. Figures 1 through 3 depict the regression of the results of the BST Reading Test on results of the Eighth-Grade Reading Test, the GRP-Eighth-Grade Reading Test, and the GRS-Eighth-Grade Reading Test, respectively.

Table 38

Estimated Parameters of Regressions of Original BST Reading Test Results on Original and Synthetic Eighth-Grade Reading Test Results, by Race

Predictor:	Prediction of BST Reading Results: White Examinees Intercept (95%CI)	White Examinees Slope (95%CI)	Black Examinees Intercept (95%CI)	Black Examinees Slope (95%CI)
Eighth-Grade Reading	16.8865 (12.993, 20.778)	0.6749 (0.639, 0.711)	12.4664 (7.786, 17.146)	0.6812 (0.632, 0.729)
GRP-Eighth-Grade Reading	13.0045 (8.195, 17.814)	0.9064 (0.850. 0.963)	6.4408 (0.790, 12.09)	0.9130 (0.841, 0.985)
GRS-Eighth-Grade Reading	14.442 (10.126, 18.758)	0.7790 (0.735, 0.824)	9.4988 (4.324, 14.673)	0.7786 (0.720, 0.837)

Table 39

Estimated Parameters of Regressions of Original BST Mathematics Test Results on Original and Synthetic Eighth-Grade Mathematics Test Results, by Race

Predictor:	Prediction of BST Mathematics Results: White Examinees Intercept (95%CI)	White Examinees Slope (95%CI)	Black Examinees Intercept (95%CI)	Black Examinees Slope (95%CI)
Eighth-Grade Mathematics	7.9720 (5.187, 10.757)	0.8051 (0.777, 0.833)	9.5151 (5.676, 13.350)	0.7517 (0.704, 0.799)
GRP-Eighth-Grade Mathematics	-5.8595 (-10.746, -0.973)	1.8246 (1.729, 1.920)	-4.3192 (-10.660, 2.021)	1.5917 (1.456, 1.728)
GRS-Eighth-Grade Mathematics	-1.7159 (-3.602, 0.1705)	1.2726 (1.218, 1.327)	1.4057 (-3.749, 6.560)	1.1129 (1.029, 1.196)

Figure 1
Regression of BST Reading on Gr. 8 Reading

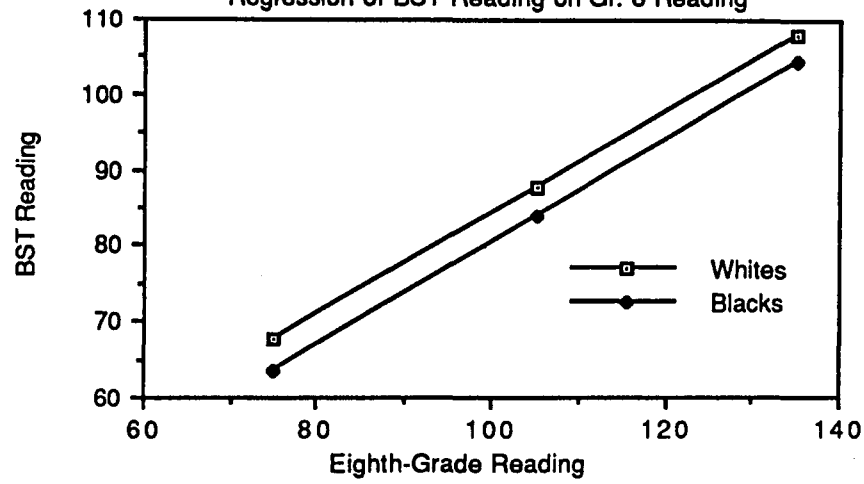


Figure 2
Regression of BST Reading on GRP-Gr. 8 Reading

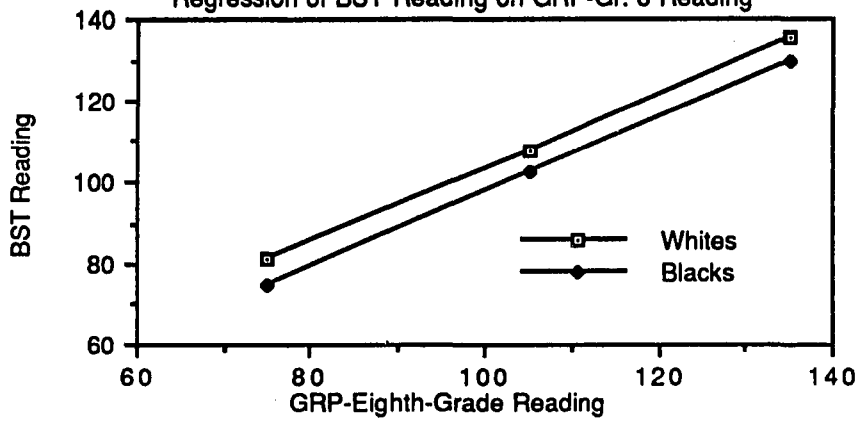
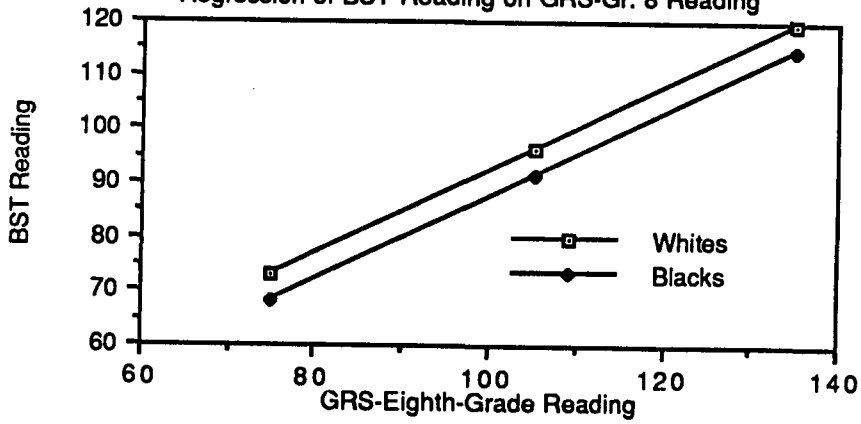


Figure 3
Regression of BST Reading on GRS-Gr. 8 Reading



As shown in Table 38 and Figure 1, when BST Reading Test results are regressed on Eighth-Grade Reading Test results, the slopes for black examinees and white examinees are very similar. The 95% confidence interval for the slope for white examinees contains the value of the slope for black examinees, thus one can conclude that the difference between the slopes is not statistically significant (at $\alpha = 0.05$) for the Eighth-Grade Reading Test. The slopes for black examinees and white examinees are also very similar when results of the GRP- or GRS-Eighth-Grade Reading Test are used to predict BST Reading Test results. Again, the 95% confidence interval on the slope for one group contains the value of the slope for the other group, and one can conclude that the differences between the slopes for black examinees and white examinees are not statistically significant at $\alpha = 0.05$.

For the Eighth-Grade Reading Test, the intercept for black examinees was lower than that for white examinees, and the slopes for black examinees and white examinees were essentially identical, indicating that use of single regression line to predict students' performances on the tenth-grade tests from their performances on the eighth-grade tests would consistently overestimate tenth-grade reading performance for black examinees, and underestimate tenth-grade reading performance for white examinees. Application of the Golden Rule procedures to the Eighth-Grade Reading Test increased, rather than decreased, the difference between the intercepts for black examinees and white examinees, and thus increased the degree of misestimation of tenth-grade reading performance, were a single regression line to be used For the Eighth-Grade Mathematics Test, the slopes of the regression lines were not parallel, and the

intercept was higher for black examinees than for white examinees, indicating that, at some score levels, use of a single regression line would overestimate the tenth-grade performance one racial group, and at other score levels, underestimate the tenth-grade performance of that group. In the range in which most eighth-grade mathematics test scores occurred, use of a single regression line tended to overestimate the tenth-grade performance of black examinees and underestimate that of white examinees. Application of the Golden Rule procedures to the Eighth-Grade Mathematics Test increased the distance between the regression lines for black examinees and white examinees over the range where most Eighth-Grade Mathematics Test scores occur, and thus increased the degree of misestimation of tenth-grade mathematics performances, as well.

Figures 4 through 6, below, depict the regression of results of the BST Mathematics Test on results of the Eighth-Grade Mathematics Test, the GRP-Eighth-Grade Mathematics Test, and the GRS-Eighth-Grade Mathematics Test. As shown in Table 39 and Figure 4, when BST Mathematics Test results are regressed on Eighth-Grade Mathematics Test results, the slopes for black examinees and white examinees are dissimilar. The 95% confidence interval on the slope for one group does not contain the value of the slope for the other group. The slopes associated with regression of the results of the GRP- and GRS-Eighth-Grade Mathematics Tests are also dissimilar. Since the confidence intervals for the groups do not overlap, one may conclude that the slopes of corresponding regressions for black examinees and white examinees are significantly different (at $\alpha = 0.05$), for the original and synthetic eighth-grade mathematics tests. The difference between the slopes for black examinees and white examinees associated with the Eighth-Grade Mathematics Test is

Figure 4
Regression of BST Mathematics on Gr. 8 Mathematics

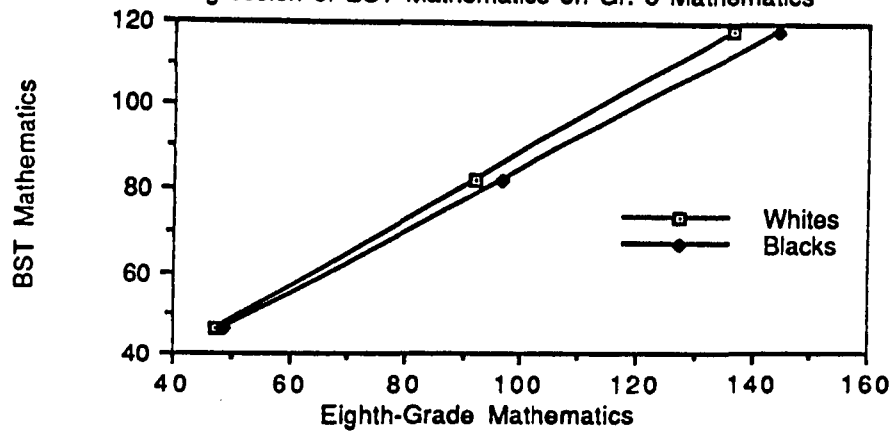


Figure 5
Regression of BST Mathematics on GRP-Gr. 8 Mathematics

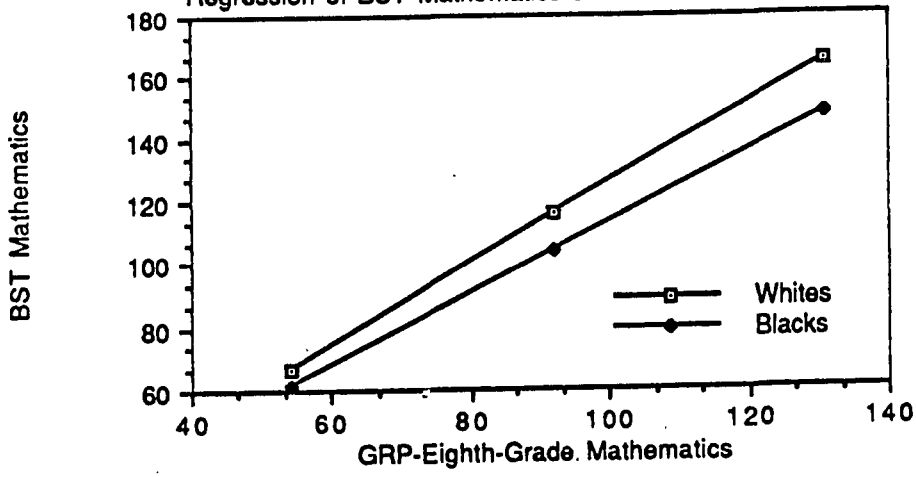
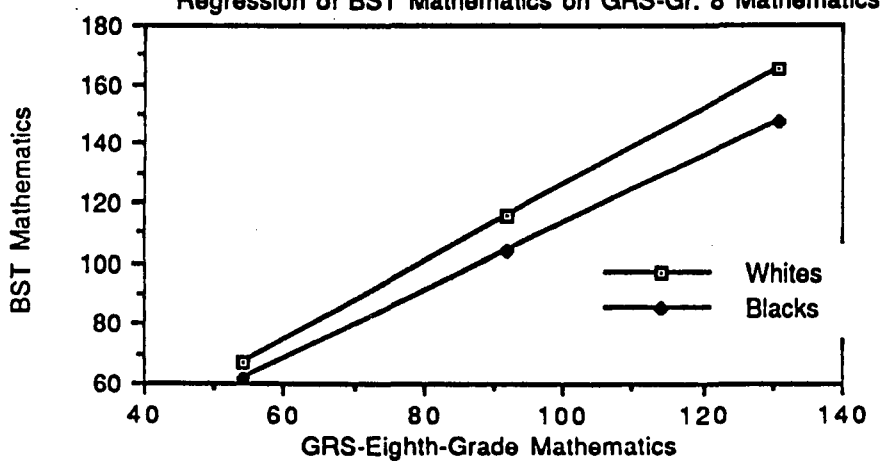


Figure 6
Regression of BST Mathematics on GRS-Gr. 8 Mathematics



approximately 0.05. The difference between the slopes for black examinees and white examinees associated with the GRP-Eighth-Grade Mathematics Test is approximately 0.20. For the GRS-Eighth-Grade Mathematics Test, the difference between the slopes is about 0.16. Thus, application of the Golden Rule procedures to the Eighth-Grade Mathematics Test exacerbated, rather than decreased, dissimilarity of corresponding regression equations for black examinees and white examinees.

The degree of differential regression for the original and synthetic eighth-grade tests was also evaluated by fitting a regression model in which ethnic-group membership was included as a predictor. (Only black and white ethnic groups were included in these analyses). This allowed evaluation of the interaction of ethnic group membership and eighth-grade test performance in the prediction of BST test performance. Table 40, below, contains estimates of

Table 40

Estimates of Parameters of Regressions of Original BST Reading Test Results on Original and Synthetic Eighth-Grade Reading Test Results and Ethnic Group Membership

Eighth-Grade Test:	Prediction of BST Reading Results: Intercept (95%CI)	Slope of Test (95%CI)	Coeff. of Ethnic (95%CI)	Coeff. of Ethnic*Test (95%CI)
Eighth-Grade Reading	16.8865 (12.84, 20.93)	0.6749 (0.638, 0.712)	-4.4201 (-10.325, 1.485)	0.0062 (-0.052, 0.064)
GRP-Eighth-Grade Reading	13.0045 (8.04, 17.97)	0.9064 (0.848, 0.965)	-6.5637 (-13.796, 0.669)	0.0067 (-0.082, 0.095)
GRS-Eighth-Grade Reading	14.4422 (9.96, 18.93)	0.7794 (0.733, 0.825)	-4.9435 (-11.484, 1.597)	-0.0008 (-0.0716, 0.0700)

parameters for regressions of BST Reading performance on original and synthetic eighth-grade reading performances and ethnic group membership. Table 41 contains corresponding information for the mathematics tests.

Table 41

Estimates of Parameters of Regressions of Original BST Mathematics Test Results on Original and Synthetic Eighth-Grade Mathematics Test Results and Ethnic Group Membership

Eighth-Grade Test:	Prediction of BST Mathematics Results: Intercept (95%CI)	Slope of Test (95%CI)	Coeff. of Ethnic (95%CI)	Coeff. of Ethnic*Test (95%CI)
Eighth-Grade Mathematics	7.9720 (5.04, 10.90)	0.80512 (0.776, 0.835)	1.5410 (-2.989, 6.071)	-0.0534 (-0.104, -0.003)
GRP-Eighth-Grade Mathematics	-5.8595 (-10.87, -0.854)	1.8246 (1.727, 1.923)	1.5403 (-6.275, 9.355)	-0.2330 (-0.396, -0.070)
GRS-Eighth-Grade Mathematics	-1.7159 (-5.71, 2.28)	1.2726 (1.216, 1.329)	3.1216 (-3.085, 9.328)	-0.1596 (-0.255, -0.0638)

As shown in Table 40, the coefficient of the interaction of ethnic group and eighth-grade test score was not significant when Eighth-Grade Reading Test results and ethnic-group membership were used to predict BST Reading Test results. The probability that the parameter corresponding to the coefficient of the interaction term was equal to zero was $p = 0.8335$. Consistent with the conclusions drawn above, the interaction of ethnic-group membership and eighth-grade test score was also not significant when GRP- and GRS- Eighth-Grade Reading Test results and ethnic-group membership were used to predict BST Reading Test results. The probabilities that the parameters corresponding to

the coefficients of the interaction terms were equal to zero were p = 0.8828 and p = 0.9823, respectively, for the two synthetic tests.

The coefficient of the interaction of ethnic-group membership and eighth-grade test score was significant ($\alpha = 0.05$) when Eighth-Grade Mathematics Test results and ethnic-group membership were used to predict BST Mathematics Test results, as shown in Table 41. The probability that the parameter corresponding to the coefficient of the interaction term was equal to zero was p = 0.0439. Consistent with the conclusions drawn above, the interaction of ethnic-group membership and eighth-grade test score was also significant ($\alpha = 0.05$) when GRP- and GRS- Eighth-Grade Mathematics Test results were used, with ethnic-group membership, to predict BST Mathematics Test results. The probabilities that the parameters corresponding to the coefficients of the interaction terms were equal to zero were p = 0.0048 and p = 0.0011, respectively, for the two synthetic tests. Applying the Golden Rule procedures to the Eighth-Grade Mathematics Test decreased, rather than increased, the similarity of the regression equations for black examinees and white examinees.

Results of Investigation of Research Question 2I

Research Question 2I addresses the effect of applying the Golden Rule procedures on the content and format of the items that compose each test. (Examination of the content representativeness of tests composed only of Type I items was discussed above, in the section addressing Research Question 2C). As described in Chapter III, a matrix of item content and item format was created for each test, and each test item was assigned to a cell of the matrix. The numbers of items in each cell, the numbers of these items that were excluded from both the GRS- and GRP- synthetic tests, and the numbers of these items that were

excluded only from the GRP- synthetic tests were then determined for each original test.

A six-by-five matrix of item content by format was used for the reading tests. The six categories of item content were: 1) literal comprehension (i.e., items that assessed whether the reader understood the literal meaning of something read), 2) deduction of main ideas, 3) inference of principles or future actions, 4) understanding figures of speech (e.g., interpretation of homilies, metaphors and similes), 5) understanding of language structure and rules, and 6) distinguishing fact from opinion. The five categories of item format were: 1) long passages followed by questions (i.e., passages with more than three sentences, often of several paragraphs in length), 2) short passages followed by questions (i.e., passages with three sentences or less), 3) poems, 4) charts or forms followed by questions, and 5) short answer questions. A seven-by-four matrix of content by format was used for the mathematics tests. The seven categories of item content were: 1) units of measurement, 2) problems involving money, 3) problems involving geometry or area, 4) problems involving fractions (common and decimal), 5) knowledge or understanding of basic facts or principles, 6) ability to perform basic arithmetic operations, and 7) problems involving statistics or probability. The four categories of item format were: 1) stories (i.e., brief narratives) followed by questions, 2) figures (e.g., charts, pictures, or graphs) followed by questions, 3) short-answer problems, and 4) math problems (i.e., equations to be solved or arithmetic to be performed, presented directly). As stated in Chapter III, the construct validity of these categories was not examined, nor was the reliability of the assignment of items to categories.

Matrices containing the numbers of items in each cell, the numbers of items excluded from both the GRS- and GRP- synthetic tests, and the numbers excluded from the GRP- synthetic tests for each original test appear in Appendix S. Inspection of Appendix S reveals no obvious patterns of item-format or content categories having disproportionate numbers of Type II items. One might presuppose that black students would have more difficulty than white students understanding items dealing with figures of speech, since these items represented phrases from the dominant culture. On the BST Reading Tests, this content category did have the largest proportion of Type II items of any content category (46%); however, on the Eighth-Grade Reading Test, this content category had a proportion of Type II items that was quite comparable to those of the other categories (21%). The content category of "deducing main ideas" had no Type II items on the BST Reading Test, but had 29% Type II items on the Eighth-Grade Reading Test. On both reading tests, about 20% to 35% of items in a given content category, and about 20% to 25% of items in a given format category were typically classified as Type II, on either the GRP- or both the GRP- and GRS- tests. Larger numbers of the items of the mathematics tests were classified as Type II, so the proportions of Type II items in each content and format category are larger, but, as with the reading tests, there was no obvious pattern of any one category having disproportionate numbers of Type II items. Since the numbers of items in many cells of the matrices for all of the tests were very small, inclusion or exclusion of a single item would have changed the proportions dramatically. Because of the empty cells in each matrix, it was not feasible to conduct Chi-square tests to determine whether application of the Golden Rule procedures would significantly alter the content and/or format of

the tests, as presented in the matrices in Appendix S. However, "ocular examination" does not suggest that the distributions of Type II items across the cells of the matrices are different from those one would expect due to chance, particularly when comparisons are made between comparable cells in the eighth- and tenth-grade tests of the same subject.

Results of Investigation of Research Question 2J

Research Question 2J addresses the effect of applying the Golden Rule procedures to both a test and the criterion it is intended to predict (where the criterion is also a test) on differences between the coefficients of prediction and the regression equations for black examinees and for white examinees. In Research Questions 2G and 2H, discussed above, original and synthetic eighth-grade test scores were used to predict original BST test scores. In the investigation of this research question, original eighth-grade tests scores were used to predict original BST scores, while synthetic eighth-grade test scores were used to predict scores on the synthetic BST tests. The similarity of coefficients of prediction and regression equations for black examinees and white examinees for the unaltered original tests and the synthetic tests which conform to the Golden Rule stipulations were then compared. The coefficients of prediction for the synthetic tests were adjusted for differences in reliability due to differences in test lengths.

Table 42, below, contains parameter estimates for the regression equations, and the coefficients of prediction (adjusted for test length for the synthetic tests). As discussed above, the slopes of the regression of BST Reading results on Eighth-Grade Reading Test results were very similar for black examinees and white examinees: the confidence intervals for corresponding

parameters overlap. This relationship was also found when the GRS-BST Reading results were regressed on GRS-Eighth-Grade Reading Test results. However, when GRP-BST Reading results were regressed on GRP-Eighth-Grade Reading Test results, the slopes for black examinees and white examinees were found to be dissimilar, and the confidence intervals on corresponding slopes did

Table 42

Estimated Parameters of Regression Equations Relating Original Eighth-Grade Reading Test Results to Original BST Reading Test Results, and Synthetic Eighth-Grade Reading Test Results to Synthetic BST Reading Test Results, by Race

Predictor: (Criterion)	Prediction of BST Reading Results: White Examinees Intercept (95%CI)	Slope (95%CI)	R_W	Black Examinees Intercept (95%CI)	Slope (95%CI)	R_b
Eighth-Grade Reading (BST Reading)	16.8865 (12.99,-20.78)	0.6749 (0.64,-0.71)	0.7212 (0.69,-0.75)	12.4664 (7.79-17.15)	0.6812 (0.63-0.73)	0.7728 (0.74,-0.81)
GRP-Eighth-Grade Reading (GRP-BST Reading)	15.8723 (12.75,-18.98)	0.5685 (0.53,-0.61)	0.6692 (0.64,-0.70)	11.4374 (7.77,-15.14)	0.5976 (0.55,-0.64)	0.7693 (0.73,-0.80)
GRS-Eighth-Grade Reading (GRS-BST Reading)	16.4229 (12.83,-20.00)	0.6374 (0.60,-0.68)	0.6984 (0.67,-0.73)	13.5323 (9.20,-17.86)	0.6369 (0.59,-0.69)	0.7540 (0.71,-0.79)
Eighth-Grade Mathematics (BST Mathematics) 0.84)	7.9720 (5.19,-10.76)	0.8051 (0.78,-0.83)	0.8443 (0.83,-0.86)	9.5151 (5.68,-13.35)	0.7517 (0.70,-0.80)	0.8080 (0.78-
GRP-Eighth-Grade Mathematics (GRP-BST Mathematics)	9.9148 (7.95,-11.87)	0.6886 (0.65,-0.73)	0.7672 (0.74,-(0.79)	9.2581 (6.43-12.09)	0.6652 (0.61-0.72)	0.7474 (0.71,-0.78)
GRS-Eighth-Grade Mathematics (GRS-BST Mathematics)	8.3162 (6.01,-10.63)	0.7081 (0.675,-0.741)	0.7931 (0.77,-0.81)	10.3357 (7.14-13.53)	0.6481 (0.60-0.69)	0.7607 (0.72,-0.79)

not overlap. In this case, application of the Golden Rule procedures to both a test and its criterion exacerbated, rather than decreased, the dissimilarity of the regression equations for black examinees and white examinees. The coefficient of prediction for black examinees was larger than that for white examinees, when Eighth-Grade Reading Test results were used to predict BST Reading Test results. Application of the Golden Rule procedures increased the dissimilarity of the coefficients of prediction for black examinees and white examinees, for both the GRS- and GRP- reading tests.

The slopes of the regressions of BST Mathematics results on Eighth-Grade Mathematics Test results are dissimilar for black examinees and white examinees: the confidence intervals around the parameter for one group does not contain the value for the other group. Application of the Golden Rule procedures to the mathematics tests had inconsistent effects: the slopes for black examinees and white examinees were more similar for the GRP- mathematics tests than for the original tests, but the slopes for black examinees and white examinees were less similar for the GRS-mathematics tests than for the original tests. The coefficients of prediction for black examinees and white examinees were more similar when synthetic eighth-grade mathematics test results were used to predict synthetic BST mathematics test results than when the original Eighth-Grade Mathematics Test results were used to predict the original BST Mathematics Test results; this was true for both the GRS- and GRP- synthetic mathematics tests.

The findings described above lead to the conclusion that applying the Golden Rule procedures to both a test and the criterion it is to predict does not consistently increase the similarity of regression equations or coefficients of

prediction for black examinees and white examinees. The results of applying the Golden Rule procedures in this way appear to be unpredictable: in some cases it might increase similarity of regression parameters for black examinees and white examinees, in others it might decrease the similarity of these parameters.

To further clarify the effect of applying the Golden Rule procedures to both a test and the criterion it is to predict, regression models were fit in which ethnic-group membership was included as a predictor. (Only black and white ethnic groups were included in these analyses). Tables 43 and 44, below, contain the results of these regressions.

Table 43

Estimates of Parameters of Regression of Original BST Reading Test Results on Original Eighth-Grade Reading Test Results and Regression of Synthetic BST Reading Test Results on Synthetic Eighth-Grade Reading Test Results

Eighth-Grade Test:	Prediction of Original and Synthetic BST Reading Results: Intercept (95%CI)	Coeff. of Test (95%CI)	Coeff. of Ethnic (95%CI)	Coeff. of Ethnic*Test (95%CI)
Eighth-Grade Reading	16.8865 (12.84, 20.93)	0.6749 (0.638, 0.712)	-4.4201 (-10.325, 1.485)	0.0062 (-0.052, 0.064)
GRP-Eighth-Grade Reading	15.8727 (12.66, 19.08)	0.5685 (0.531, 0.606)	-4.4348 (-9.119, 0.249)	0.0291 (-0.028, 0.087)
GRS-Eighth-Grade Reading	16.4229 (12.69, 20.16)	0.6373 (0.599, 0.675)	-2.8905 (-8.341, 2.561)	-0.0004 (-0.059, 0.058)

As shown in Table 43, the coefficient of the interaction of ethnic group and eighth-grade test score was not significant when Eighth-Grade Reading Test results were used to predict original BST Reading Test results. The probability that the parameter corresponding to the coefficient of the interaction term was

equal to zero was p = 0.8335. Consistent with the conclusions drawn earlier, the interaction of ethnic-group membership and eighth-grade test score also was not significant when GRP- and GRS- Eighth-Grade Reading Test results were used to predict GRP- and GRS-BST Reading Test results. The probability that the parameters corresponding to the coefficients of the interaction terms were equal to zero were p = 0.3209 and p = 0.9892, respectively, for the synthetic tests.

Table 44

Estimates of Parameters of Regression of Original BST Mathematics Test Results on Original Eighth-Grade Mathematics Test Results and Regression of Synthetic BST Mathematics Test Results on Synthetic Eighth-Grade Mathematics Test Results

Eighth-Grade Test:	Prediction of Original and Synthetic BST Mathematics Results: Intercept (95%CI)	Coeff. of Test (95%CI)	Coeff. of Ethnic (95%CI)	Coeff. of Ethnic*Test (95%CI)
Eighth-Grade Mathematics	7.9720 (5.04, 10.90)	0.80512 (0.776, 0.835)	1.5410 (-2.989, 6.071)	-0.0534 (-0.104, -0.003)
GRP-Eighth-Grade Mathematics	9.9148 (7.83, 12.00)	0.6886 (0.648, 0.729)	-0.6567 (-3.910, 2.597)	-0.0234 (-.090, 0.044)
GRS-Eighth-Grade Mathematics	8.3162 (5.91, 10.73)	0.7080 (0.673, 0.743)	2.0195 (-1.753, 5.773)	-0.0600 (-0.118, -0.002)

The coefficients of the interaction of ethnic-group membership and eighth-grade test score was significant (α=0.05) when Eighth-Grade Mathematics Test results were used to predict BST Mathematics Test results, as shown in Table 41. The probability that the parameters corresponding to the coefficients of the interaction term was equal to zero was p = 0.0439. Consistent with the conclusions drawn earlier, the interaction of ethnic-group membership and eighth-grade test score was also significant when GRS-Eighth-Grade

Mathematics Test results were used to predict GRS-BST Mathematics Test results, but not when GRP-Eighth-Grade Mathematics Test results were used to predict GRP-BST Mathematics Test results. The probabilities that the parameters corresponding to the coefficients of the interaction terms were equal to zero were $p = 0.0431$ and $p = 0.4974$, respectively, for the GRS- and GRP- synthetic tests. Application of the Golden Rule procedures to both the Eighth-Grade and BST Mathematics Tests had inconsistent effects on the similarity of the regression equations for black examinees and white examinees.

Summary of Results of Investigation of Research Questions

The empirical component of this dissertation addressed two major research questions. Research Question 1 addressed whether application of the Golden Rule procedures is effective in reducing the adverse impact of test use. Research Question 2 addressed whether application of the Golden Rule procedures is effective in reducing test bias. Research Question 2 was addressed through investigation of a series of subsidiary research questions. In investigating these research questions, synthetic tests composed solely of Golden Rule Type I items were created in two ways: in the first, the observed item-correct-answer rates were treated as population parameters (the GRP-synthetic tests) ; in the second, they were treated as sample statistics (the GRS-synthetic tests). Despite the fact that these two methods of composing the synthetic tests led to the exclusion of different numbers of items from the original tests, conclusions based on the results of investigation of the research questions were often consistent across the two method used to create synthetic tests.

The results of this study suggest that an affirmative answer is indicated for Research Question 1: applying the Golden Rule procedures substantially (and

with statistical significance) reduced the difference between the average total test scores of black examinees and white examinees. This was true for both the GRS- and GRP- synthetic tests.

The question of whether application of the Golden Rule procedures is effective in reducing test bias (Research Question 2) is complex, and a monosyllabic answer is inappropriate. However, the evidence examined in this dissertation suggests that the Golden Rule procedures are not effective in reducing test bias, and, in some circumstances may exacerbate test bias. Further, the findings of this study suggest that the Golden Rule procedures undermine important psychometric properties of tests.

Application of the Golden Rule procedures to the tests examined in this study resulted in the same or lower values of Coefficient Alpha reliability, even when reliability estimates for the synthetic tests were adjusted for decreased length. This finding was consistent across grade-levels, subjects, and methods used to create the synthetic tests.

When test reliability was examined separately by race, application of the Golden Rule procedures to the tests examined in this study had inconsistent effects. The difference between the reliability estimates for black examinees and white examinees was greater for the GRS-Eighth-Grade Reading Test than for the original Eighth-Grade Reading Test. The differences between the reliability estimates for black examinees and white examinees were also greater for the GRP-Eighth-Grade Reading Test than for the original Eighth-Grade Reading Test, and for the GRP-Eighth-Grade Mathematics Test than for the original Eighth-Grade Mathematics Test. For the remaining five comparisons of original

to synthetic test, applying the Golden Rule procedures increased the similarity of reliability estimates for black examinees and white examinees.

The numbers of items addressing each objective in the synthetic tests were not statistically different from what one would expect from random samples of items of sizes equal to the lengths of the synthetic test forms, selected from the original tests. The classification of items as Type II would thus appear to be random, as far as test objectives were concerned. There was also no identifiable common format or content category for items identified as Type II. Applying the Golden Rule procedures does not appear to exclude items of any predictable format, objective designation, or subject content. This conclusion was consistent for all tests examined in this study.

Applying the Golden Rule procedures significantly raised the average item difficulty level of each test examined in this study (i.e., resulted in less difficult tests). This finding is consistent with the earlier finding that application of the procedures increased the average total score. However, a change in the average item difficulty level of a test may alter the ability of the test to discriminate between test takers who possess certain levels of an attribute, such as knowledge or achievement. It might thus alter the usefulness of the test.

Application of the Golden Rule procedures reduced the average item-total score point-biserial correlation for every synthetic test created. For all but the two synthetic tests corresponding to the Eighth-Grade Reading Test, this decrease was statistically significant. Items retained in the synthetic test were thus less effective in discriminating among examinees of differing abilities, on average, than were the items of the original tests, on average.

When eighth-grade test results were used to predict unaltered BST results, the overall predictive validity application of the Golden Rule procedures to the eighth-grade tests reduced their predictive validities significantly in all four cases examined (i.e., GRS- and GRP- Eighth-Grade Reading Tests, and GRS- and GRP-Eighth-Grade Mathematics Tests). In these comparisons, the predictive validities of the synthetic tests were adjusted to compensate for differences attributable to differences in reliabilities due to test length.

When predictive validities of the eighth-grade tests were examined by race, it was found that applying the Golden Rule procedures reduced predictive validity for both black examinees and white examinees, although the impact of the procedures on predictive validity for white examinees was more marked. This result produced erratic effects on differences between predictive validities for black examinees and white examinees: for both the GRS- and GRP- Eighth-Grade reading tests, differences between predictive validities for black examinees and white examinees were increased; for the GRS- and GRP-Eighth-Grade Mathematics Tests, differences between predictive validities for black examinees and white examinees were decreased.

When eighth-grade test results were used to predict unaltered BST results, the Eighth-Grade Reading Test did not exhibit differential regression slopes: the slopes for black examinees and white examinees were not statistically different. This was also true for the GRP- and GRS- Eighth-Grade Reading Tests. The slopes associated with the original Eighth-Grade Mathematics results were dissimilar for black examinees and white examinees, and applying the Golden Rule procedures increased, rather than decreased, this dissimilarity.

In investigating the final research question of this study, the Golden rule procedures were applied to the predictor (i.e., the eighth-grade tests) and the criterion (i.e., the BST tests) in various prediction equations. The differential predictive validities and differential regressions of these relationships were then compared to those of the original eighth-grade and BST tests. As stated above, the original Eighth-Grade Reading Test did not exhibit differential regression in slope (when used to predict the original BST Reading Test). The GRS-Eighth-Grade Reading Test also did not exhibit differential regression in slope (when used to predict the GRS-BST Reading Test), however, the GRP-Eighth-Grade Reading Test did exhibit differential regression in slope (when used to predict the GRP-BST Reading Test). Application of the Golden Rule procedures to both eighth-grade and BST mathematics test results had inconsistent effects: for the GRP-synthetic mathematics tests, differential regression in slope was reduced; for the GRS- synthetic mathematics tests, differential regression in slope was increased. Predictive validities were decreased when the Golden Rule procedures were applied to both the Eighth-Grade and BST Reading Tests, for both GRS- and GRP- versions of these tests. However, predictive validities were increased when the Golden Rule procedures were applied to both the Eighth-Grade and BST Mathematics Tests, for both GRS- and GRP- versions of these tests.

CHAPTER V
DISCUSSION

In previous chapters of this dissertation, the background and context of the Golden Rule procedures were presented, a set of research questions for investigating selected effects of applying the procedures was derived, methodology used in investigating the research questions was described, and results of investigating the research questions were presented. This chapter contains a discussion of the results of investigation. This chapter is divided into three sections. The first section contains an outline of the limitations of the study and a discussion of the generalizability of findings of the study. The second section contains a discussion of implications of the findings of this study for use in Georgia of the tests examined in this study. The final section contains a discussion of implications of the results of this study for testing in other contexts. The conceptual issues of test bias and adverse impact ground this discussion.

Limitations of this Study

As discussed in Chapter III, the sample used in this study may be biased, in that students with poorest performance may be underrepresented: Subjects initially selected in the eighth grade were chosen through a process similar to linear systematic sampling, but the final sample was selected in such a way as to exclude students who moved to a different school district, were retained in grade, dropped out of school, changed their names, or recorded different ethnic group memberships, names, or genders on the tests

administered in the eighth and tenth grades. It is likely that these excluded students had lower academic achievement, on average, than did the students in the final sample. It is also possible that black students were disproportionately excluded from the sample because of these selection rules. Had a simple random sample of students from either year been used, it is possible that even more items would have failed to meet the criteria for Golden Rule Type I items (as discussed subsequently in this chapter.) The size of these effects are not estimable, and generalization of the results of this study to groups of students who are more transient or less consistent in their progress through school than were students used in this study might not be warranted.

Application of variants of the Golden Rule procedures has been proposed for tests of a variety of types and purposes, including admissions tests, certification tests, licensure tests, promotion tests, competency tests, and placement tests. The tests used in this study certainly do not constitute a representative sample of the types of tests to which application of the Golden Rule procedures has been proposed. Certain psychometric characteristics of a test, such as its average item difficulty level, vary with the intended uses and purposes of the test. These characteristics are relevant to the effects of applying the Golden Rule procedures. In this study, the tests used to investigate the results of applying the Golden Rule procedures were designed for use in determining whether students had mastered knowledge deemed essential by educators. If a single group of students was to be identified through use of these tests, it was presumably those of lowest achievement. Consistent with this purpose, these tests were composed of items at a variety

of difficulty levels, with a majority of items somewhat easier than middle level (i.e., $p>0.05$) in difficulty. Tests intended for other purposes, such as to discriminate those of highest ability from other examinees, would likely be composed of relatively difficult items. Such differences in test purposes and attendant test item characteristics would have profound effects on the numbers of items classified as Type II because of correct answer rates lower than 40%. In turn, this might impact the effects of applying the procedures on other psychometric properties of the tests.

Other characteristics of the tests examined in this study are also unlikely to be representative of all tests to which application of the Golden Rule procedures has been suggested. The consequences of applying the procedures to other types of tests might, therefore, be very different. For example, the tests used in this study exhibited substantial differences in average total scores between black examinees and white examinees, and there is reason to believe that, to some extent, these differences reflected actual differences in knowledge levels between black examinees and white examinees. The knowledge or abilities measured by some tests to which the Golden Rule procedures might be applied will likely be more comparably distributed between black examinees and white examinees, while the knowledge or abilities measured by other tests will likely be less comparably distributed between black examinees and white examinees. Such variations might significantly affect the consequences of applying the Golden Rule procedures to these tests.

The tests examined in this study were, by traditional psychometric standards, fairly soundly constructed; there was little evidence that these tests

were biased against black examinees, as the term "biased" is used by psychometricians and others in the measurement profession. The consequences of applying the Golden Rule procedures to tests that are less soundly constructed, or to tests with compelling evidence of racial bias, would likely be very different.

Implications of Results of this Study for Use of the Georgia Eighth-Grade Criterion Referenced Tests and the Georgia Basic Skills Tests

The tests used to investigate the results of applying the Golden Rule procedures were designed for use in determining whether students had mastered knowledge deemed essential by educators. According to the Georgia Department of Education, the tests are intended to serve several purposes, including public accountability; functioning as an aid to school personnel in individualization of education for students with knowledge deficits; functioning as an aid to school personnel in the identification of students who are at risk for failing the Basic Skills Tests (in the case of the eighth-grade tests); and as a method for protecting the integrity of the high school diploma (in the case of the Basic Skills Tests). These tests were thus intended to accurately characterize the performance of all students. If a single group of students is to be identified through use of these tests, it would presumably be those of lowest, rather than highest, achievement. These testing purposes would best be served by composing tests of items at a variety of difficulty levels, with a majority of items at middle levels, or somewhat easier than of middle levels, of difficulty.

The tests examined in this dissertation had substantial disparate impact on the samples of students used in this study, in that black examinees had

substantially lower average scores than did white examinees, on all four original tests. However, there was little evidence that these tests were biased against black examinees, as the term "biased" is used by psychometricians and others in the measurement profession. All four tests had high indices of internal-consistency reliability for both black examinees and white examinees. The indices of internal-consistency reliability were very similar for black examinees and white examinees. (Differences between the reliability indices for black examinees and white examinees for three of the four tests were less than 0.005; the difference for the remaining test was 0.01). One of the original tests, the Eighth-Grade Reading Test, was slightly more reliable for black examinees than for white examinees. The other three tests were slightly more reliable for white examinees. When eighth-grade test results were used to predict students' performances on the tenth-grade tests, coefficients of predictive validity were also relatively similar for black examinees and white examinees; one was slightly higher for black examinees, the other was slightly higher for white examinees (these differences were statistically significant, in light of the sample sizes used in this study). When students' eighth-grade test results were used to predict their performances on the tenth-grade tests, the slopes associated with the Eighth-Grade Reading Tests were not significantly different for black examinees and white examinees. The difference between the slopes for black examinees and white examinees, when Eighth-Grade Mathematics Test results were used to predict BST Mathematics Test results were small, although they were statistically significant, again, as a result of large sample sizes. (For the Eighth-Grade Reading Test, the intercept for black examinees was lower than for white examinees, and the slopes for black

examinees and white examinees were essentially identical, indicating that use of a single regression line to predict students' performances on the tenth-grade tests from their performances on the eighth-grade tests would consistently overestimate tenth-grade reading performance for black examinees, and underestimate tenth-grade reading performance for white examinees. For the Eighth-Grade Mathematics Test, the slopes were not parallel, and the intercept was higher for black examinees than for white examinees, indicating that, at some score levels, use of a single regression line would overestimate tenth-grade performance one racial group, and at other score levels, underestimate tenth-grade performance of that group. In the range in which most eighth-grade mathematics test scores occurred, use of a single regression line tended to overestimate the tenth-grade performance of black examinees and underestimate that of white examinees). The Georgia tests used in this study were thus initially of relatively sound construction and demonstrated no compelling evidence of bias against black examinees.

Inherent in the Golden Rule procedures is the assumption that it is ultimately desirable to use tests composed solely of Type I items. Application of the Golden Rule procedures to the standardized tests used in this study led to the classification of many items as Type II. This was true when the observed item-correct-answer rates were treated as sample statistics, and when they were treated as population parameters. More items were classified as Type II because the difference between observed correct answer rates of black examinees and white examinees was greater than 0.15, than because the observed correct answer rates were less than 0.40 for black examinees or white examinees, or for all examinees. For the Reading tests, 11% to 30% of items

were classified as Type II; for the Mathematics tests, 32% to 53% were classified as Type II. For specific test objectives, up to 86% of the items that addressed the objective on the original test were classified as Type II. (Had this study used a random sample of students from either grade level, it is likely that even more items would have failed to meet the Golden Rule criteria for Type I items). This has profound implications for test developers in Georgia. To compose eighth- and tenth-grade criterion referenced tests solely of Type I items while maintaining content representativeness, item pools would have to be increased substantially.

Application of the Golden Rule procedures to the tests examined in this study did reduce the disparate impact of the tests, in that differences between average total scores for black examinees and white examinees were significantly and substantially reduced. These reductions in disparate impact were accomplished through increases in average total scores for both black examinees and white examinees, with a proportionately greater increase realized by black examinees.

The question remains as to whether this reduction in disparate impact would be of benefit to the test user or the test taker, given the intended uses of the tests. The primary aim of the statewide testing program is stated to be: "providing information to teachers, students, parents, concerned citizens, and educational policy and decision makers" (Georgia Department of Education, 1988b, p. I-1). This information is ostensibly provided "to aid teachers and administrators in instructional planning, to aid students and their parents in personal decision-making, and to aid educators and the general public in evaluating the effectiveness of educational programs" (Georgia Department

of Education, 1988b, p. I-1). According to the Georgia Student Assessment Handbook, the criterion-referenced tests are "primarily used to customize each student's learning program with his or her specific needs" (Georgia Department of Education, 1988b, p. I-1). The Eighth-Grade Criterion Referenced Tests in Mathematics and Reading are also specifically intended to "identify students who may need additional learning experiences in the basic skills before taking the High School Basic Skills Tests ...in grade 10" (Georgia Department of Education, 1987a, p.1). The Basic Skills Tests are assertedly intended to protect the integrity of the high school diploma, and are used so that "educators, parents and students can be assured that a student who attains a Georgia high school diploma possesses at least minimal levels of many important basic tools of lifelong learning" (Georgia Department of Education, 1982b. p. 1). It seems unlikely that raising the average total score of all groups of examinees would severely undermine the public-accountability purpose of the tests: The Georgia Department of Education publishes the average scores and, in the case of the Basic Skills Test, the proportion of passing examinees, of all schools and school districts; and the press often makes comparisons between schools and districts on the basis of these test results. A comparable increase in average total scores for all schools would not effect the rank-ordering of the schools, thus would not affect comparisons between schools or districts. Increasing the total scores of black students to a greater extent than those of white students (i.e., reducing the difference between average total scores) would, superficially, seem to be of benefit to black students: it would possibly increase their self-esteem, result in less frequent assignment to low-achievement educational tracks, and result in the

promotion, and graduation with diploma, of a more comparable proportion of black students. Increasing the proportion of black students who graduate with a high school diploma would result in employment and economic gains for blacks, as well (cf., Jaeger, 1989). This would appear to be a just result, from a Rawlsian perspective, since the concerns of the least advantaged group (i.e., black students) should outweigh the concerns of the more advantaged group (e.g., the public, which wished information for purposes of accountability). However, it is possible that reducing the disparity in average total scores would actually be detrimental to black students. If the scores of black students were increased without regard to an increase in knowledge or ability of these students, it might serve to obfuscate the inadequacy of educational provision to black students in some schools: schools which provided adequate instruction to white students but inadequate instruction to black students might show comparable test scores for the two groups, leading to the conclusion that educational provision was adequate for both groups. Similarly, individual black students and their parents might compare their test scores to the the published average scores for schools, school districts, or the state, (most of which are composed predominantly of white students), and conclude that their knowledge (i.e., educational preparation) was comparable to that of white students, when in fact it was not. To the extent that eighth-grade test results are useful in identifying students in need of extra preparation for the Basic Skills Test, and to the extent that such identified needs are actually addressed, increasing the total scores of black students on the eighth-grade tests might lead to denial of needed remedial work for some black students. However, this consideration must be tempered by the fact that

remedial basic skills education programs are not universally agreed to be of benefit to the participants.

Increasing the total scores of black students without concomitant changes in their educational attainment might be argued to undermine the integrity of the high school diploma, which is antithetical to a stated focus of the Basic Skills Tests. However, possession of "at least minimal levels of many important basic tools of lifelong learning" (Georgia Department of Education, 1982b. p. 1) is a rather vague construct, and it is difficult to envision that an increase in the proportion of black students achieving a diploma would seriously undermine the public trust. As Jaeger (1989, p. 510) has noted, performance on a minimum-competency test has not been shown to be a good predictor of a graduate's ability to obtain a job or earn money, provided that person is awarded a high school diploma. To the extent that increasing the proportion of black students achieving a diploma did undermine the public accountability function of the tests (e.g., that employers felt less confident that employees possessing a diploma had certain desired skills), this consideration must be weighed against the benefit to blacks (the least advantaged group in a Rawlsian scheme) of increased access to jobs and other opportunities contingent upon possession of a diploma. Withholding a diploma has profound impact on an individual's economic well-being. Jaeger (1989, p. 511) concluded that "it is by denying students a high school diploma, rather than failing to assure their possession of the skills assessed by minimum-competency tests, that our schools endanger student's economic survival." The Golden Rule procedure is just one of many possible ways in which the proportion of black students denied a high school diploma might

be reduced, and it is not obviously the best one. This study did not evaluate the effects of other strategies (which could include, for example, elimination of the requirement that students pass a basic skills competency test), but it seems possible that there exists an alternative mechanism for decreasing the proportion of black students denied a high school diploma which would not have the negative consequences attendant upon the Golden Rule procedures.

Other factors must also be considered in evaluating the desirability of reducing the disparate impact of use of the tests by applying the Golden Rule procedures. Type II items exhibited no common format or content. To be able to compose a test solely of Type I items, Georgia test developers would be required to increase the size of their item pool substantially, since there is no obvious way to avoid writing a large proportion of Type II items. This has financial implications for test developers in Georgia, and increases in cost would probably be passed along to test users (i.e., the Georgia Department of Education and, ultimately, Georgia tax payers). In this time of tight budget constraints, an increase in funds allocated to one governmental function is often at the expense of allocations to another, perhaps more valuable, function. However, such considerations are not likely to be overriding.

Another consideration in evaluating the desirability of reducing the disparate impact of use of these tests by applying the Golden Rule procedures is that this study found no evidence that such application was effective in reducing test bias, as the term is used by professionals in the measurement community. In addition, application of the Golden Rule procedures to the tests examined in this study did not consistently make the internal consistency reliabilities of the tests for black examinees and white examinees

more similar. In some cases examined in this study, application of the Golden Rule procedures increased the difference between reliability coefficients for black examinees and white examinees; in other cases, it decreased the difference.

Georgia's eighth-grade tests are intended to predict students' performances on the Basic Skills Tests. In this study, application of the Golden Rule procedures to the eighth-grade tests reduced the predictive validities of the tests for both black examinees and white examinees in every case examined. Predictive validities were generally reduced more for white examinees than for black examinees. Nonetheless, differences between predictive validities for black examinees and white examinees were not consistently reduced by applying the Golden Rule procedures. Application of the Golden Rule procedures to the Eighth-Grade Reading Test resulted in an increase in the disparity between the coefficients of reliability for black examinees and white examinees, while application of the procedures to the Eighth-Grade Mathematics Test resulted in an decrease in the disparity between the coefficients of reliability for black examinees and white examinees. Application of the Golden Rule procedures to the eighth-grade tests was not effective in reducing differential regression for the tests examined in this study. The original Eighth-Grade Reading Test did not demonstrate differential regression in prediction of BST Reading Test scores, in that the slopes for black examinees and white examinees were not significantly different. The original Eighth-Grade Mathematics Test did demonstrate differential regression in prediction of BST Mathematics Test scores, in that the slopes for black examinees and white examinees were

significantly different. However, application of the Golden Rule procedures increased, rather than decreased, the difference between the slopes of these regression equations. For the Eighth-Grade Reading Test, the intercept for black examinees was lower than that for white examinees, and the slopes for black examinees and white examinees were essentially identical, indicating that use of a single regression line to predict students' performances on the tenth-grade tests from their performances on the eighth-grade tests would consistently overestimate tenth-grade reading performance for black examinees, and underestimate tenth-grade reading performance for white examinees. Application of the Golden Rule procedures to the Eighth-Grade Reading Test increased, rather than decreased, the difference between the intercepts for black examinees and white examinees, and thus increased the degree of misestimation of tenth-grade reading performance, were a single regression line to be used For the Eighth-Grade Mathematics Test, the slopes of the regression lines were not parallel, and the intercept was higher for black examinees than for white examinees, indicating that, at some score levels, use of a single regression line would overestimate the tenth-grade performance one racial group, and at other score levels, underestimate the tenth-grade performance of that group. In the range in which most eighth-grade mathematics test scores occurred, use of a single regression line tended to overestimate the tenth-grade performance of black examinees and underestimate that of white examinees. Application of the Golden Rule procedures to the Eighth-Grade Mathematics Test increased the distance between the regression lines for black examinees and white examinees over the range where most Eighth-Grade Mathematics Test scores occur, and thus

increased the degree of misestimation of tenth-grade mathematics performances, as well.

The effects of applying the Golden Rule procedures on indices of test bias were inconsistent; in some cases, indices of bias were increased, rather than decreased, by applying the procedures. Shepard (1987, p. 1) and others have expressed concern that adoption of the Golden Rule procedures would "undermine legitimate efforts to screen tests for bias," and the results of this study support this concern for the tests examined here. The Golden Rule procedures might undermine such efforts in several ways. First, use of the procedures might increase test bias. Second, use of the procedures might satisfy test users and the public that bias was being effectively addressed, and thus might reduce the demand for more effective actions. Third, test developers would be required to divert to increasing the sizes of item pools funds which might otherwise be used to conduct more effective analyses of item bias or to protect the psychometric integrity of tests. Finally, use of the procedures might obfuscate existing inequities in educational provision, which the tests are supposedly used to detect.

Not only did this study provide evidence that application of the Golden Rule procedures was ineffective in reducing bias in the tests examined, its application undermined important psychometric properties of the tests. Application of the Golden Rule procedures to the tests used in this study consistently lowered the overall reliability of the tests, even after adjustments had been made for differences in reliability due to differences in test lengths. The amounts by which the Alpha coefficients were decreased were small. However, the reader is reminded that in this study the "item

pools" to which the Golden Rule procedures were applied consisted of items which had already met certain psychometric criteria. If, in developing future versions of the tests, the Golden Rule procedures were applied to an unscreened item pool, it would likely result in a far more dramatic decrease in reliability, provided items were also held to comparable standards for discrimination statistics, and other psychometric criteria. Application of the procedures significantly increased the average item difficulty of the tests examined in this study, which might alter the discrimination of the tests. In most cases examined, application of the procedures significantly reduced the average item-total score correlations of the tests. Application of the Golden Rule procedures also significantly reduced the overall predictive validities of the tests examined in this study, even after the predictive validities of the altered tests had been adjusted to compensate for differences in reliability between the original and the altered tests due to differences in length. When the Golden Rule procedures were applied to both the predictor and criterion tests (i.e., to both the eighth-grade and BST tests), the overall predictive validities were also decreased in some cases examined in this study.

In summary, the results of this study suggest that application of the Golden Rule procedures to the Georgia Eighth-Grade Criterion Referenced Tests and the Georgia Basic Skills Tests would sacrifice crucial psychometric properties of tests without achieving compensating benefits to test users or the public. Achievement of the purposes of the tests, as stated by the Georgia Department of Education, would not be enhanced by application of the procedures. The only obvious advantage to test takers resulting from application of the Golden Rule procedures is that some examinees would be

granted a high school diploma which would be denied them using the current tests. However, there are other ways to decrease the proportion of students denied diplomas which would probably not have the negative consequences that result from applying the Golden Rule procedures. The results of this study suggest that, as a mechanism for reducing test bias in these Georgia tests, the Golden Rule procedures are worse than ineffective: they are inconsistent and sometimes detrimental. While the results of this study suggest that application of the procedures does reduce disparate impact (i.e., reduce the difference between the average total scores for black examinees and white examinees), suggesting that the procedures are tenable as mechanisms of affirmative action, the costs of this reduction of disparate impact are steep, and the benefits are questionable.

Implications of Results of this Study for Testing in Other Contexts

As stated above, application of variants of the Golden Rule procedures has been proposed for tests of a variety of types and purposes, including admissions tests, certification tests, licensure tests, promotion tests, competency tests, and placement tests. And, as already noted, the tests used in this study do not constitute a representative sample of the types of tests to which application of the Golden Rule procedures has been proposed. Generalization of the results of this study to other testing situations therefore might not be warranted. Previous research on the effects of applying the Golden Rule procedures has used data from college and graduate-level admissions tests. Such tests are intended to distinguish a superior group from a pool of self-selected applicants. To best distinguish those of highest ability or aptitude from other examinees, admissions tests are composed of

relatively difficult items. In this study, the tests used to investigate the results of applying the Golden Rule procedures were designed to help educators determine whether students had mastered knowledge they deemed to be essential. These tests were intended to characterize accurately the performances of all students. If a single group of students was to be identified through use of these tests, it was presumably those of lowest, rather than highest, achievement. These purposes would best be served by composing the tests of items at a variety of difficulty levels, with a majority of items at middle levels, or somewhat easier than middle levels, of difficulty. Although direct generalization of the results of this study to other testing contexts is unjustified, this study affords an opportunity to explore the impact of the Golden Rule procedures on tests at the opposite end of a continuum of test purposes and intended test users from those employed in previous research on these procedures.

The tests examined in this dissertation exhibited substantial disparate impact on the samples of students used in this study, in that black examinees had substantially lower average scores than did white examinees, on all four original tests. As discussed above, there was little evidence that these tests were biased against black examinees, as the term "biased" is used by psychometricians and others in the measurement profession. In addition the original tests were soundly constructed. This study thus also afforded an opportunity to examine the impact of applying of the Golden Rule strategy to a test "item pool" that was relatively sound from a psychometric perspective and, in addition, exhibited no compelling evidence of racial bias.

Inherent in the Golden Rule procedures is the assumption that it is ultimately desirable to use tests composed solely of Type I items. Despite the initial soundness of the tests examined in this study, application of the Golden Rule procedures led to the classification of many items as Type II. For these tests, which were presumably intended to focus on distinguishing examinees of lowest achievement from other examinees, more items were classified as Type II because the difference between observed correct answer rates of black examinees and white examinees was greater than 0.15, than because the observed correct answer rates were less than 0.40 for black examinees or white examinees, or for all examinees. For specific test objectives, up to 86% of the items that addressed the objective on the original test were classified as Type II. There was no common format or content of Type II items. This could have profound implications for test developers. To compose a test solely of Type I items, while maintaining content representativeness, an item pool would have to be increased substantially, since there is no obvious way to avoid writing a large proportion of Type II items. This finding has financial implications for test developers, and increases in cost would probably be passed along to the test users. However, such considerations are not likely to be overriding. Of groups with a vested interest in the use of standardized tests (i.e., test developers, test users, the general public, and test takers), minority test takers are clearly the least advantaged and least powerful group in most situations. From a Rawlsian perspective, their interests should be considered and protected above those of other groups. Thus, an increase in costs to test developers or test users could

be justified if increased costs were accompanied by demonstrable benefits to minority test takers.

Application of the Golden Rule procedures to the tests examined in this study reduced the disparate impact of the tests, in that differences between average total scores for black examinees and white examinees were significantly and substantially reduced. These reductions in disparate impact were accomplished through increases in average total scores for both black examinees and white examinees, with a proportionately greater increase realized by black examinees. However, since there was no compelling evidence that the original tests were racially biased, the differences between average total scores for black examinees and white examinees on the original tests must be taken to represent real differences in knowledge or abilities. Thus, reduction of the disparate impact of the tests obscures real differences between examinees on the constructs the tests were intended to measure. Such an action might be considered justified by those who feel that justice requires society to guarantee equality of outcomes as well as equality of opportunity and equality of treatment. However, a more commonly held view is that, while society should guarantee equality of opportunity and equality of treatment, society does not have an obligation to guarantee equality of outcome, particularly if that equality is only apparent. Adherents of this view often believe that achieving equality of outcome artificially, as when the Golden Rule procedures were applied to the tests in this study, might be harmful, in that it might obscure real inequalities in treatment or opportunity.

In summary, the results of this study suggest that application of the Golden Rule procedures might sacrifice crucial psychometric properties of tests without major benefit to test users, test takers, or the public. The results of this study suggest that, as a mechanism for reducing test bias, the Golden Rule procedures are worse than ineffective: they are inconsistent and sometimes detrimental. Although application of the procedures was found to reduce disparate impact in some testing situations, the costs of this reduction are likely to be steep, and the benefits questionable. If each individual's test performance is measured less reliably and less validly, the usefulness of the measurement to the individual is reduced proportionately. Most important, use of the Golden Rule procedures as a mechanism of affirmative action seems likely to camouflage true inequities in educational provision. Although other mechanisms for achieving affirmative action and for detection of racial bias were not examined in this study, the results of this study suggest that, at least for some tests, the Golden Rule procedures provide a bad approach to ameliorating these crucial social problems. Thus the Golden Rule procedures should not be considered an adequate solution to these problems, or lead to the abandonment of other efforts to develop mechanisms for reducing test bias or developing just and effective affirmative action programs.

REFERENCES

Adler, M.J. (1981). *Six great ideas.* New York: Macmillan Publishing Company.

Adler, M.J. & Gorman, W. (Eds.) (1952). *The great ideas: A synopticon of the great books of the western world.* Chicago: William Benton.

Allen, J. and Yen, W.M. (1979). *Introduction to measurement theory.* Montery, Ca.: Brooks/Cole Publishing Company.

Allen et al. v. Alabama State Board of Education et al. (Docket No 81-697-N. U.S. District Court Middle District of Alabama, Northern Division, 1985).

Amaker, N.C. (1989). Civil rights and the Reagan administration. *Harvard Law Review, 102,* 2056.

Angoff, W.H. (1982). Uses of difficulty and discrimination indices for detecting item bias. In R.A. Berk (Ed.) *Handbook of methods for detecting test bias.* (pp. 96-116). Baltimore: Johns Hopkins University Press.

Angoff, W.H. (1989). *Perspectives on the theory and application of differential item functioning methodology.* Princeton, N.J.: Educational Testing Service.

Anrig, G.R. (1987a, January). "Golden Rule": Second thoughts. *APA Monitor,* p. 3.

Anrig, G.R. (1987b). ETS on "Golden Rule." *Educational Measurement: Issues and Practice, 6(2),* 24-27.

Anrig, G.R. (1988). ETS replies to Golden Rule on "Golden Rule." *Educational Measurement: Issues and Practice, 7(1),* 20-21.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, D.C.: American Psychological Association.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1974). *Standards for educational and psychological tests.* Washington, D.C.: American Psychological Association.

Baker, F.B. (1981). A criticism of Scheuneman's item bias technique. *Journal of Educational Measurement, 18,* 59-62.

Berk, R.A. (1982). Introduction. In R.A. Berk (Ed.) *Handbook of Methods for Detecting Test Bias.* (pp. 1-9). Baltimore: Johns Hopkins University Press,

Bond, L. (1981). Bias in mental tests. In B. F. Green (Ed.) *New Directions for Testing and Measurement: Issues in Testing-Coaching, Disclosure, and Ethnic Bias,* no. 11. (pp. 55-77).San Francisco: Jossey-Bass.

Bond, L. (1987). The Golden Rule settlement: A minority perspective. *Educational Measurement: Issues and Practice, 6(2),* 18-20.

Cleary, A. (1968). Test bias: Prediction of grades of negro and white students in integrated colleges. *Journal of Educational Measrument, 5,* 115-124.

Cochran, W.G. (1977). *Sampling techniques,* (3rd. ed.). New York: John Wiley and Sons.

Cole, N.S. & Moss, P.A. (1989). Bias in test use. In R.L. Linn (Ed.) *Educational Measurement* (pp. 201-219). New York: Macmillan Publishing Company.

Columbus Board of Education v. Penick, 443 U.S. 499 (1979).

Cohen, D.K. & Haney, W. (1980). Minimums, competency testing, and social policy. In R.M. Jaeger & C.K. Tittle. *Minimum competency achievement testing: Motives, models, measures, and consequences.* (pp. 5-22) Berkley: McCutchan Publishing Company.

Coyle, M. (1990). Undoing another's handiwork: Congress, high court clash on civil rights. *National Law Journal, 12,* 1-4.

Cronback, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrica, 16,* 297-334.

Cronback, L.J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.) Intelligence: measurement theory and public policy. (Proceedings of a symposium in honor of Lloyd G. Humphreys). Urbana, Il.: University of Illinois Press.

Cronbach, L.J. (1976). Equity in selection- where psychometrics and political philosophy meet. *Journal of Educational Measurement, 13,* 31-42.

Debra P. v. Turlington, 474 F. Supp. 244, 247 (M.D. Fla. 1979).

Diaz v. San Jose Unified School District, 733 F. 2d. 660 (9th Cir 1984).

Dorans, N.J. (1984, December). *Approximate IRT formula score and scaled score standard errors of measurement at different ability levels.* (Statistical Report No. SR-84-118). Princeton, New Jersey: Educational Testing Service.

Edmonds, R. (1979) Effective schools for the urban poor. *Educational Leadership, 37*(1), 15-27.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice. (August 1978). Adoption by four agencies of uniform guidelines on employee selection procedures. *Federal Register, 43,* 38290-38315.

Faggen, J. (1987). Golden Rule revisited: Introduction. *Educational Measurement: Issues and Practice, 6(2),* 5-7.

Flaugher, R.L. (1978). The many definitions of test bias. *American Psychologist, 33*(7), 671-679.

Flaugher, R.L. & Schrader, W.B. (1978, March). *Eliminating differentially difficult items as an approach to test bias.* (Research Bulletin). Princeton, N.J.: Educational Testing Service.

First, J.M. & Cardenas, J. (1986). A minority view on testing. *Educational Measurement: Issues and Practice, 5,* 6-11.

Frierson, H.T. (1990) The situation of black educational researchers: Continuation of a crisis. *Educational Researcher, 19*(2), 12-17.

Gay, L.R. (1976). *Educational research: Competencies for analysis and application.* Columbus, Oh: Charles E. Merrill Publishing Company.

Georgia Department of Education, (1982a). *Georgia basic skills test: Information for students.* Atlanta: Division of Standards and Assessment, Office of Planning and Evaluation, Georgia Department of Education.

Georgia Department of Education, (1982b). *Georgia high school basic skills tassessment*. Atlanta: Division of Standards and Assessment, Office of Planning and Evaluation, Georgia Department of Education.

Georgia Department of Education, (1987a). *Georgia criterion-referenced rests: Grade 8 teacher's interpretive guide*. Atlanta: Division of Standards and Assessment, Office of Planning and Evaluation, Georgia Department of Education.

Georgia Department of Education, (1987b). *Student Assessment in Georgia 1986-1987* Atlanta: Division of Standards and Assessment, Office of Planning and Evaluation, Georgia Department of Education.

Georgia Department of Education, (1988a). *Georgia Basic Skills Test Fall 1987-Spring 1988 interpretive guide.* Atlanta: Division of Standards and Assessment, Office of Planning and Evaluation, Georgia Department of Education.

Georgia Department of Education, (1988b). *Student assessment handbook.* Atlanta: Division of Assessment, Office of Evaluation and Personnel Development, Georgia Department of Education.

Glass, G.V & Hopkins, K.D. (1984). *Statistical methods in education and psychology*, (2nd ed.). Englewood Cliffs: Prentice-Hall, Inc.

Glickman, C.D. & Pellegrini, A.D. (1988). The case against kindergarten testing for promotion. *Principal, 68*, 34-35.

Golden Rule Insurance Company, et al. v. Duncan, et al., 419-76 (order for dismissal, filed in the Circuit Court of the Seventh Judicial Circuit, Sangamon County, IL, 1978)

Golden Rule Insurance Company, et al. v. Mathias, et al., (order for dismissal of amended case, filed in the Circuit Court of the Seventh Judicial Circuit, Sangamon County, IL, 1979)

Golden Rule Insurance Company, et al. v. Mathias, et al., 86 Ill. App. 3rd. 323 (order affirming in part, reversing in part, and remanding, 1980)

Golden Rule Insurance Company, et al. v. Washburn, et al., 419-76 (stipulation for dismissal and order dismissing case, filed in the Circuit Court of the Seventh Judicial Circuit, Sangamon County, IL, 1984).

Goldman, R.D., & Hewett, M.H.A. (1975). An investigation of test bias for Mexican-American college students. *Journal of Educational Measurment, 12*(3), 187-196.

Goldman, R.D., & Richards, R. (1974). The SAT prediction of grades for Mexican-American versus Anglo-American students at the University of California, Riverside. *Journal of Educational Measurment, 11*(2), 129-136.

Goldman, R.D. & Widawski, M.H. (1976). An analysis of types of errors in selection of minority college students. *Journal of Educational Mearsurement, 13*(3), 185-200.

Gulliksen, H. (1976). When high validity may indicate a faulty criterion. (RM 76-10). Princeton, N.J.: Educational Testing Service.

Hackett, R.K., Holland, P.W., Pearlman, M., & Thayer, D.T. (1987, July). *Test construction manipulating score differences between black and white examinees: Properties of the resulting tests.* (Research Report). Princeton, New Jersey: Educational Testing Service.

Hambleton, R.K. (1989). Principles and selected applications of item response theory. In R.L. Linn (Ed.) *Educational Measurement* (pp. 147-200). New York: Macmillan Publishing Company .

Haney, W.M. & Reidy, E.F. (1987). Editorial. *Educational Measurement: Issues and Practice, 6(2),* 4.

Hardigan, J.A. & Wigdor, A.K. (Eds.) (1989). *Fairness in employment testing: Validity generalization, minority issues, and the General Aptitude Test Battery.* Washington, DC: National Academy Press.

Hemeryck, S., Butts, C., Jehl, L., Koch, A., & Sloan, M. (1990). Reconstruction, deconstruction and legislative response: the 1988 Supreme Court term and the Civil Rights Act of 1990. *Harvard Civil Rights-Civil Liberties Law Review, 25,* 475-590.

Hills, J.R. (1989). Screening for potentially biased items in testing programs. *Educational Measurement: Issues and Practice, 8*(4), 5-11.

Hobbes, T. (1966). Leviathan, or the matter, form, and power of a commonwealth, ecclesiastical and civil. In W. Molesworth (Ed.) *The English works of Thomas Hobbes of Malmesbury. Vol. 3.* London: John Bohn. (Original work published 1651).

Holland, P.W., & Thayer, D.T. (1985, October). *An alternate definition of the ETS Delta Scale of Item Difficulty.* (Technical Report No. 85-64). Princeton, New Jersey: Educational Testing Service.

Howe, H. & Edleman, M. (1985). *Barriers to excellence: Our children at risk.* Boston: National Coalition of Advocates for Students.

Indep. Fed'n of Flight Attendants v. Zipes, 109 S. Ct. 2732 (1989)

Ironson, G.H. (1982). Use of Chi-square and latent trait approaches for detecting item bias.. In R.A. Berk (Ed.) *Handbook of Methods for Detecting Test Bias* (pp. 117-160). Baltimore: Johns Hopkins University Press.

Ironson, G.H. (1983). Using item response theory to measure bias. In R.K. Hambleton (Ed.) *Applications of item response theory* (pp. 155-174). Vancover, BC: Educational Research Institute of British Columbia.

Jaynes, G.D., & Williams, R.M. (Eds.). (1989). *A common destiny: Blacks and American society.* Report of the Committee on the status of Black Americans, Commission on Behavioral and Social Sciences and Education, National Research Council. Washington, DC: National Academy Press.

Jaeger, R.M. (Ed.) (1976). Bias in selection. *Journal of Educational Measurement, 13*(1).

Jaeger, R.M. (1987). NCME opposition to proposed golden rule legislation. *Educational Measurement: Issues and Practice, 6(2),* , 21-22.

Jaeger, R.M. (1984). *Sampling in education and the social sciences.* New York: Longman.

Jaeger, R.M. (in press). Competency testing. In *Encyclopedia of Educational Research,* 6th ed.

Jaeger, R.M. & Tittle, C.K. (Eds.) (1980). *Minimum competency achievement testing: Motives, models, measures, and consequences.* Berkley: McCutchan Publishing Company.

Jensen, A.R. (1980). *Bias in mental testing.* New York: Free Press.

Jett v. Dallas Indep. School District, 109 S. Ct. 2702 (1989)

Karmel, P. (1985). Quality and equality in education. *Australian Journal of Education, 29,* 279-293.

Kluegel, J.R., & Smith, E.R. (1986). *Beliefs About inequality: American's views of what is and what ought to be.* New York: Aldine De Gruyter.

Kok, F.G., Mellenbergh, G.F., & Van Der Flier, H. (1985). Detecting experimentally induced item bias using the interactive logit method. *Journal of Educational Measurement, 22*(4), 295-303.

Lehmann, I.J. (1990). The state of NCME: Remembering the past, looking to the future. *Educational measurement: Issues and practice. 9* (1), 3-10.

Linn, R.L. & Drasgow, F.D. (1987, January). Implications of the Golden Rule agreement. *The Score,* pp. 4,8.

Linn, R.L. & Drasgow, F.D. (1987) Implications of the Golden Rule settlement for test construction. *Educational Measurement: Issues and Practice, 6(2),* 5-7.

Linn, R.L. & Harnisch, D.L. (1981). Interactions between item content and group membreship on achievement test items. *Journal of Educational Measurement, 18,* 109-118.

Linn, R.L., Madaus, G.F., & Peudlla, J. (1982). Minimum competency testing: Cautions on the state of the art. *American Journal of Education, 91,* 1-35.

Locke, John. (1964). Two treatises of government. In P. Laslett (Ed.) *Two treatises of government by John Locke.* Cambridge: University Press. (Original work published 1689)

Lorance v. AT & T Technologies, 109 S. Ct. 2261 (1989)

Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum.

LULAC, GI Forum, & NAACP v. State of Texas, (1985, August 27). Memorandum decision in the United States District Court for the Eastern District of Texas, Tyler Division, Civil Action No. 5281.

McAllister, P.H. (1987, April). *Overview of state legislative initiatives concerning standardized testing.* Paper presented at the annual meeting of the American Educational Research Association: Washington, D.C.

McNemar, Q. (1975). On so-called test bias. *American Psychologist, 30,* 848-851.

Marasculio, L.A., & Slaughter, R.E. (1981). Statistical procedures for identifying possible sources of item bias based on χ^2 statistics. *Journal of Educational Measurement, 18*(4), 229-248.

Marco, G. L. (1987, April). *Does the use of test assembly procedures proposed in legislation make any difference in test properties and in the test performance of black and white test takers?* Paper presented at the annual meeting of the American Educational research Association: Washington, D.C.

Marco, G. L. (1988). Does the use of test assembly procedures proposed in legislature make any difference in test properties and in the test performance of black and white test takers? *Applied Measurement in Education,* 1(2), 109-133.

Martin v. Wilks, 109 S. Ct. 2180 (1989)

Menacker, J. (1987). *School law: Theoretical and case perspectives.* Englewood Cliffs: Prentice Hall, Inc.

Messick, S. (1989). Validity. In R.L. Linn (Ed.) *Educational Measurement* (pp. 13-103). New York: Macmillan Publishing Company.

National Center for Educational Statistics. (1989). *Digest of educational statistics.* Washington, D.C.:U.S. Department of Education.

National Center for Educational Statistics. (1988). *Digest of educational statistics.* Washington, D.C.:U.S. Department of Education.

National Commission on Testing and Public Policy. (1990). From gatekeeper to gateway: Transforming testing in America. Chestnut Hill, MA: Boston College.

Novick, M.R. & Peterson, N.S. (1976). Toward equalizing educational and employment opportunity. *Journal of Educational Measurement,* 13, 77-88.

Parents in Action on Special Education (PASE) v. Hannon, 506 F. Supp. 831 (1980) at 836-37.

Patterson v. McLean Credit Union, 109 S. Ct. 2363 (1989).

Personal Administrator v. Feeney, 442 U.S. 256 (1979).

Peterson, N.S. & Flesher, R.B. (1982, November). *Test analysis of College Board Scholastic Aptitude Test March 1982 administration.* (Statistical Report No. SR-82-103). Princeton, New Jersey: Educational Testing Service.

Peterson, N.S., & Novick, M.R. (1976). An evaluation of some models for culture-fair selection. *Journal of Educational Measurement, 13,* 3-29.

Pipho, C. (1985, May 2). Tracking the reforms, part 5: Testing- Can it measure the success of the reform movement? *Education Week,* p. 19.

Plake, B.S. (1980) A comparison of a statistical and subjective procedure to ascertain item validity: One step in the test validation process. *Educational and Psychological Measurement, 40,* 397-404.

Price v. Denison Independent School District, 694 F. 2nd 334 (5th Cir. 1982).

Price-Waterhouse v. Hopkins, 109 S. Ct. 1775 (1989)

Rawls, J. (1971). *A theory of justice.* Cambridge: Belknapp Press of Harvard University Press.

Reutter, E.E., Jr. (1985). *The law of public education*, 3rd ed. New York: Foundation Press.

Reynolds, C.R.(1982a). Methods for detecting construct and predictive bias. In R.A. Berk (Ed.) *Handbook of methods for detecting test bias* (pp. 199-227). Baltimore: Johns Hopkins University Press.

Reynolds, C.R. (1982b). The problem of bias in psychological assessment. In. C.R. Reynolds & T.B. Gutkin (Eds.) *The handbook of school psychology* (pp. 178-208). New York: Wiley.

Reynolds, C.R., & Brown, R.T. (1984). *Perspectives on bias in mental testing.* New York: Llenum Press.

Rogers, J., Dorans, N.J., & Schmitt, A.P. (1986, January). *Assessing unexpected differential item performance of black candidates on SAT form 3GSA08 and TSWE form E43.* (Statistical Report No. SR-86-22). Princeton, New Jersey: Educational Testing Service.

Romberg, T.A., Zarinnia, E.A., & Williams, S.R. (1989). The influence of mandated testing on mathematics instruction: Grade 8 teachers' perceptions. Madison, WI: Wisconsin Center for Educational Research, University of Wisconsin-Madison.

Rooney, J.P. (1987a). Golden Rule on "Golden Rule". *Educational Measurement: Issues and Practice, 6(2),* 9-12.

Rooney, J.P. (1987b). A response from Golden Rule to "ETS on 'Golden Rule'". *Educational Measurement: Issues and Practice, 6(4),* 19-23.

Rottenberg, C, & Smith, M.L. (1990, April). Unintended effects of external testing in elementary schools. Paper presented at the annual meeting of the American Educational Research Association, Boston.

Rovell, S. (1962, October). *Binomial confidence limits.* Unpublished manuscript.

Sandel, M.J. (1982). *Liberalism and the limits of justice.* New York: Cambridge University Press.

SAS Institute. (1985a). SAS user's guide: Basics, version 5 edition . Cary, N.C.; SAS Institute.

SAS Institute. (1985b). SAS user's guide: Statistics, version 5 edition. Cary, N.C.; SAS Institute.

Schmeiser, C.B., & Ferguson, R.L. (1978). Performance of black and white students on test materials containing content based on black and white cultures. *Journal of Educational Measurement, 15*(3), 193-200.

Scheuneman, J.A. (1976). A method for assessing bias in test items. *Journal of Educational Measurement, 16,* 143-152.

Schmitt, A.P. & Dorans, N.J. (1987, March). *Differential item functioning on the Scholastic Aptitude Test.* (Research Memorandum). Princeton, New Jersey: Educational Testing Service.

Scriven, M. (1988). Philosophical inquiry methods in education. In. R.M. Jaeger, Ed. *Complementary Methods for Research in Education* (pp. 131-148) Washington, D.C.: American Educational Research Association.

Shapiro, W. (1989, August 7). Unfinished business. *Time.* pp. 12-15.

Shepard, L.A. (1981). Identifying bias in test items. In B.F. Green (Ed.) *New Directions for Testing and Measurement: Issues in Testing- Coaching , Disclosure, and ethnic Bias, 11,* San Francisco: Jossey-Bass.

Shepard, L.A. (1982). Definitions of bias. In R.A. Berk (Ed.) *Handbook of Methods for Detecting Test Bias.* (pp. 9-31). Baltimore: Johns Hopkins University Press.

Shepard, L.A. (1987, January). The Golden Rule agreement: Bad law, bad science. *The Score,* pp. 7-9.

Siegel, S. & Castellan, N.J. Jr. (1988). *Nonparametric Statistics for the behavioral sciences.* New York: McGraw-Hill Book Company.

State Assessment Programs (1988). *Student assessment handbook.* Atlanta, Georgia: Division of Assessment, Office of Evaluation and Personnel Development, Georgia Department of Education.

Thorndike, R.L. (1971a). Concepts of culture-fairness. *Journal of Educational Measurement, 8*(2), 63-70.

Thorndike, R.L. (1971b). *Educational measurement* (2nd ed.). Washington, D.C.: American Council on Education.

Thorndike, R.L. (1982). *Applied psychometrics.* Boston: Houghton Mifflin Company.

Tittle, C.K. (1982). Use of judgmental methods in item bias studies. In R.A. Berk (Ed.) *Handbook of Methods for Detecting Test Bias.* (pp. 31-63). Baltimore: Johns Hopkins University Press.

Titus, H.H. (1970). *Living issues in philosophy.* New York: Van Nostrand Reinhold Company.

United States v. LULAC, 793 F. 2nd 636 (5th Cir. 1986).

Village of Arlington Heights v. Metropolitan Housing Development Corp., 429 U.S. 252 (1977).

Wards Cove Packing Co. v. Antonio, 109 S. Ct. 2115 (1989)

Weiss, J. (1987). The Golden Rule bias reduction principle: A practical reform. *Educational Measurement: Issues and Practice, 6(2),* 23-25.

Werner, E. (1988, September). *On Golden Rule: A critical examination of its background, procedures, and program impact.* Paper presented at the annual meeting of the National CLEAR Conference: Washington, D.C.

Wisconsin Assembly Bill No. 855. (1986, February 18).

Wolff, R.P. (1977). *Understanding Rawls: A reconstruction and critique of 'A Theory of Justice'.* Princeton: Princeton University Press.

Zoreff, L. & Williams, P. (1980). A look at content bias in IQ tests. *Journal of Educational Measurement, 17*(4), 313-320.

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics, 15 (3),* 185-197.

APPENDIX A

Legislative Initiatives to Regulate Standardized Testing
with Stipulations Similar to those of
Golden Rule Insurance Company, et al. v. Washburn, et al., (1984).

State	Year	Bill Number	Sponsor
CA	1986	A.4045	Moore (D)
CA	1986	A.4046	Moore (D)
MA	1985	S.2530	Houston (D) & Melconian (D)
MA	1987	S.683	(reintroduction of S. 2530)
NY	1986	S.8985/A.11023	LaValle (R) & Eve (D)
NY	1986	S.9020/A.11029	LaValle (R) & Eve (D)
NY	1986	S.9015/A.11028	LaValle (R) & Eve (D)
NY	1987	S.3614/A.5582	(reintroduction of S. 8995/A.11023)
NY	1987	S.3624/A.5600	(reintroduction of S.9020/A.11029)
NY	1987	S.3623/A.5601	(reintroduction of S.9015/A.11028)
TX	1985	S.657/A.1441	Truan (D) & Moreno (D)
TX	1985	S.624/H.1377	Truan (D) & Olivera (D)
TX	1987	S.29/A.325	Luna (D) (reintroduction of S.657/A.1441)
TX	1987	S.28/A.240	Luna (D) (reintroduction of S.624/H.1377)
WI	1986	A.885	Becker (D)

APPENDIX B

SKILL AREAS AND OBJECTIVES MEASURED BY THE GEORGIA EIGHTH GRADE CRITERION-REFERENCED TEST IN MATHEMATICS *

SKILL AREA: CONCEPT IDENTIFICATION — This skill area contains the basic vocabulary of mathematics and the interrelationships of different kinds of numbers.

Objective 1. The student translates forms of rational numbers in the context of academic tasks, everyday tasks, or employment activities.

Objective 2. The student identifies relations or properties of sets of numbers and operations in the context of academic tasks, everyday tasks, or employment activities.

Objective 3. The student selects customary or metric units to measure length, area, volume, weight, time and temperature in the context of academic tasks, everyday tasks, or employment activities.

Objective 4. The student identifies relations and properties of sets of points in the context of academic tasks, everyday tasks, or employment activities.

SKILL AREA: COMPONENT OPERATIONS — This skill area involves actions on numbers and focuses on addition, subtraction, multiplication and division, as well as using units of measurement. The student:

Objective 5. The student determines probabilities in the context of academic tasks, everyday tasks, or employment activities.

Objective 6. The student computes with whole numbers, fractions, decimals, integers and percents in the context of academic tasks, everyday tasks, or employment activities.

Objective 7. The student applies formulas or units of measure to determine length, area, volume, weight, time, temperature and amounts of money in the context of academic tasks, everyday tasks, or employment activities.

APPENDIX B, continued

SKILL AREA: PROBLEM SOLVING — This skill area requires the student to select or apply the appropriate concepts and/or operations necessary to solve problems. The student:

Objective 8. The student selects tie appropriate operation for a given problem situation and the reverse in the context of academic tasks, everyday tasks, or employment activities.

Objective 9. The student solves word problems in the context of academic tasks, everyday tasks, or employment activities.

Objective 10. The student organizes data in the context of academic tasks, everyday tasks, or employment activities.

Objective 11. The student interprets data which have been organized in the context of academic tasks, everyday tasks, or employment activities.

Objective 12. The student estimates results in the context of academic tasks, everyday tasks, or employment activities.

*Source: Georgia Criterion-Referenced Tests Grade 8: Interpretive Guide (GDE, 1987)

APPENDIX C

SKILL AREAS AND OBJECTIVES MEASURED BY THE GEORGIA EIGHTH GRADE CRITERION-REFERENCED TEST IN READING*

SKILL AREA: LITERAL COMPREHENSION — This area involves understanding information which is explicitly stated in written material.

Objective 1. The student distinguishes between fact and opinion in the context of academic, everyday, or employment materials.
Objective 2. The student recognizes explicitly stated main ideas, details, sequences of events, and cause and effect relationships in the context of academic, everyday, or employment materials.
Objective 3. The student interprets instructions in the context of academic, everyday, or employment materials.

SKILL AREA: INFERENTIAL COMPREHENSION — This area involves understanding information that can be determined from written material even though it is not directly stated. The student:

Objective 4. The student recognizes implicitly stated main ideas, details, sequences of events, and cause and effect relationships in the context of academic, everyday, or employment materials.
Objective 5. The student interprets word meanings and patterns of language in the context of academic, everyday, or employment materials.
Objective 6. The student interprets figurative language in the context of academic, everyday, or employment materials.
Objective 7. The student recognizes propaganda techniques in the context of academic, everyday, or employment materials.

SKILL AREA: PROBLEM SOLVING — This skill area involves locating, recognizing, interpreting or evaluating information needed to make decisions or solve problems. The student:

Objective 8. The student uses reference sources in the context of academic, everyday, or employment materials.
Objective 9. The student makes generalizations and draws conclusions in the context of academic, everyday, or employment materials.
Objective 10. The student makes predictions and comparisons in the context of academic, everyday, or employment materials.
Objective 11. The student recognizes relevance of data in the context of academic, everyday, or employment materials.

*Source: Georgia Criterion-Referenced Tests Grade 8: Interpretive Guide (GDE, 1987)

APPENDIX D

SKILL AREAS AND OBJECTIVES MEASURED BY THE GEORGIA BASIC SKILLS TEST IN READING*

SKILL AREA: LITERAL COMPREHENSION — This area involves understanding information which is explicitly stated in written material.

Objective 1: The student distinguishes between fact and opinion. Answers to fact and opinion questions are based only on information contained in a written passage. Personal opinions or values of the examinee are not to be considered in answering the questions.

Objective 2. The student recognizes explicitly stated main ideas, details, sequences of events and cause and effect relationships. Main emphasis of these questions is on information directly stated in written material. Four types of questions may be asked (I) The main idea is the major point or purpose of the entire passage. (2) Selected detail' are necessary for or relevant to understanding the passage. (3) Sequences of events point to a series of steps or happenings in the passage. (4) Cause and effect describes a kind of relationship between two or more events.

Objective 3: The student interprets instructions (may be in narrative style or in contexts such as forms, lists, steps, non-numerical tables, now charts, labels or applications). Items ask for interpretation of information found in material such as labels and forms, etc. Interpretation of numerical graphs, tables and charts is included in Mathematics Objective 4.

SKILL AREA: INFERENTIAL COMPREHENSION — This skill area involves understanding information which can be determined from written material even though it is not directly stated in the material.

Objective 4: The student recognizes implicitly stated main ideas, details, sequences of events and cause and effect relationships. Questions ask for information about a topic when answers are not directly stated in the material. Question types are similar to those for Objective 2.

Objective 5: The student interprets semantic relationships. Written passages are used to give clues to aid in determining the meaning of words or phrases, based on how they are used in context. Some items ask for identification of appropriate connotations or paraphrases of words or phrases used in context.

APPENDIX D, continued

Objective 6: The student interprets figurative language. Items ask for identification of the meaning of figurative phrases (e.g., metaphors or similes) used in the context of a passage.

Objective 7: The student recognizes propaganda techniques. Items require the student to recognize an underlying, but not directly stated, intent to mislead, misinform or persuade. Names of specific propaganda techniques are not asked.

SKILL AREA: PROBLEM SOLVING — This skill area involves locating, recognizing, interpreting, and evaluating information needed for making decisions or solving problems.

Objective 8: The student locates information in reference sources. Items require selection of appropriate information needed from sources such as an index, table of contents or directory; or selection of appropriate steps to find information.

Objective 9: The student makes generalizations and draws conclusions. Items ask for determining a possible result based on information given in written material. Conclusion items require the student to use multiple pieces of information to determine a specific result which brings closure to a passage. Generalization items require that there be application to other situations points in time, people or events. (9-II)

Objective 10: The student makes predictions and comparisons. Items require the identification of the next most likely event (prediction), or may ask students questions involving comparisons. The latter may involve similarities or differences, or evaluating alternate problem solving solutions presented in a passage.

Objective 11: The student recognizes relevance of data. Items ask for identification or recognition of information necessary to make a decision about a problem presented in a written passage. In some cases, information may be omitted and the student must identify what is missing.

Objective 12: The student recognizes appropriate reference sources. Items ask for identification of the most appropriate reference (e.g., yellow pages, encyclopedia, government agency) to help solve a problem presented in a written passage.

*Source: Georgia Basic Skills Tests Interpretive Guide (GDE, 1987)

APPENDIX E

SKILL AREAS AND OBJECTIVES MEASURED BY THE GEORGIA BASIC SKILLS TEST IN MATHEMATICS*

SKILL AREA: CONCEPT IDENTIFICATION — This skill area concerns the basic vocabulary of mathematics and the interrelationship of different kinds of numbers.

Objective 1: The student translates numerical forms of rational numbers. Items may require matching words to numerals, changing numerals with decimals to percents, changing fractions to percents, changing fractions to decimals or the reverse of any of these operations.

Objective 2: The student orders fractions, decimals or percents. Items give several sets of fractions, decimals or percents and ask for the selection of the set which is arranged in order, from least to greatest or greatest to least.

Objective 3: The student identifies customary or metric units to measure length, area, volume, weight, time and temperature. Items in this indicator present a measurement problem. The student may be asked to select the best type and/ or size of unit. Both metric and customary systems of measurement are used.

Objective 4: The student identifies sets of points using standard names or Cartesian coordinates. Items may ask for identification of standard geometric shapes, including plane figures such as triangles and rectangles, and solid figures such as cubes, cylinders and cones. Other items ask the student to describe the location of points on a graph or map by using a set of letters and numbers.

Objective 5: The student identifies geometric relations and properties. The items in this indicator require recognition of relations between sets of points such as parallel and perpendicular lines, identification of properties such as degrees in a right angle, or identification of transformed shapes.

APPENDIX E, continued

SKILL AREA: COMPONENT OPERATIONS — This skill area involves actions using numbers. The student must be able to add, subtract, multiply, and divide numbers as well as to use units of measurement.

Objective 6: The student applies formulas and proportions. Items require the student to solve word problems by use of an appropriate formula or by setting up and solving a proportion, or may require substituting numbers for variables in a formula.

Objective 7: The student computes with whole numbers, fractions, decimals and percents. Items from this indicator ask the student to add, subtract, multiply, and divide using whole numbers, fractions, or decimals. A few items are percent problems in which the student must find the percentage, the rate, or the hase. There are no word problems for this indicator. However, there are items included which test the students understanding of the mathematical properties of numerical operations (associative, commutative, distributive, identity, and inverse).

Objective 11: The student determines amounts of money. The items in this indicator emphasize the use of money. For example, the student must be able to count money, make change, and find the amount of sales tax.

Objective 9: The student applies customary or metric units of measurement to determine length, area, volume, weight, time, and temperature. There are several basic types of items in this indicator. One type requires changing from one unit of measurement to another (i.e., feet to inches). A second type asks the student to find the perimeter, area, or volume of described figures. Some formulas are not given. Other items require application of units to solve simple problems.

SKILL AREA: PROBLEM SOLVING — This skill area requires the student to select and/or apply the appropriate concepts and/or operations necessary to solve problems.

Objective 10: The student estimates results using rounded numbers, with or without units of measurement. The student is asked to make a reasonable guess to answer a problem rather than working it out. Some items ask for the estimation of the number of units in a drawing or may ask which of several strategies would provide the best estimate.

APPENDIX E, continued

Objective 11: The student selects appropriate operations for a given problem situation. The items in this indicator are word problems. The student is asked to choose the steps needed to solve the problem. The steps may or may not be written as mathematical expressions.

Objective 12: The student solves simple word problems. The student is asked to solve a word problem by using whole numbers, fractions, decimals, or percents. Some problems may require two to four steps or may require making a judgment about whether various solutions are appropriate.

Objective 13: The student organizes data into tables, charts, and graphs. Items for this indicator include a list of data. The student is asked to select the best arrangement of the data. The data is grouped into either tables, charts, or graphs. A few items use data collection as a step in solving a specific problem.

Objective 14: The student interprets data which has been organized. The items from this indicator include data that is in tables, charts or graphs. The student is asked to use this data to answer a question. The answer may be found by simply reading the table, chart or graph, by making an interpretation, by finding the mean or median of the data or by determining the likelihood of a specific event occurring.

*Source: Georgia Basic Skills Tests: Interpretive Guide (GDE, 1987)

APPENDIX F

Correct Answer Rates, by Race, for the Eighth-Grade Reading Test

This table presents the correct answer rates for white examinees (P_W), the t statistic for testing H_O: P_w=0.40, the correct answer rates for black examinees (P_b), the t statistic for testing H_O: P_b=.40, for items of the Eighth-Grade Reading Test. It also presents the difference between the correct answer rate of white and black examinees for each item, and the t statistic for testing H_O:d=.15. Observed correct answer rates less than 0.40 are shown in italics, as are observed differences of greater than 0.15. Statistically significant values of the t statistics are indicated by astericks.

Item #	White Examinees P_W	t statistic for H_O: P_w=.40	Black Examinees P_b	t statistic for H_O: P_b=.40	difference d =(P_w - P_b)	t statistic for H_O: d=.15
1	0.970055	119.101	0.955513	61.2043	0.014543	-13.2011
2	0.982664	158.962	0.969052	74.6429	0.013611	-16.1234
3	0.925138	71.056	0.762089	19.3166	***0.163049***	0.6476
4	0.981875	155.320	0.938104	50.7266	0.043771	-9.4426
5	0.936170	78.104	0.856867	29.6338	0.079304	-4.1891
6	0.972419	124.464	0.938104	50.7266	0.034315	-10.0057
7	0.957447	98.342	0.911025	40.7726	0.046422	-7.5298
8	0.811663	37.493	0.644101	11.5812	***0.167562***	0.7390
9	0.697400	23.053	0.593810	8.9642	0.103589	-1.8434
10	0.781718	32.905	0.622824	10.4432	***0.158894***	0.3662
11	0.967691	114.325	0.897485	37.2562	0.070206	-5.6010
12	0.923562	70.168	0.837524	26.9423	0.086038	-3.5790
13	0.969267	117.450	0.932302	48.1300	0.036965	-9.3609
14	0.944050	84.295	0.843327	27.7047	0.100724	-2.8559
15	0.685579	21.903	0.533849	6.0949	***0.151730***	0.0677
16	0.969267	117.450	0.887814	35.1115	0.081453	-4.6585
17	0.921198	68.884	0.810445	23.7876	0.110753	-2.0831
18	0.887313	54.877	0.752418	18.5478	0.134895	-0.7202
19	0.971631	122.604	0.932302	48.1300	0.039329	-9.2208
20	0.960599	102.609	0.885880	34.7125	0.074719	-5.0101
21	0.973995	128.429	0.951644	58.4147	0.022351	-12.2178

(continued)

* t > 1.645, p < 0.05

** t < -1.645, p < 0.05

Appendix F, continued

Item #	White Examinees P_W	t statistic for H_0: P_w=.40	Black Examinees P_b	t statistic for H_0: P_b=.40	difference $d = (P_w - P_b)$	t statistic for H_0: d=.15
22	0.932230	75.401	0.835590	26.6957	0.096640	-3.0014
23	0.837667	42.263	0.775629	20.4538	0.062039	-4.1721
24	0.911742	64.239	0.845261	27.9669	0.066480	-4.6914
25	0.725768	26.002	0.735010	17.2433	-0.009241	-6.8883
26	0.967691	114.325	0.928433	46.5676	0.039258	-8.9405
27	0.943262	83.621	0.833656	26.4529	0.109607	-2.2907
28	0.976359	135.089	0.938104	50.7266	0.038255	-9.7732
29	0.959811	101.497	0.936170	49.8240	0.023641	-10.4495
30	0.883373	53.625	0.787234	21.4929	0.096139	-2.6736
31	0.907013	62.167	0.814313	24.2029	0.092700	-3.0219
32	0.860520	47.334	0.733075	17.1040	0.127445	-1.0361
33	0.812451	37.625	0.705996	15.2568	0.106455	-1.9051
34	0.881797	53.140	0.733075	17.1040	0.148721	-0.0595
35	0.903073	60.549	0.767892	19.7947	0.135182	-0.7279
36	0.925138	71.056	0.804642	23.1835	0.120496	-1.5566
37	0.929078	73.394	0.891683	35.9382	0.037395	-7.2816
38	0.869976	49.759	0.713733	15.7664	***0.156243***	0.2834
39	0.949567	89.425	0.911025	40.7726	0.038541	-7.9846
40	0.951143	91.041	0.918762	43.1332	0.032381	-8.7354
41	0.970843	120.818	0.916828	42.5146	0.054015	-7.3594
42	0.951143	91.041	0.847195	28.2333	0.103947	-2.7159
43	0.962963	106.149	0.916828	42.5146	0.046135	-7.8312
44	0.944050	84.295	0.899420	37.7184	0.044631	-7.1534
45	0.765957	30.778	0.564797	7.5506	***0.201161***	2.0584*
46	0.906225	61.836	0.841393	27.4466	0.064833	-4.7195
47	0.936958	78.673	0.793037	22.0376	0.143921	-0.3183
48	0.897557	58.429	0.754352	18.6989	0.143205	-0.3271
49	0.959023	100.416	0.911025	40.7726	0.047998	-7.4377
50	0.873128	50.619	0.731141	16.9658	0.141987	-0.3703
51	0.977935	140.099	0.949710	57.1375	0.028226	-11.6331
52	0.973207	126.404	0.932302	48.1300	0.040905	-9.1268

(continued)

* $t > 1.645$, $p < 0.05$

** $t < -1.645$, $p < 0.05$

Appendix F, continued

Item #	White Examinees P_W	t statistic for H_0: P_w=.40	Black Examinees P_b	t statistic for H_0: P_b=.40	difference $d = (P_w - P_b)$	t statistic for H_0: d=.15
53	0.772262	31.609	0.576402	8.1094	***0.195859***	1.8539*
54	0.698188	23.131	0.529981	5.9158	***0.168207***	0.7147
55	0.906225	61.836	0.814313	24.2029	0.091912	-3.0613
56	0.839243	42.583	0.744681	17.9562	0.094563	-2.5440
57	0.914106	65.333	0.773694	20.2866	0.140411	-0.4787
58	0.941686	82.313	0.856867	29.6338	0.084820	-3.8884
59	0.963751	107.403	0.862669	30.5344	0.101082	-3.0506
60	0.736013	27.144	0.582205	8.3920	***0.153808***	0.1523
61	0.829787	40.722	0.589942	8.7724	***0.239845***	3.7300*
62	0.975571	132.764	0.936170	49.8240	0.039401	-9.5330
63	0.946414	86.401	0.889749	35.5199	0.056666	-6.1529
64	0.909377	63.184	0.760155	19.1601	0.149223	-0.0380
65	0.962963	106.149	0.895551	36.8058	0.067412	-5.7072
66	0.799842	35.584	0.686654	14.0379	0.113189	-1.5794
67	0.904649	61.185	0.758220	19.0050	0.146429	-0.1736
68	0.962175	104.934	0.907157	39.6964	0.055018	-6.8560
69	0.962963	106.149	0.918762	43.1332	0.044201	-8.0490
70	0.861308	47.528	0.578337	8.2034	***0.282972***	5.5852*
71	0.845548	43.902	0.649903	11.9009	***0.195644***	1.9571*
72	0.824271	39.696	0.595745	9.0606	***0.228526***	3.2579*
73	0.860520	47.334	0.624758	10.5446	***0.235762***	3.6603*
74	0.799842	35.584	0.727273	16.6925	0.072570	-3.4265
75	0.919622	68.057	0.843327	27.7047	0.076295	-4.1571
76	0.958235	99.365	0.941973	52.6584	0.016262	-11.4055
77	0.957447	98.342	0.885880	34.7125	0.071567	-5.1937
78	0.620961	16.218	0.458414	2.6630	***0.162547***	0.4859
79	0.788810	33.921	0.729207	16.8287	0.059603	-3.9870
80	0.821119	39.127	0.669246	12.9995	***0.151873***	0.0803
81	0.960599	102.609	0.899420	37.7184	0.061179	-6.2010
82	0.780142	32.685	0.508704	4.9393	***0.271438***	4.8786*
83	0.791962	34.386	0.586074	8.5817	***0.205889***	2.2815*

(continued)

* $t > 1.645$, $p < 0.05$

** $t < -1.645$, $p < 0.05$

Appendix F, continued

Item #	White Examinees P_W	t statistic for H_0: P_w=.40	Black Examinees P_b	t statistic for H_0: P_b=.40	difference $d = (P_w - P_b)$	t statistic for H_0: d=.15
84	0.825847	39.985	0.752418	18.5478	0.073429	-3.5154
85	0.926714	71.970	0.796905	22.4109	0.129809	-1.0537
86	0.924350	70.609	0.858801	29.9286	0.065549	-4.9578
87	0.857368	46.573	0.680851	13.6861	***0.176517***	1.1656
88	0.806935	36.712	0.624758	10.5446	***0.182176***	1.3393
89	0.617021	15.897	0.526112	5.7373	0.090909	-2.2837
90	0.959023	100.416	0.891683	35.9382	0.067340	-5.5962
91	0.928290	72.912	0.870406	31.8159	0.057884	-5.5946
92	0.842396	43.234	0.798839	22.6007	0.043556	-5.2180
93	0.947991	87.880	0.891683	35.9382	0.056308	-6.2314
94	0.965327	110.034	0.889749	35.5199	0.075578	-5.0578
95	0.920410	68.467	0.738878	17.5251	***0.181532***	1.5176
96	0.906225	61.836	0.671180	13.1124	***0.235045***	3.8236*
97	0.802994	36.080	0.537718	6.2746	***0.265277***	4.6809*
98	0.893617	57.008	0.839458	27.1925	0.054159	-5.2274
99	0.907801	62.502	0.883946	34.3225	0.023856	-7.7517
100	0.953507	93.611	0.903288	38.6802	0.050219	-6.9816
101	0.905437	61.509	0.831721	26.2135	0.073716	-4.1446
102	0.957447	98.342	0.885880	34.7125	0.071567	-5.1937
103	0.938534	79.842	0.856867	29.6338	0.081668	-4.0606
104	0.925138	71.056	0.847195	28.2333	0.077942	-4.1226
105	0.903073	60.549	0.856867	29.6338	0.046207	-5.9265
106	0.939322	80.443	0.820116	24.8462	0.119206	-1.6930
107	0.925926	71.509	0.893617	36.3666	0.032309	-7.6236
108	0.943262	83.621	0.854932	29.3441	0.088330	-3.6687
109	0.892829	56.733	0.678917	13.5701	***0.213912***	2.8642*
110	0.869976	49.759	0.667311	12.8872	***0.202665***	2.3107*
111	0.917258	66.859	0.831721	26.2135	0.085536	-3.5427
112	0.847124	44.243	0.669246	12.9995	***0.177878***	1.2097
113	0.836879	42.105	0.667311	12.8872	***0.169568***	0.8437
114	0.578408	12.865	***0.307544***	-4.5511**	***0.270865***	4.9137*
115	0.944050	84.295	0.829787	25.9776	0.114263	-2.0123

(continued)

* $t > 1.645$, $p < 0.05$

** $t < -1.645$, $p < 0.05$

APPENDIX G

Correct Answer Rates for the Eighth-Grade Reading Test

This table presents the correct answer rates for the entire sample of 1807 examinees (P), and the t statistic for testing H_0: P=0.40, for the Eighth-Grade Reading Test. There were no observed correct answer rates of less than 0.40 for the entire sample for this test.

Item #	All Examinees P	All Examinees t statistic for H_0: P=.40	Item #	All Examinees P	All Examinees t statistic for H_0: P=.40
1	0.966242	133.239	29	0.952961	110.990
2	0.978417	169.155	30	0.855562	55.073
3	0.877698	61.962	31	0.881018	63.138
4	0.968456	138.216	32	0.824018	47.319
5	0.912562	77.112	33	0.783619	39.591
6	0.962369	125.584	34	0.839513	50.886
7	0.944660	101.234	35	0.863863	57.483
8	0.762590	36.214	36	0.889873	66.501
9	0.665191	23.881	37	0.916436	79.308
10	0.737133	32.548	38	0.825125	47.561
11	0.946873	103.620	39	0.936912	93.851
12	0.898727	70.252	40	0.942446	98.980
13	0.959048	119.881	41	0.954621	113.243
14	0.915329	78.666	42	0.920863	81.997
15	0.643055	21.559	43	0.949087	106.153
16	0.946320	103.010	44	0.930825	88.900
17	0.889319	66.281	45	0.707803	28.763
18	0.848368	53.126	46	0.887659	65.627
19	0.959048	119.881	47	0.895407	68.795
20	0.939126	95.823	48	0.856115	55.228
21	0.967903	136.925	49	0.945213	101.817
22	0.904261	72.832	50	0.832319	49.179
23	0.819037	46.256	51	0.969563	140.900
24	0.893193	67.858	52	0.961815	124.584
25	0.728832	31.434	53	0.715551	29.724
26	0.956281	115.618	54	0.651356	22.415
27	0.912009	76.810	55	0.880465	62.939
28	0.965689	132.069	56	0.812396	44.892

(continued)

Appendix G, continued

Item #	All Examinees P	t statistic for H_0: P=.40
57	0.872717	60.275
58	0.916436	79.308
59	0.933592	91.071
60	0.692861	26.979
61	0.759823	35.795
62	0.964029	128.717
63	0.930271	88.480
64	0.866076	58.158
65	0.942999	99.532
66	0.768677	37.156
67	0.862203	56.986
68	0.945767	102.409
69	0.949640	106.811
70	0.779745	38.941
71	0.787493	40.254
72	0.756503	35.300
73	0.791920	41.030
74	0.779192	38.850
75	0.897067	69.516
76	0.952961	110.990
77	0.936912	93.851
78	0.574986	15.043
79	0.771998	37.681
80	0.777532	38.576
81	0.942999	99.532
82	0.703376	28.226
83	0.733259	32.023
84	0.805755	43.586
85	0.889319	66.281
86	0.904815	73.101
87	0.807416	43.907
88	0.753735	34.892
89	0.591035	16.513
90	0.938572	95.321
91	0.912562	77.112
92	0.829552	48.546
93	0.932485	90.187
94	0.943553	100.091
95	0.868290	58.848
96	0.838406	50.617
97	0.728279	31.361
98	0.877144	61.770
99	0.900387	71.006
100	0.939126	95.823
101	0.883785	64.152
102	0.936912	93.851
103	0.914776	78.350
104	0.902601	72.037
105	0.890426	66.724
106	0.904261	72.832
107	0.916436	79.308
108	0.917543	79.961
109	0.830105	48.672
110	0.810183	44.450
111	0.892640	67.628
112	0.794134	41.425
113	0.788600	40.446
114	0.500830	8.570
115	0.910349	75.918
116	0.869397	59.199
117	0.916990	79.633
118	0.757609	35.464
119	0.680133	25.524
120	0.726619	31.143
121	0.832872	49.307
122	0.730492	31.654

APPENDIX H

Correct Answer Rates, by Race, for the Eighth-Grade Mathematics Test

This table presents the correct answer rates for white examinees (P_W), the t statistic for testing H_O: P_w=0.40; the correct answer rates for black examinees (P_b), the t statistic for testing H_O: P_b=.40, for items of the Eighth-Grade Mathematics test. It also presents the difference between the correct answer rate of white and black examinees for each item, and the t statistic for testing H_O:d=.15. Observed correct answer rates less than 0.40 are shown in italics, as are observed differences of greater than 0.15. Statistically significant values of the t statistics are indicated by astericks.

Item #	White Examinees P_W	t statistic for H_O: P_w=.40	Black Examinees P_b	t statistic for H_O: P_b=.40	difference $d = (P_w - P_b)$	t statistic for H_O: d=.15
1	0.822695	39.410	0.669246	12.9995	*0.153449*	0.1479
2	0.872340	50.402	0.796905	22.4109	0.075435	-3.7214
3	0.847124	44.243	0.640232	11.3704	*0.206892*	2.4292*
4	0.824271	39.696	0.595745	9.0606	*0.228526*	3.2579*
5	0.603625	14.824	*0.392650*	-0.3419	*0.210975*	2.3901*
6	0.855004	46.016	0.659574	12.4435	*0.195430*	1.9679*
7	0.855004	46.016	0.789168	21.6725	0.065836	-4.1057
8	0.939322	80.443	0.760155	19.1601	*0.179168*	1.4615
9	0.962175	104.934	0.916828	42.5146	0.045347	-7.8777
10	0.611505	15.452	0.411992	0.5535	*0.199513*	1.9319*
11	0.734437	26.966	0.531915	6.0053	*0.202522*	2.0821*
12	0.462569	4.469	*0.228240*	-9.2963**	0.234329	3.6376
13	0.804571	36.331	0.508704	4.9393	*0.295866*	5.9139*
14	0.825059	39.840	0.733075	17.1040	0:091984	-2.6128
15	0.710796	24.410	0.533849	6.0949	*0.176947*	1.0615
16	0.951143	91.041	0.905222	39.1811	0.045920	-7.3065
17	0.632782	17.196	*0.382979*	-0.7954	*0.249803*	3.9413*
18	0.691095	22.434	0.491296	4.1483	*0.199799*	1.9492*
19	0.734437	26.966	0.615087	10.0413	0.119350	-1.2383
20	0.490938	6.477	*0.257253*	-7.4181**	*0.233684*	3.5132*

(continued)

* $t > 1.645$, $p < 0.05$

** $t < -1.645$, $p < 0.05$

Appendix H, continued

Item #	White Examinees P_W	t statistic for H_0: P_w=.40	Black Examinees P_b	t statistic for H_0: P_b=.40	difference $d = (P_w - P_b)$	t statistic for H_0: d=.15
21	0.630418	16.998	0.450677	2.3136	***0.179741***	1.1546
22	0.918046	67.253	0.835590	26.6957	0.082456	-3.7434
23	0.880221	52.664	0.818182	24.6290	0.062039	-4.5640
24	0.687155	22.054	0.537718	6.2746	0.149438	-0.0220
25	0.855004	46.016	0.558994	7.2741	***0.296010***	6.0863*
26	0.643814	18.130	0.439072	1.7884	***0.204742***	2.1338*
27	0.598897	14.451	0.441006	1.8760	***0.157891***	0.3055
28	0.914894	65.707	0.735010	17.2433	***0.179884***	1.4265
29	0.687155	22.054	0.560928	7.3661	0.126227	-0.9347
30	0.883373	53.625	0.837524	26.9423	0.045849	-5.6076
31	0.984240	167.038	0.965184	70.0355	0.019056	-14.8880
32	0.825059	39.840	0.640232	11.3704	***0.184827***	1.4714
33	0.611505	15.452	0.452611	2.4010	***0.158894***	0.3442
34	0.750985	28.901	0.483559	3.7982	***0.267426***	4.6730
35	0.905437	61.509	0.806576	23.3824	0.098861	-2.6591
36	0.511426	7.938	***0.361702***	-1.8106**	0.149724	-0.0109
37	0.729708	26.436	0.562863	7.4582	***0.166846***	0.6699
38	0.893617	57.008	0.806576	23.3824	0.087041	-3.2412
39	0.757289	29.676	0.495164	4.3236	***0.262125***	4.4693*
40	0.899133	59.019	0.856867	29.6338	0.042267	-6.1266
41	0.860520	47.334	0.802708	22.9869	0.057812	-4.6004
42	0.750985	28.901	0.603482	9.4490	0.147503	-0.1010
43	0.672971	20.720	0.609284	9.7436	0.063686	-3.4255
44	0.601261	14.637	0.526112	5.7373	0.075149	-2.8869
45	0.970843	120.818	0.974855	83.4039	-0.004012	-18.4304
46	0.937746	79.252	0.882012	33.9411	0.055735	-5.9892
47	0.866036	48.721	0.762089	19.3166	0.103947	-2.1884
48	0.890465	55.922	0.688588	14.1565	***0.201877***	2.3376*
49	0.877069	51.736	0.721470	16.2900	***0.155599***	0.2570
50	0.966115	111.416	0.852998	29.0593	0.113117	-2.2495
51	0.783294	33.128	0.630561	10.8511	***0.152733***	0.1130
52	0.780142	32.685	0.665377	12.7755	0.114765	-1.4800

(continued)

* $t > 1.645$, $p < 0.05$

** $t < -1.645$, $p < 0.05$

Appendix H, continued

Item #	White Examinees P_W	t statistic for H_0: P_w=.40	Black Examinees P_b	t statistic for H_0: P_b=.40	difference $d = (P_w - P_b)$	t statistic for H_0: d=.15
53	0.884949	54.119	0.798839	22.6007	0.086109	-3.2281
54	0.820331	38.987	0.802708	22.9869	0.017623	-6.4353
55	0.561072	11.558	0.524178	5.6482	0.036894	-4.3452
56	0.944050	84.295	0.891683	35.9382	0.052368	-6.4540
57	0.516942	8.333	***0.394584***	-0.2517	0.122358	-1.0761
58	0.918834	67.652	0.864603	30.8457	0.054230	-5.6661
59	0.492514	6.589	***0.326886***	-3.5407**	***0.165628***	0.6259
60	0.982664	158.962	0.970986	77.2758	0.011677	-16.7703
61	0.552403	10.914	0.460348	2.7504	0.092055	-2.2279
62	0.845548	43.902	0.750484	18.3981	0.095064	-2.5451
63	0.846336	44.072	0.622824	10.4432	***0.223512***	3.1125*
64	0.619385	16.090	0.444874	2.0512	***0.174511***	0.9508
65	0.866036	48.721	0.669246	12.9995	***0.196791***	2.0510*
66	0.838455	42.423	0.570600	7.8290	***0.267856***	4.8867*
67	0.888101	55.135	0.717602	16.0263	***0.170499***	0.9444
68	0.682427	21.603	***0.251451***	-7.7778**	***0.430976***	12.1398*
69	0.528763	9.185	***0.396518***	-0.1617	0.132244	-0.6910
70	0.564224	11.793	0.425532	1.1730	0.138692	-0.4376
71	0.836879	42.105	0.671180	13.1124	***0.165700***	0.6785
72	0.968479	115.859	0.920696	43.7728	0.047783	-7.9437
73	0.859732	47.142	0.636364	11.1614	***0.223368***	3.1469*
74	0.938534	79.842	0.907157	39.6964	0.031378	-8.2108
75	0.641450	17.928	0.541586	6.4548	0.099864	-1.9478
76	0.829787	40.722	0.481625	3.7108	***0.348163***	8.1223*
77	0.833727	41.481	0.717602	16.0263	0.116126	-1.5118
78	0.698188	23.131	0.549323	6.8172	0.148865	-0.0447
79	0.981087	151.905	0.930368	47.3335	0.050720	-8.3852
80	0.557132	11.264	0.499033	4.4992	0.058099	-3.5266
81	***0.397951***	-0.149	***0.172147***	-13.7105**	***0.225804***	3.5148*
82	0.708432	24.166	0.597679	9.1572	0.110753	-1.5650
83	0.566588	11.971	0.411992	0.5535	***0.154596***	0.1785
84	0.732861	26.788	0.568665	7.7360	***0.164195***	0.5657

(continued)

* $t > 1.645$, $p < 0.05$

** $t < -1.645$, $p < 0.05$

Appendix H, continued

Item #	White Examinees P_W	t statistic for H_0: P_w=.40	Black Examinees P_b	t statistic for H_0: P_b=.40	difference $d = (P_w - P_b)$	t statistic for H_0: d=.15
85	0.644602	18.198	***0.388781***	-0.5228	***0.255821***	4.1790*
86	0.444444	3.185	***0.208897***	-10.6785**	***0.235547***	3.7697*
87	0.995272	309.001	0.980658	95.7705	0.014614	-21.2814
88	0.976359	135.089	0.907157	39.6964	0.069203	-5.9985
89	0.802994	36.080	0.661509	12.5536	0.141486	-0.3602
90	0.743105	27.963	0.493230	4.2360	***0.249875***	3.9635*
91	0.948779	88.644	0.961315	66.1194	-0.012537	-15.4694
92	0.910165	63.531	0.738878	17.5251	***0.171287***	1.0167
93	0.736013	27.144	0.471954	3.2741	***0.264059***	4.5220*
94	0.683215	21.678	***0.288201***	-5.6071**	***0.395014***	10.2783*
95	0.636722	17.527	0.468085	3.0995	***0.168637***	0.7227
96	0.899133	59.019	0.889749	35.5199	0.009385	-8.6933
97	0.661151	19.647	***0.313346***	-4.2436**	***0.347804***	8.1183*
98	0.882585	53.382	0.769826	19.9571	0.112759	-1.806
99	0.704492	23.764	0.609284	9.7436	0.095207	-2.1908
100	0.907801	62.502	0.825919	25.5157	0.081883	-3.6692
101	0.687943	22.130	0.444874	2.0512	***0.243069***	3.6563*
102	0.984240	167.038	0.980658	95.7705	0.003582	-20.9184
103	0.808511	36.970	0.644101	11.5812	***0.164410***	0.6055
104	0.775414	32.034	0.644101	11.5812	0.131313	-0.7749
105	0.741529	27.779	0.738878	17.5251	0.002651	-6.4304
106	0.713948	24.738	0.485493	3.8857	***0.228455***	3.0888*
107	0.781718	32.905	0.665377	12.7755	0.116341	-1.4147
108	0.690307	22.358	0.576402	8.1094	0.113905	-1.4248
109	0.991332	227.151	0.955513	61.2043	0.035819	-12.0925
110	0.844760	43.734	0.593810	8.9642	***0.250949***	4.2251*
111	0.731284	26.612	0.516441	5.2929	***0.214844***	2.5653*
112	0.851064	45.115	0.692456	14.3958	***0.158607***	0.3801
113	0.715524	24.903	0.489362	4.0607	***0.226162***	2.9993*
114	0.946414	86.401	0.914894	41.9156	0.031521	-8.5752
115	0.452325	3.744	***0.317215***	-4.0407**	0.135110	-0.6004
116	0.770686	31.399	0.682785	13.8027	0.087900	-2.6263

(continued)

* $t > 1.645$, $p < 0.05$

** $t < -1.645$, $p < 0.05$

Appendix H, continued

Item #	White Examinees P_W	t statistic for H_0: P_w=.40	Black Examinees P_b	t statistic for H_0: P_b=.40	difference $d = (P_w - P_b)$	t statistic for H_0: d=.15
117	0.841608	43.070	0.684720	13.9199	***0.156888***	0.3010
118	0.776202	32.141	0.508704	4.9393	***0.267498***	4.7137*
119	0.824271	39.696	0.671180	13.1124	***0.153091***	0.1328
120	0.629630	16.933	0.305609	-4.6545	***0.324020***	7.1331*
121	0.887313	54.877	0.762089	19.3166	0.125224	-1.1945
122	0.589441	13.713	***0.301741***	-4.8626**	***0.287700***	5.6255*
123	0.799842	35.584	0.680851	13.6861	0.118991	-1.3254
124	0.715524	24.903	0.535783	6.1847	***0.179741***	1.1733
125	0.412924	0.935	***0.355899***	-2.0923**	0.057024	-3.6884
126	0.837667	42.263	0.721470	16.2900	0.116197	-1.5168
127	0.588652	13.652	0.444874	2.0512	0.143778	-0.2405

* $t > 1.645$, $p < 0.05$

** $t < -1.645$, $p < 0.05$

APPENDIX I

Correct Answer Rates for the Eighth-Grade Mathematics Test

This table presents the correct answer rates for the entire sample of 1807 examinees (P), and the t statistic for testing H_0: P=0.40, for items of the Eighth-Grade Mathematics test. Observed correct answer rates less than 0.40 are shown in italics. Statistically significant values of the t statistic are indicated by an asterick.

Item #	All Examinees P	All Examinees t statistic for H_0: P=.40
1	0.780299	39.033
2	0.849474	53.417
3	0.788046	40.350
4	0.758163	35.546
5	0.542335	12.141
6	0.799115	42.333
7	0.835639	49.955
8	0.887659	65.627
9	0.948533	105.505
10	0.555617	13.309
11	0.676812	25.153
12	*0.395683*	-0.375
13	0.720531	30.355
14	0.798561	42.231
15	0.660764	23.406
16	0.938572	95.321
17	0.562258	13.899
18	0.630327	20.277
19	0.701716	28.026
20	0.423354	2.009
21	0.578860	15.395
22	0.894300	68.324
23	0.862756	57.151
24	0.645268	21.786
25	0.769231	37.243
26	0.586608	16.104
27	0.553957	13.162
28	0.862756	57.151
29	0.652463	22.531
30	0.870504	59.554
31	0.978971	171.482
32	0.772551	37.769
33	0.566685	14.295
34	0.675152	24.968
35	0.877144	61.770
36	0.469286	5.900
37	0.682900	25.835
38	0.868290	58.848
39	0.682900	25.835
40	0.886552	65.199
41	0.843940	51.985
42	0.709463	28.967
43	0.655230	22.821
44	0.579967	15.496
45	0.971776	146.722
46	0.921417	82.348
47	0.836193	50.086
48	0.832872	49.307
49	0.832872	49.307
50	0.933592	91.071

(continued)

* $t < -1.645$, $p < 0.05$

Appendix I, continued

Item #	All Examinees P	All Examinees t statistic for H_0: P=.40
51	0.739900	32.927
52	0.748201	34.092
53	0.858329	55.856
54	0.815717	45.566
55	0.550636	12.869
56	0.929718	88.065
57	0.482014	6.975
58	0.904261	72.832
59	0.446043	3.936
60	0.978971	171.482
61	0.524626	10.605
62	0.817930	46.024
63	0.781959	39.311
64	0.570559	14.643
65	0.809629	44.341
66	0.760376	35.879
67	0.838406	50.617
68	0.558937	13.604
69	0.490869	7.725
70	0.526287	10.748
71	0.788046	40.350
72	0.952407	110.265
73	0.795241	41.625
74	0.928611	87.249
75	0.612618	18.548
76	0.729386	31.507
77	0.798561	42.231
78	0.655783	22.879
79	0.965136	130.926
80	0.539568	11.900
81	***0.333149***	-6.027*

Item #	All Examinees P	All Examinees t statistic for H_0: P=.40
82	0.674045	24.846
83	0.524073	10.558
84	0.687327	26.340
85	0.570005	14.593
86	***0.380188***	-1.734*
87	0.991145	268.166
88	0.955728	114.812
89	0.762590	36.214
90	0.671832	24.603
91	0.952407	110.265
92	0.860542	56.496
93	0.660210	23.347
94	0.571112	14.693
95	0.587714	16.206
96	0.896514	69.274
97	0.561705	13.850
98	0.850028	53.564
99	0.677919	25.276
100	0.884892	64.566
101	0.619258	19.190
102	0.982844	190.751
103	0.762037	36.130
104	0.737687	32.623
105	0.741007	33.080
106	0.648589	22.128
107	0.749862	34.330
108	0.657443	23.054
109	0.980631	179.041
110	0.772551	37.769
111	0.669618	24.361
112	0.805755	43.586

* $t < -1.645$, $p < 0.05$

APPENDIX J

Correct Answer Rates, by Race, for the Basic Skills Test in Reading

This table presents the correct answer rates for white examinees (P_W), the t statistic for testing H_O: P_W=0.40, the correct answer rates for black examinees (P_b), the t statistic for testing H_O: P_b=.40, for items of the Basic Skills Test in Reading. It also presents the difference between the correct answer rate of white and black examinees for each item, and the t statistic for testing H_O:d=.15. Observed correct answer rates less than 0.40 are shown in italics, as are observed differences of greater than 0.15. Statistically significant values of the t statistics are indicated by astericks.

Item #	White Examinees P_W	t statistic for H_O: P_W=.40	Black Examinees P_b	t statistic for H_O: P_b=.40	difference $d = (P_W - P_b)$	t statistic for H_O: d=.15
1	0.847912	44.415	0.781431	20.9653	0.066480	-4.0151
2	0.828999	40.573	0.735010	17.2433	0.093990	-2.5322
3	0.786446	33.578	0.595745	9.0606	*0.190701*	1.6628*
4	0.758865	29.873	0.518375	5.3816	*0.240490*	3.6105*
5	0.853428	45.652	0.715667	15.8959	0.137761	-0.5512
6	0.810875	37.361	0.622824	10.4432	*0.188051*	1.5852
7	0.849488	44.762	0.767892	19.7947	0.081596	-3.2381
8	0.957447	98.342	0.949710	57.1375	0.007737	-12.7401
9	0.961387	103.754	0.940039	51.6703	0.021348	-10.9313
10	0.827423	40.277	0.626692	10.6464	*0.200731*	2.1324*
11	0.829787	40.722	0.646035	11.6873	*0.183752*	1.4333
12	0.946414	86.401	0.862669	30.5344	0.083745	-4.0352
13	0.657998	19.366	*0.359768*	-1.9042**	*0.298231*	5.9346*
14	0.859732	47.142	0.717602	16.0263	0.142131	-0.3563
15	0.929866	73.884	0.779497	20.7931	*0.150369*	0.0188
16	0.905437	61.509	0.882012	33.9411	0.023426	-7.7144
17	0.899133	59.019	0.800774	22.7927	0.098360	-2.6467
18	0.719464	25.321	0.547389	6.7263	*0.172075*	0.8731
19	0.716312	24.986	0.504836	4.7630	*0.211476*	2.4212*
20	0.933806	76.455	0.837524	26.9423	0.096282	-3.0389
21	0.941686	82.313	0.851064	28.7794	0.090623	-3.4931

(continued)

* $t > 1.645$, $p < 0.05$

** $t < -1.645$, $p < 0.05$

Appendix J, continued

Item #	White Examinees P_W	t statistic for H_0: P_w=.40	Black Examinees P_b	t statistic for H_0: P_b=.40	difference $d = (P_w - P_b)$	t statistic for H_0: d=.15
22	0.936170	78.104	0.905222	39.1811	0.030948	-8.1498
23	0.947991	87.880	0.887814	35.1115	0.060176	-5.8984
24	0.731284	26.612	0.566731	7.6432	***0.164553***	0.5794
25	0.945626	85.684	0.845261	27.9669	0.100365	-2.8946
26	0.806147	36.585	0.667311	12.8872	0.138835	-0.4746
27	0.943262	83.621	0.833656	26.4529	0.109607	-2.2907
28	0.821907	39.268	0.702128	15.0069	0.119779	-1.3243
29	0.931442	74.887	0.883946	34.3225	0.047496	-6.4937
30	0.902285	60.236	0.833656	26.4529	0.068630	-4.4241
31	0.947203	87.132	0.893617	36.3666	0.053585	-6.4466
32	0.935382	77.545	0.827853	25.7450	0.107529	-2.3600
33	0.903073	60.549	0.682785	13.8027	***0.220288***	3.1793*
34	0.895981	57.852	0.777563	20.6226	0.118418	-1.5622
35	0.914894	65.707	0.858801	29.9286	0.056093	-5.4545
36	0.809299	37.100	0.582205	8.3920	***0.227094***	3.1655*
37	0.942474	82.961	0.837524	26.9423	0.104950	-2.5733
38	0.919622	68.057	0.852998	29.0593	0.066624	-4.8033
39	0.907801	62.502	0.787234	21.4929	0.120567	-1.4892
40	0.911742	64.239	0.796905	22.4109	0.114836	-1.8107
41	0.788022	33.807	0.578337	8.2034	***0.209686***	2.4279*
42	0.557920	11.323	***0.259188***	-7.2997**	***0.298732***	6.2482*
43	0.914894	65.707	0.876209	32.8453	0.038685	-6.7543
44	0.658786	19.436	0.516441	5.2929	0.142345	-0.2977
45	0.794326	34.740	0.649903	11.9009	0.144423	-0.2336
46	0.891253	56.190	0.765957	19.6339	0.125295	-1.2000
47	0.923562	70.168	0.827853	25.7450	0.095709	-2.9802
48	0.757289	29.676	0.595745	9.0606	***0.161545***	0.4668
49	0.810875	37.361	0.574468	8.0157	***0.236407***	3.5432*
50	0.922774	69.734	0.843327	27.7047	0.079447	-3.9926
51	0.903861	60.865	0.748549	18.2496	***0.155312***	0.2552
52	0.900709	59.621	0.812379	23.9939	0.088330	-3.2239
53	0.879433	52.429	0.822050	25.0664	0.057382	-4.8338

(continued)

* $t > 1.645$, $p < 0.05$

** $t < -1.645$, $p < 0.05$

Appendix J, continued

Item #	White Examinees P_W	t statistic for H_0: P_w=.40	Black Examinees P_b	t statistic for H_0: P_b=.40	difference $d = (P_w - P_b)$	t statistic for H_0: d=.15
54	0.754925	29.383	0.638298	11.2657	0.116627	-1.3701
55	0.869188	49.548	0.814313	24.2029	0.054875	-4.8625
56	0.929866	73.884	0.868472	31.4863	0.061394	-5.3646
57	0.933018	75.924	0.914894	41.9156	0.018125	-9.3207
58	0.900709	59.621	0.812379	23.9939	0.088330	-3.2239
59	0.796690	35.098	0.736944	17.3837	0.059746	-4.0225
60	0.861308	47.528	0.671180	13.1124	***0.190128***	1.7565*
61	0.887313	54.877	0.767892	19.7947	0.119421	-1.4846
62	0.966903	112.845	0.930368	47.3335	0.036536	-9.2401
63	0.884949	54.119	0.777563	20.6226	0.107386	-2.0906
64	0.799054	35.462	0.535783	6.1847	***0.263271***	4.5913*
65	0.952719	92.734	0.901354	38.1928	0.051365	-6.8417
66	0.904649	61.185	0.818182	24.6290	0.086468	-3.3657
67	0.813239	37.758	0.686654	14.0379	0.126585	-1.0107
68	0.655634	19.157	0.557060	7.1824	0.098574	-2.0075
69	0.783294	33.128	0.634429	11.0576	0.148865	-0.0470
70	0.858156	46.761	0.690522	14.2758	***0.167634***	0.7807
71	0.758077	29.774	0.618955	10.2415	0.139122	-0.4435
72	0.747045	28.428	0.578337	8.2034	***0.168708***	0.7504
73	0.805359	36.457	0.584139	8.4867	***0.221219***	2.9212*
74	0.929078	73.394	0.752418	18.5478	***0.176660***	1.3119
75	0.970843	120.818	0.880077	33.5680	0.090766	-3.9327
76	0.875493	51.284	0.783366	21.1394	0.092127	-2.8414
77	0.643814	18.130	0.611219	9.8425	0.032595	-4.6358
78	0.717100	25.070	0.595745	9.0606	0.121355	-1.1442
79	0.940110	81.054	0.889749	35.5199	0.050362	-6.5064
80	0.873128	50.619	0.839458	27.1925	0.033670	-6.2311
81	0.805359	36.457	0.582205	8.3920	***0.223153***	2.9989*
82	0.929078	73.394	0.899420	37.7184	0.029658	-7.9824
83	0.866824	48.925	0.762089	19.3166	0.104735	-2.1520
84	0.810875	37.361	0.653772	12.1164	***0.157103***	0.3003
85	0.859732	47.142	0.746615	18.1023	0.113117	-1.7164

(continued)

* $t > 1.645$, $p < 0.05$

** $t < -1.645$, $p < 0.05$

Appendix J, continued

Item #	White Examinees P_W	t statistic for H_0: P_w=.40	Black Examinees P_b	t statistic for H_0: P_b=.40	difference $d = (P_w - P_b)$	t statistic for H_0: d=.15
86	0.895193	57.568	0.812379	23.9939	0.082814	-3.4958
87	0.938534	79.842	0.816248	24.4146	0.122287	-1.5115
88	0.838455	42.423	0.682785	13.8027	***0.155670***	0.2471
89	0.955083	95.431	0.920696	43.7728	0.034386	-8.7312
90	0.848700	44.588	0.731141	16.9658	0.117559	-1.4773
91	0.753349	29.189	0.624758	10.5446	0.128591	-0.8734
92	0.851852	45.292	0.733075	17.1040	0.118776	-1.4270
93	0.827423	40.277	0.653772	12.1164	***0.173651***	1.0073
94	0.948779	88.644	0.847195	28.2333	0.101583	-2.8470
95	0.970843	120.818	0.940039	51.6703	0.030805	-10.3920
96	0.907801	62.502	0.854932	29.3441	0.052869	-5.5493
97	0.853428	45.652	0.692456	14.3958	***0.160971***	0.4852
98	0.817179	38.434	0.653772	12.1164	***0.163407***	0.5683
99	0.835303	41.791	0.752418	18.5478	0.082886	-3.0974
100	0.835303	41.791	0.659574	12.4435	***0.175729***	1.1035
101	0.901497	59.927	0.680851	13.6861	***0.220646***	3.1878*
102	0.916470	66.470	0.822050	25.0664	0.094419	-2.9973
103	0.825847	39.985	0.725339	16.5574	0.100509	-2.2144
104	0.772262	31.609	0.733075	17.1040	0.039186	-4.8693
105	0.842396	43.234	0.688588	14.1565	***0.153808***	0.1669

* $t > 1.645$, $p < 0.05$

** $t < -1.645$, $p < 0.05$

APPENDIX K

Correct Answer Rates for the Basic Skills Test in Reading

This table presents the correct answer rates for the entire sample of 1807 examinees (P), and the t statistic for testing H_0: P=0.40, for items of the Basic Skills Test in Reading. There were no items on this test for which the observed correct answer rates were less than 0.40.

Item #	All Examinees P	All Examinees t statistic for H_0: P=.40	Item #	All Examinees P	All Examinees t statistic for H_0: P=.40
1	0.827892	48.173	28	0.786940	40.159
2	0.801328	42.745	29	0.918650	80.627
3	0.731046	31.727	30	0.883232	63.946
4	0.688987	26.530	31	0.931378	89.324
5	0.812950	45.003	32	0.904261	72.832
6	0.757609	35.464	33	0.840620	51.157
7	0.826785	47.927	34	0.862756	57.151
8	0.955728	114.812	35	0.899834	70.753
9	0.954621	113.243	36	0.745434	33.699
10	0.770338	37.417	37	0.910902	76.213
11	0.775872	38.305	38	0.901494	71.518
12	0.921417	82.348	39	0.872717	60.275
13	0.573879	14.943	40	0.878805	62.349
14	0.818484	46.140	41	0.727726	31.288
15	0.885999	64.986	42	0.471500	6.087
16	0.898727	70.252	43	0.903708	72.565
17	0.871057	59.732	44	0.617045	18.975
18	0.669618	24.361	45	0.751522	34.570
19	0.653016	22.589	46	0.856115	55.228
20	0.905921	73.646	47	0.895407	68.795
21	0.915883	78.985	48	0.711677	29.240
22	0.927504	86.451	49	0.743774	33.466
23	0.930271	88.480	50	0.900387	71.006
24	0.683453	25.898	51	0.859435	56.174
25	0.916990	79.633	52	0.875484	61.201
26	0.765357	36.639	53	0.862203	56.986
27	0.912562	77.112	54	0.722192	30.568

(continued)

Appendix K, continued

Item #	All Examinees P	All Examinees t statistic for H_0: P=.40
55	0.853348	54.461
56	0.912562	77.112
57	0.927504	86.451
58	0.874931	61.014
59	0.779192	38.850
60	0.805202	43.480
61	0.854455	54.765
62	0.956834	116.439
63	0.855008	54.919
64	0.725512	30.999
65	0.937465	94.335
66	0.879358	62.544
67	0.776425	38.395
68	0.627006	19.948
69	0.739347	32.851
70	0.809629	44.341
71	0.717211	29.933
72	0.698395	27.630
73	0.742114	33.234
74	0.878251	62.155
75	0.944660	101.234
76	0.847261	52.837
77	0.633647	20.608
78	0.680686	25.586
79	0.925844	85.285
80	0.862756	57.151
81	0.742667	33.311
82	0.919756	81.305
83	0.837853	50.483
84	0.767017	36.896
85	0.827338	48.050
86	0.871057	59.732
87	0.903154	72.300
88	0.793580	41.326
89	0.945767	102.409
90	0.814610	45.340
91	0.716657	29.863
92	0.817377	45.909
93	0.779745	38.941
94	0.917543	79.961
95	0.962369	125.584
96	0.892086	67.400
97	0.807969	44.015
98	0.768124	37.069
99	0.811843	44.781
100	0.785280	39.874
101	0.838406	50.617
102	0.889319	66.281
103	0.795241	41.625
104	0.762590	36.214
105	0.798008	42.129

APPENDIX L

Correct Answer Rates, by Race, for the Basic Skills Test in Mathematics

This table presents the correct answer rates for white examinees (P_W), the t statistic for testing H_0: P_W=0.40; the correct answer rates for black examinees (P_b), the t statistic for testing H_0: P_b=.40, for items of the Basic Skills Test in Mathematics. It also presents the difference between the correct answer rate of white and black examinees for each item, and the t statistic for testing H_0:d=.15. Observed correct answer rates less than 0.40 are shown in italics, as are observed differences of greater than 0.15. Statistically significant values of the t statistics are indicated by astericks.

Item #	White Examinees P_W	t statistic for H_0: P_W=.40	Black Examinees P_b	t statistic for H_0: P_b=.40	difference d =(P_W - P_b)	t statistic for H_0: d=.15
1	0.940110	81.054	0.905222	39.1811	0.034888	-7.9308
2	0.950355	90.224	0.858801	29.9286	0.091554	-3.5424
3	0.716312	24.986	0.578337	8.2034	0.137976	-0.4780
4	0.871552	50.186	0.789168	21.6725	0.082384	-3.3363
5	0.822695	39.410	0.696325	14.6380	0.126370	-1.0315
6	0.836091	41.948	0.504836	4.7630	*0.331256*	7.4462*
7	0.895193	57.568	0.758220	19.0050	0.136973	-0.6288
8	0.918046	67.253	0.736944	17.3837	*0.181102*	1.4912
9	0.750985	28.901	0.626692	10.6464	0.124293	-1.0487
10	0.673759	20.792	*0.363636*	-1.7171	*0.310122*	6.4213*
11	0.482269	5.863	*0.377176*	-1.0697	0.105093	-1.7584
12	0.950355	90.224	0.897485	37.2562	0.052869	-6.6164
13	0.866824	48.925	0.800774	22.7927	0.066051	-4.1963
14	0.915682	66.086	0.723404	16.4232	*0.192277*	1.9959*
15	0.750985	28.901	0.651838	12.0084	0.099147	-2.0984
16	0.465721	4.692	*0.280464*	-6.0445**	*0.185257*	1.4548
17	0.787234	33.692	0.671180	13.1124	0.116054	-1.4347
18	0.965327	110.034	0.920696	43.7728	0.044631	-8.1319
19	0.506698	7.600	*0.386847*	-0.6135	0.119851	-1.1764
20	0.757289	29.676	0.634429	11.0576	0.122860	-1.1132

(continued)

* $t > 1.645$, $p < 0.05$

** $t < -1.645$, $p < 0.05$

Appendix L, continued

Item #	White Examinees P_W	t statistic for H_o: P_w=.40	Black Examinees P_b	t statistic for H_o: P_b=.40	difference $d = (P_w - P_b)$	t statistic for H_o: d=.15
21	0.884161	53.871	0.849130	28.5041	0.035031	-6.3380
22	0.916470	66.470	0.901354	38.1928	0.015116	-8.8425
23	0.586288	13.469	0.464217	2.9249	0.122072	-1.0763
24	0.602049	14.699	***0.361702***	-1.8106**	***0.240347***	3.5814*
25	0.872340	50.402	0.682785	13.8027	***0.189555***	1.7557*
26	0.534279	9.586	***0.208897***	-10.6785**	***0.325382***	7.7170*
27	0.810875	37.361	0.750484	18.3981	0.060391	-4.0738
28	0.783294	33.128	0.673114	13.2259	0.110180	-1.6823
29	0.716312	24.986	0.441006	1.8760	***0.275306***	4.9609*
30	0.953507	93.611	0.852998	29.0593	0.100509	-2.9685
31	0.664303	19.930	0.400387	0.0179	***0.263916***	4.4989*
32	0.898345	58.722	0.785300	21.3152	0.113045	-1.8506
33	0.810087	37.230	0.638298	11.2657	***0.171789***	0.9136
34	0.904649	61.185	0.825919	25.5157	0.078731	-3.8278
35	0.861308	47.528	0.748549	18.2496	0.112759	-1.7383
36	0.698188	23.131	0.545455	6.6357	***0.152733***	0.1075
37	0.643814	18.130	0.415861	0.7310	***0.227953***	3.0538*
38	0.804571	36.331	0.676983	13.4548	0.127588	-0.9576
39	0.871552	50.186	0.787234	21.4929	0.084318	-3.2324
40	0.889677	55.657	0.823985	25.2895	0.065692	-4.4528
41	0.693459	22.665	0.502901	4.6750	***0.190558***	1.5882
42	0.715524	24.903	***0.369439***	-1.4383	***0.346085***	7.9263*
43	0.852640	45.472	0.572534	7.9222	***0.280106***	5.4334*
44	0.963751	107.403	0.938104	50.7266	0.025647	-10.5068
45	0.813239	37.758	0.663443	12.6643	0.149796	-0.0087
46	0.649330	18.606	***0.334623***	-3.1473**	***0.314707***	6.6630**
47	0.742317	27.871	0.611219	9.8425	0.131098	-0.7644
48	0.721828	25.575	0.584139	8.4867	0.137689	-0.4908
49	0.947991	87.880	0.901354	38.1928	0.046637	-7.1125
50	0.565012	11.852	***0.352031***	-2.2815*	***0.212981***	2.4976*
51	0.754137	29.286	0.547389	6.7263	***0.206748***	2.2674*
52	0.838455	42.423	0.707930	15.3829	0.130525	-0.8645

(continued)

* $t > 1.645$, $p < 0.05$

** $t < -1.645$, $p < 0.05$

Appendix L, continued

Item #	White Examinees P_W	t statistic for H_0: P_w=.40	Black Examinees P_b	t statistic for H_0: P_b=.40	difference $d = (P_w - P_b)$	t statistic for H_0: d=.15
53	0.542159	10.160	0.417795	0.8196	0.124364	-0.9925
54	0.707644	24.085	***0.394584***	-0.2517	***0.313060***	6.5166*
55	0.684003	21.753	0.460348	2.7504	***0.223655***	2.8848*
56	0.958235	99.365	0.860735	30.2288	0.097500	-3.2320
57	0.919622	68.057	0.810445	23.7876	0.109177	-2.1636
58	0.683215	21.678	0.558994	7.2741	0.124221	-1.0124
59	0.818755	38.709	0.593810	8.9642	***0.224944***	3.1000*
60	0.825847	39.985	0.721470	16.2900	0.104377	-2.0345
61	0.747045	28.428	0.595745	9.0606	***0.151300***	0.0524
62	0.978723	142.807	0.959381	64.3684	0.019342	-13.6261
63	0.621749	16.283	***0.355899***	-2.0923**	***0.265850***	4.6166*
64	0.933018	75.924	0.845261	27.9669	0.087757	-3.5772
65	0.721040	25.490	0.493230	4.2360	***0.227810***	3.0684*
66	0.874704	51.060	0.777563	20.6226	0.097142	-2.5743
67	0.729708	26.436	0.588008	8.6769	0.141701	-0.3320
68	0.509850	7.825	***0.330754***	-3.3433**	***0.179096***	1.1628
69	0.909377	63.184	0.823985	25.2895	0.085393	-3.4730
70	0.762017	30.271	0.524178	5.6482	***0.237839***	3.5097*
71	0.915682	66.086	0.787234	21.4929	0.128448	-1.0977
72	0.745469	28.241	0.533849	6.0949	***0.211620***	2.4513*
73	0.517730	8.390	***0.396518***	-0.1617	0.121212	-1.1200
74	0.809299	37.100	0.661509	12.5536	0.147790	-0.0938
75	0.516942	8.333	0.442940	1.9636	0.074002	-2.9249
76	0.756501	29.578	0.557060	7.1824	***0.199441***	1.9801*
77	0.784870	33.352	0.595745	9.0606	***0.189125***	1.5974
78	0.884949	54.119	0.603482	9.4490	***0.281467***	5.6364*
79	0.737589	27.324	0.508704	4.9393	***0.228885***	3.1255*
80	0.433412	2.401	***0.222437***	-9.6985**	***0.210975***	2.6514*
81	0.639086	17.727	***0.398453***	-0.0718	***0.240633***	3.5648*
82	0.758865	29.873	0.700193	14.8832	0.058672	-3.8902
83	0.937746	79.252	0.804642	23.1835	0.133104	-0.9022
84	0.722616	25.660	0.522244	5.5592	***0.200373***	1.9886*

(continued)

* $t > 1.645$, $p < 0.05$

** $t < -1.645$, $p < 0.05$

Appendix L, continued

Item #	White Examinees P_W	t statistic for H_0: P_w=.40	Black Examinees P_b	t statistic for H_0: P_b=.40	difference $d = (P_w - P_b)$	t statistic for H_0: d=.15
85	0.878645	52.196	0.676983	13.4548	***0.201662***	2.2924*
86	0.701340	23.446	0.497099	4.4114	***0.204241***	2.1280*
87	0.778566	32.466	0.545455	6.6357	***0.233111***	3.3474*
88	0.710796	24.410	0.576402	8.1094	0.134394	-0.6192
89	0.724980	25.916	0.537718	6.2746	***0.187263***	1.4741
90	0.799054	35.462	0.636364	11.1614	***0.162691***	0.5292
91	0.898345	58.722	0.694391	14.5166	***0.203954***	2.4543*
92	0.646966	18.401	0.402321	0.1075	***0.244645***	3.7234*
93	0.973207	126.404	0.951644	58.4147	0.021563	-12.2602
94	0.731284	26.612	0.448743	2.2262	***0.282542***	5.2623*
95	0.568952	12.148	***0.355899***	-2.0923**	***0.213052***	2.4969*
96	0.560284	11.499	***0.396518***	-0.1617	***0.163765***	0.5366
97	0.733649	26.877	0.524178	5.6482	***0.209471***	2.3554*
98	0.905437	61.509	0.804642	23.1835	0.100795	-2.5506
99	0.941686	82.313	0.874275	32.4952	0.067412	-5.1585
100	0.680063	21.380	0.514507	5.2044	***0.165556***	0.6075
101	0.502758	7.318	***0.375242***	-1.1615	0.127516	-0.8809
102	0.848700	44.588	0.675048	13.3400	***0.173651***	1.0309
103	0.657998	19.366	0.406190	0.2863	***0.251809***	4.0090*
104	0.672183	20.647	0.437137	1.7007	***0.235045***	3.3342*
105	0.828211	40.425	0.653772	12.1164	***0.174439***	1.0413
106	0.812451	37.625	0.659574	12.4435	***0.152876***	0.1221
107	0.843972	43.566	0.640232	11.3704	***0.203740***	2.2910*
108	0.731284	26.612	0.553191	6.9994	***0.178093***	1.1157
109	0.814027	37.892	0.562863	7.4582	***0.251164***	4.1430*
110	0.929866	73.884	0.899420	37.7184	0.030446	-7.9394
111	0.884949	54.119	0.800774	22.7927	0.084175	-3.3354
112	0.940110	81.054	0.847195	28.2333	0.092915	-3.3220

* $t > 1.645$, $p < 0.05$

** $t < -1.645$, $p < 0.05$

APPENDIX M

Correct Answer Rates for the Basic Skills Test in Mathematics

This table presents the correct answer rates for the entire sample of 1807 examinees (P), and the t statistic for testing H_0: P=0.40, for items of the Basic Skills Test in Mathematics. Observed correct answer rates less than 0.40 are shown in italics. Statistically significant values of the t statistic are indicated by an asterick.

Item #	All Examinees P	All Examinees t statistic for H_0: P=.40
1	0.929718	88.065
2	0.923630	83.787
3	0.675706	25.030
4	0.848368	53.126
5	0.785833	39.968
6	0.740454	33.004
7	0.855562	55.073
8	0.864416	57.650
9	0.717764	30.003
10	0.584947	15.951
11	0.455451	4.732
12	0.935252	92.436
13	0.847261	52.837
14	0.860542	56.496
15	0.723298	30.711
16	0.412839	1.108
17	0.754289	34.973
18	0.951301	108.849
19	0.474820	6.367
20	0.720531	30.355
21	0.874931	61.014
22	0.912562	77.112
23	0.551743	12.967
24	0.533481	11.371
25	0.817930	46.024
26	0.441063	3.515
27	0.793580	41.326
28	0.753735	34.892
29	0.636967	20.942
30	0.924184	84.155
31	0.587161	16.155
32	0.864970	57.819
33	0.760376	35.879
34	0.881018	63.138
35	0.828445	48.297
36	0.654123	22.705
37	0.579967	15.496
38	0.766464	36.810
39	0.847814	52.981
40	0.869950	59.376
41	0.639181	21.166
42	0.618152	19.082
43	0.772551	37.769
44	0.956281	115.618
45	0.770891	37.505
46	0.558384	13.554
47	0.705036	28.426
48	0.682900	25.835
49	0.934698	91.975
50	0.505257	8.947
51	0.695628	27.303
52	0.799668	42.435
53	0.505257	8.947
54	0.618152	19.082

(continued)

Appendix M, continued

Item #	All Examinees P	All Examinees t statistic for H_0: P=.40
55	0.619812	19.243
56	0.930825	88.900
57	0.888766	66.061
58	0.648589	22.128
59	0.754289	34.973
60	0.796347	41.825
61	0.702822	28.159
62	0.973437	151.548
63	0.545102	12.383
64	0.907582	74.481
65	0.655230	22.821
66	0.846154	52.550
67	0.687881	26.403
68	0.460985	5.199
69	0.883785	64.152
70	0.692861	26.979
71	0.877698	61.962
72	0.684007	25.961
73	0.483121	7.069
74	0.767017	36.896
75	0.496956	8.241
76	0.699502	27.762
77	0.730492	31.654
78	0.803542	43.163
79	0.673492	24.785
80	***0.373547***	-2.324*
81	0.570005	14.593
82	0.741561	33.157
83	0.899281	70.502
84	0.665744	23.940
85	0.819037	46.256
86	0.641395	21.390
87	0.711677	29.240
88	0.672939	24.724
89	0.670725	24.481
90	0.750969	34.490
91	0.839513	50.886
92	0.578307	15.344
93	0.966242	133.239
94	0.650249	22.300
95	0.509131	9.277
96	0.514665	9.750
97	0.674599	24.907
98	0.877144	61.770
99	0.921417	82.348
100	0.634200	20.664
101	0.466519	5.666
102	0.798008	42.129
103	0.585501	16.002
104	0.603763	17.704
105	0.778639	38.758
106	0.769231	37.243
107	0.785833	39.968
108	0.681240	25.648
109	0.740454	33.004
110	0.920310	81.649
111	0.861096	56.659
112	0.914222	78.036

* $t < -1.645$, $p < 0.05$

APPENDIX N

Calculation of σ_{dTot} for Hypothesis Tests for Research Question 1A

As described in Chapter III, the standard deviation of the differences between black and white examinees' total scores on random samples of n items selected from N possible items is given by the formula:

$$\sigma_{dTot} = \sqrt{n\,\sigma^2_{dp}\,(1- n/N)}$$

where σ^2_{dp} is the variance of the differences between the proportions of correct answers for black and white examinees across all N items on the original test, calculated using the formula:

$$\sigma^2_{dp} = \frac{\sum_{i=1}^{N} (dp_i-\mu_{dp})^2}{(N-1)}$$

[Note the use of (N-1), rather than N, in the denominator of the formula (Jaeger, 1984, p. 42; Cochran, 1977, p. 23).]

The difference between the proportions of correct answers for black and white examinees for each of the N items of original tests are listed in Appendices F, H, J, and L. For each of the synthetic tests, the table below shows the number of items in the synthetic test (n), the number of items in the corresponding original test (N), the variance of the differences between the proportions of correct answers for black and white examinees across all N items on the original test, and the standard deviation of the differences between black and white examinees' total scores on random samples of n items selected from N possible items.

Synthetic Test	N	n	σ^2_{dp}	σ_{dTot}
GRP- Eighth-Grade Reading	122	93	0.004151225	0.30293
GRP-Eighth-Grade Mathematics	127	61	0.007714109	0.49451
GRP-BST Reading	105	73	0.004053869	0.30031
GRP-BST Mathematics	112	53	0.006004700	0.40945

(continued)

APPENDIX N
continued

Synthetic Test	N	n	σ^2_{dp}	σ_{dTot}
GRS- Eighth-Grade Reading	122	108	0.004151225	0.22682
GRS-Eighth-Grade Mathematics	127	86	0.007714109	0.46279
GRS-BST Reading	105	90	0.004053869	0.22830
GRS-BST Mathematics	112	71	0.006004700	0.39506

APPENDIX O

Correspondence of Test Items to Test Objectives

The tables below show the correspondence of items of the four original standardized tests used in this study to the objectives of the tests. Items which were excluded from the GRS- and GRP- synthetic tests are indicated.

Correspondence of Items of Georgia Eighth-Grade Criterion-Referenced Test (Edition 3) in Reading to Test Skill Areas and Objectives*

Skill Area/Objective	Corresponding Items
Literal Comprehension	
Objective 1	35, 36, 47, 57, 68, 90, 115
Objective 2	1, 5, 7, 11, 32, 41, 42, 43, 44, 45[2], 48, 84, 86, 113[2], 116, 122
Objective 3	30, 31, 33, 34, 53 ,56, 63, 93, 94, 118, 119, 120[1]
Inferential Comprehension	
Objective 4	6, 10[1] , 14, 22, 23, 49, 54[1] , 58, 67, 71[2], 72[2], 73[2], 83[2], 87[1], 89, 98, 99, 100, 110[2], 111, 112[1], 117
Objective 5	15[1], 46, 51, 55, 62, 85, 91, 121
Objective 6	19, 25, 28,37, 61[2], 96[2], 105
Objective 7	59, 60[1], 69, 77, 88[1], 92
Problem Solving	
Objective 8	12, 13, 18, 38[1], 39, 40, 70[2], 74, 75, 76, 82[2], 108
Objective 9	4, 8[1], 16, 24, 29, 64, 66, 95[1], 97[2], 107, 114[2]
Objective 10	2, 9, 20, 21, 26, 79, 81, 101, 102, 103, 104, 106
Objective 11	3[1], 17, 27, 50, 78[1], 80[1], 109[2]

* Information is not available on which objectives items 52 & 65 are intended to address.

[1] Not included in GRP– Eighth-Grade Reading Test .

[2] Not included in GRS- Eighth-Grade Reading Test or GRP– Eighth-Grade Reading Test .

APPENDIX O, continued

Correspondence of Items of Georgia Eighth-Grade Criterion-Referenced Test (Edition 3) in Mathematics to Test Skill Areas and Objectives*

Skill Area/Objective	Corresponding Items
Concept Identification	
Objective 1	4^2, 15^1, 23, 46, 58, 73^2, 77, 78, 93^2, 98, 117^1
Objective 2	7, 29, 33^1, 39^2, 51^1, 53, 63^2, 70, 71^1, 76^2, 83^1, 99, 108, 118^2, 126
Objective 3	1^1, 26^2, 31, 43, 50, 67^1, 87, 95^1, 107
Objective 4	5^2, 10^2, 14, 28^1, 41, 47, 49^1, 57^1, 60, 74, 97^2, 100, 109, 115^2
Component Operations	
Objective 5	35, 66^2, 88, 110^2
Objective 6	6^2, 11^2, 17^2, 27^1, 37^1, 42, 45, 54, 65^2, 84^1, 89, 91, 106^2, 111^2, 114, 122^2
Objective 7	2, 12^2, 16, 18^2, 24, 30, 34^2, 36^2, 48^2, 55, 61, 62, 69^1, 81^2, 86^2, 94^2, 101^2, 103^1, 104, 112^1, 124^1, 12^25
Problem Solving	
Objective 8	25^2, 32^1, 52, 68^2, 85^2, 90^2, 92^1, 123
Objective 9	6, 8^1, 21^1, 64^1, 80, 113^2, 116, 121
Objective 10	13^2, 44, 79, 96, 127
Objective 11	20^2, 38, 56, 59^2, 102, 105, 120^2
Objective 12	3^2, 40, 82, 119^1

* Information is not available on which objectives items 19, 22, 72, & 75 are intended to address.

[1] Not included in GRP– Eighth-Grade Mathematics Test .

[2] Not included in GRS- Eighth-Grade Reading Test or GRP– Eighth-Grade Mathematics Test .

APPENDIX O, continued

Correspondence of Items of Georgia Basic Skills Test (Form 01, Edition 13) in Mathematics to Test Skill Areas and Objectives*

Skill Area/Objective	Corresponding Items
Concept Identification	
Objective 1	4, 13, 15, 20, 41[1], 49, 58, 60, 66, 75, 76[2], 82, 100[1]
Objective 2	48, 68[2], 73[1], 92[2], 102[1]
Objective 3	10[2], 16[2], 35, 40, 47, 50[2], 56, 78[2], 91[2], 109[2]
Objective 4	18, 28, 84[2]
Objective 5	7, 45, 55[2], 64, 103[2]
Component Operations	
Objective 6	6[2], 17, 37[2], 51[2], 59[2], 69
Objective 7	1, 5, 9, 11[1], 19[2], 21, 23, 27, 38, 52, 62, 72[2], 77[1], 79[2], 89[1], 93, 101[1], 107[2], 110
Objective 8	22, 26[2], 34, 61[1], 105[1]
Objective 9	3, 36[1], 42[2], 43[2], 74, 81[2], 95[2], 96[2], 104[2]
Problem Solving	
Objective 10	44, 90[1], 97[2], 108[1]
Objective 11	8[1], 31[2], 33[1], 46[2], 83, 98
Objective 12	12, 24[2], 29[2], 54[2], 65[2], 70[2], 85[2], 88
Objective 13	2, 57, 86[2], 99, 106[1], 112
Objective 14	25[2], 30, 32, 39, 53, 63[2], 67, 71, 80[2], 87[2], 94[2], 111

* Information is not available on which objective item 14 is intended to address.

[1] Not included in GRP-- BST Mathematics Test .

[2] Not included in GRS- BST Mathematics Test or GRP-- BST Mathematics Test .

APPENDIX O, continued

Correspondence of Items of Georgia Basic Skills Test (Form 01, Edition 13) in Reading to Test Skill Areas and Objectives

Skill Area/Objective	Corresponding Items
Literal Comprehension	
Objective 1	31, 54, 81[2], 92, 101[2]
Objective 2	9, 12, 13[2], 25, 35, 48[2], 52, 53, 55, 57, 65, 66, 69, 94, 95
Objective 3	3[2], 4[2], 32, 42[2], 43, 61, 90, 102, 103
Inferential Comprehension	
Objective 4	1, 2, 17, 18[1], 19[2], 24[1], 34, 56, 63, 68, 77, 79, 82, 97[1], 99,
Objective 5	14, 20, 44, 45, 58, 70[1], 75, 93[1]
Objective 6	5, 21, 30, 36[2], 50, 96
Objective 7	8, 28, 39, 40, 86, 104, 105[1]
Problem Solving	
Objective 8	6[1], 7, 27, 67, 85, 89
Objective 9	10[2], 22, 26, 37, 39, 71, 74[1], 83, 87, 91, 98[1]
Objective 10	23, 46, 49[2], 59, 62, 72[1], 73[2], 78, 84[1]
Objective 11	11[1], 16, 29, 47, 64[2], 80, 88[1]
Objective 12	15[1], 33[2], 41[2], 51[1], 60[2], 76, 100[1]

[1] Not included in GRP-- BST Reading Test .

[2] Not included in GRS- BST Reading Test or GRP-- BST Reading Test .

APPENDIX P

χ^2 Goodness of Fit Tests for Content Representativeness of Synthetic Tests Composed Solely of Type I Items

For each comparison of original and synthetic test, this table presents the expected number of items per objective (found by multiplying the proportion of items addressing the objective on the original test by the ratio of the lengths of the synthetic and original tests), the number of items observed per objective on the synthetic test, the value of the χ^2 statistic, and the probability of the χ^2 statistic. (Note: the "expected" values below have been rounded: four decimal places were carried in the calculations of the χ^2 statistics.)

Goodness of Fit Test for the GRS-Eighth-Grade Reading Test

	Objectives:										
	1	2	3	4	5	6	7	8	9	10	11
Expected (from Eighth-Grade Reading)	6.19	14.17	10.62	19.47	7.08	6.20	5.31	10.62	9.74	10.62	6.20
Observed on GRS-Eighth Grade Reading	7	15	11	17	8	5	6	10	9	12	6

$\chi^2 = 1.1980$, d.f. = 10, prob. > 0.99

Goodness of Fit Test for the GRP-Eighth-Grade Reading Test

	Objectives:										
	1	2	3	4	5	6	7	8	9	10	11
Expected (from Eighth-Grade Reading)	5.34	12.20	9.15	16.77	6.10	5.33	4.57	9.15	8.39	9.15	5.34
Observed on GRP-Eighth Grade Reading	7	14	10	13	7	5	4	9	7	12	3

$\chi^2 = 4.0825$, d.f. = 10, prob. > 0.95

APPENDIX P, continued

Goodness of Fit Test for the GRS-Eighth-Grade Mathematics Test

	Objectives:											
	1	2	3	4	5	6	7	8	9	10	11	12
Expected (from Eighth-Gr. Mathematics)	7.45	10.16	6.09	9.48	2.71	10.83	14.90	5.42	5.42	3.39	4.74	2.71
Observed on GRS-Eighth Gr. Mathematics	8	11	8	10	2	9	12	4	7	4	4	3

$\chi^2 = 2.8861$, d.f. = 11, prob. > 0.99

Goodness of Fit Test for the GRP-Eighth-Grade Mathematics Test

	Objectives:											
	1	2	3	4	5	6	7	8	9	10	11	12
Expected (from Eighth-Gr. Mathematics)	5.28	7.20	4.32	6.72	1.92	7.69	10.57	3.84	3.84	2.40	3.36	1.92
Observed on GRP-Eighth Gr. Mathematics	6	7	5	7	2	6	8	2	4	4	4	2

$\chi^2 = 3.2947$, d.f. = 11, prob. > 0.98

Goodness of Fit Test for the GRP-BST Mathematics Test

	Objectives:													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Expected (from BST Mathematics)	6.15	2.37	4.73	1.42	2.37	2.84	8.99	2.37	4.26	1.89	2.84	3.79	2.84	5.68
Observed on GRP-BST Mathematics	10	1	4	2	3	2	11	2	2	1	2	2	4	7

$\chi^2 = 7.9616$, d.f. = 13, prob. > 0.80

APPENDIX P, continued

Goodness of Fit Test for the GRS-BST Mathematics Test

	Objectives:													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Expected (from BST Mathematics)	8.24	3.17	6.34	1.90	3.16	3.80	12.04	3.17	5.71	2.54	3.80	5.07	3.80	7.61
Observed on GRS-BST Mathematics	12	3	4	2	3	2	16	4	4	3	4	2	5	7

$\chi^2 = 7.8625$, d.f. = 13, prob. > 0.80

Goodness of Fit Test for the GRP-BST Reading Test

	Objectives:											
	1	2	3	4	5	6	7	8	9	10	11	12
Expected (from BST-Reading)	3.48	10.43	6.26	10.43	5.56	4.17	4.87	4.17	7.65	6.25	4.87	4.87
Observed on GRP-BST-Reading	3	13	6	11	6	5	6	5	8	5	4	1

$\chi^2 = 4.8641$, d.f. = 11, prob. > 0.90

Goodness of Fit Test for the GRS-BST Reading Test

	Objectives:											
	1	2	3	4	5	6	7	8	9	10	11	12
Expected (from BST-Reading)	4.29	12.86	7.71	12.86	6.86	5.14	6.00	5.14	9.43	7.71	6.00	6.00
Observed on GRS-BST-Reading	3	14	6	14	8	5	7	6	10	7	6	4

$\chi^2 = 2.2412$, d.f. = 11, prob. > 0.99

APPENDIX Q

Proportions of Items Addressing Test Objectives, for Original and Synthetic Tests

The tables below show the number and proportion of items addressing each objective of each of the synthetic tests, and the corresponding original test. A 95% confidence interval is also given for the proportion of items on the synthetic tests addressing each objective.

Table 1
Number and Proportion of Items of Eighth-Grade Reading Test and GRS-Eighth-Grade Reading Test Addressing Each Test Objective

	Original Test		Synthetic Test		
Skill Area/Objective	Number of Items	Proportion of Items	Number of Items	Proportion of Items	95%C.I
Literal Comprehension					
Objective 1	7	0.057	7	0.065	[0.029, 0.139]
Objective 2	16	0.131	15	0.139	[0.087, 0.235]
Objective 3	12	0.098	11	0.102	[0.056, 0.188]
Inferential Comprehension					
Objective 4	22	0.180	17	0.157	[0.102, 0.258]
Objective 5	8	0.066	8	0.074	[0.035, 0.152]
Objective 6	7	0.057	5	0.046	[0.016, 0.113]
Objective 7	6	0.049	6	0.056	[0.022, 0.126]
Problem Solving					
Objective 8	12	0.098	10	0.093	[0.049, 0.176]
Objective 9	11	0.090	9	0.083	[0.042, 0.164]
Objective 10	12	0.098	12	0.111	[0.064, 0.200]
Objective 11	7	0.057	6	0.056	[0.022, 0.126]
Unknown Objective	2	0.016	2	0.019	[0.002, 0.070]
Total Number of Items	122		108		

APPENDIX Q, continued

Table 2
Number and Proportion of Items of Eighth-Grade Reading Test and GRP-Eighth-Grade Reading Test Addressing Each Test Objective

	Original Test		Synthetic Test		
Skill Area/Objective	Number of Items	Proportion of Items	Number of Items	Proportion of Items	95%C.I
Literal Comprehension					
Objective 1	7	0.057	7	0.075	[0.031, 0.149]
Objective 2	16	0.131	14	0.151	[0.085, 0.240]
Objective 3	12	0.098	10	0.108	[0.053, 0.189]
Inferential Comprehension					
Objective 4	22	0.180	13	0.140	[0.077, 0.227]
Objective 5	8	0.066	7	0.075	[0.031, 0.149]
Objective 6	7	0.057	5	0.054	[0.018, 0.121]
Objective 7	6	0.049	4	0.043	[0.012, 0.106]
Problem Solving					
Objective 8	12	0.098	9	0.097	[0.045, 0.176]
Objective 9	11	0.090	7	0.075	[0.031, 0.149]
Objective 10	12	0.098	12	0.129	[0.069, 0.215]
Objective 11	7	0.057	3	0.032	[0.007, 0.091]
Unknown Objective	2	0.016	2	0.022	[0.003, 0.076]
Total Number of Items	122		93		

APPENDIX Q, continued

Table 3
Number and Proportion of Items of Eighth-Grade Mathematics Test and GRS-Eighth-Grade Mathematics Test Addressing Each Test Objective

	Original Test		Synthetic Test		
Skill Area/Objective	Number of Items	Proportion of Items	Number of Items	Proportion of Items	95%C.I
Concept Identification					
Objective 1	11	0.087	8	0.093	[0.041, 0.175]
Objective 2	15	0.118	11	0.128	[0.066, 0.217]
Objective 3	9	0.071	8	0.093	[0.041, 0.175]
Objective 4	14	0.110	10	0.116	[0.057, 0.203]
Component Operations					
Objective 5	4	0.031	2	0.023	[0.003, 0.081]
Objective 6	16	0.126	9	0.105	[0.049, 0.189]
Objective 7	22	0.173	12	0.140	[0.074, 0.231]
Problem Solving					
Objective 8	8	0.063	4	0.047	[0.013, 0.115]
Objective 9	8	0.063	7	0.081	[0.033, 0.161]
Objective 10	5	0.039	4	0.047	[0.013, 0.115]
Objective 11	7	0.055	4	0.047	[0.013, 0.115]
Objective 12	4	0.031	3	0.035	[0.007, 0.099]
Unknown Objective	4	0.031	4	0.047	[0.013, 0.115]
Total Number of Items	127		86		

APPENDIX Q, continued

Table 4
Number and Proportion of Items of Eighth-Grade Mathematics Test and GRP-Eighth-Grade Mathematics Test Addressing Each Test Objective

	Original Test		Synthetic Test		
Skill Area/Objective	Number of Items	Proportion of Items	Number of Items	Proportion of Items	95%C.I
Concept Identification					
Objective 1	11	0.087	6	0.098	[0.037, 0.202]
Objective 2	15	0.118	7	0.115	[0.048, 0.222]
Objective 3	9	0.071	5	0.082	[0.027, 0.181]
Objective 4	14	0.110	7	0.115	[0.048, 0.222]
Component Operations					
Objective 5	4	0.031	2	0.033	[0.004, 0.113]
Objective 6	16	0.126	6	0.098	[0.037, 0.202]
Objective 7	22	0.173	8	0.131	[0.058, 0.242]
Problem Solving					
Objective 8	8	0.063	2	0.033	[0.004, 0.113]
Objective 9	8	0.063	4	0.066	[0.018, 0.159]
Objective 10	5	0.039	4	0.066	[0.018, 0.159]
Objective 11	7	0.055	6	0.066	[0.037, 0.202]
Objective 12	4	0.031	2	0.033	[0.004, 0.113]
Unknown Objective	4	0.031	4	0.066	[0.018, 0.159]
Total Number of Items	127		61		

APPENDIX Q, continued

Table 5
Number and Proportion of Items of Basic Skills Test in Reading and GRS-Basic Skills Test in Reading Addressing Each Test Objective

	Original Test		Synthetic Test		
Skill Area/Objective	Number of Items	Proportion of Items	Number of Items	Proportion of Items	95%C.I
Literal Comprehension					
Objective 1	5	0.048	3	0.033	[0.007, 0.094]
Objective 2	15	0.143	14	0.156	[0.088, 0.247]
Objective 3	9	0.086	6	0.067	[0.025, 0.139]
Inferential Comprehension					
Objective 4	15	0.143	14	0.156	[0.088, 0.247]
Objective 5	8	0.076	8	0.089	[0.039, 0.168]
Objective 6	6	0.057	5	0.056	[0.018, 0.125]
Objective 7	7	0.067	7	0.078	[0.032, 0.154]
Problem Solving					
Objective 8	6	0.057	6	0.067	[0.025, 0.139]
Objective 9	11	0.105	10	0.111	[0.055, 0.195]
Objective 10	9	0.086	7	0.078	[0.032, 0.154]
Objective 11	7	0.067	6	0.067	[0.025, 0.139]
Objective 12	7	0.067	4	0.044	[0.012, 0.110]
Total Number of Items	105		90		

APPENDIX Q, continued

Table 6
Number and Proportion of Items of Basic Skills Test in Reading and GRP-Basic Skills Test in Reading Addressing Each Test Objective

	Original Test		Synthetic Test		
Skill Area/Objective	Number of Items	Proportion of Items	Number of Items	Proportion of Items	95%C.I
Literal Comprehension					
Objective 1	5	0.048	3	0.041	[0.009, 0.115]
Objective 2	15	0.143	13	0.178	[0.098, 0.285]
Objective 3	9	0.086	6	0.082	[0.031, 0.170]
Inferential Comprehension					
Objective 4	15	0.143	11	0.151	[0.078, 0.254]
Objective 5	8	0.076	6	0.082	[0.031, 0.170]
Objective 6	6	0.057	5	0.068	[0.023, 0.153]
Objective 7	7	0.067	6	0.082	[0.031, 0.170]
Problem Solving					
Objective 8	6	0.057	5	0.068	[0.023, 0.153]
Objective 9	11	0.105	8	0.110	[0.049, 0.205]
Objective 10	9	0.086	5	0.068	[0.023, 0.153]
Objective 11	7	0.067	4	0.055	[0.015, 0.134]
Objective 12	7	0.067	1	0.014	[0.0003,0.074]
Total Number of Items	105		73		

APPENDIX Q, continued

Table 7
Number and Proportion of Items of Basic Skills Test in Mathematics and GRS-Basic Skills Test in Mathematics Addressing Each Test Objective

	Original Test		Synthetic Test		
Skill Area/Objective	Number of Items	Proportion of Items	Number of Items	Proportion of Items	95%C.I
Concept Identification					
Objective 1	13	0.116	12	0.169	[0.091, 0.277]
Objective 2	5	0.045	3	0.042	[0.009, 0.119]
Objective 3	10	0.089	4	0.056	[0.016, 0.138]
Objective 4	3	0.027	2	0.028	[0.003, 0.098]
Objective 5	5	0.045	3	0.042	[0.009, 0.119]
Component Operations					
Objective 6	6	0.055	2	0.028	[0.003, 0.098]
Objective 7	19	0.170	16	0.225	[0.135, 0.340]
Objective 8	5	0.045	4	0.056	[0.016, 0.138]
Objective 9	9	0.080	4	0.056	[0.016, 0.138]
Problem Solving					
Objective 10	4	0.036	3	0.042	[0.009, 0.119]
Objective 11	6	0.055	4	0.056	[0.016, 0.138]
Objective 12	8	0.071	2	0.028	[0.003, 0.098]
Objective 13	6	0.055	5	0.070	[0.023, 0.157]
Objective 14	12	0.107	7	0.099	[0.041, 0.193]
Unknown Objective	1	0.009	0	0.000	[0.000, 0.051]
Total Number of Items	112		71		

APPENDIX Q, continued

Table 8
Number and Proportion of Items of Basic Skills Test in Mathematics and GRP-Basic Skills Test in Mathematics Addressing Each Test Objective

	Original Test		Synthetic Test		
Skill Area/Objective	Number of Items	Proportion of Items	Number of Items	Proportion of Items	95%C.I
Concept Identification					
Objective 1	13	0.116	10	0.189	[0.094, 0.320]
Objective 2	5	0.045	1	0.019	[0.001, 0.101]
Objective 3	10	0.089	4	0.075	[0.021, 0.182]
Objective 4	3	0.027	2	0.038	[0.005, 0.130]
Objective 5	5	0.045	3	0.057	[0.012, 0.157]
Component Operations					
Objective 6	6	0.055	2	0.038	[0.005, 0.130]
Objective 7	19	0.170	11	0.208	[0.108, 0.341]
Objective 8	5	0.045	2	0.038	[0.005, 0.130]
Objective 9	9	0.080	2	0.038	[0.005, 0.130]
Problem Solving					
Objective 10	4	0.036	1	0.019	[0.001, 0.101]
Objective 11	6	0.055	2	0.038	[0.005, 0.130]
Objective 12	8	0.071	2	0.038	[0.005, 0.130]
Objective 13	6	0.055	4	0.075	[0.021, 0.182]
Objective 14	12	0.107	7	0.132	[0.055, 0.253]
Unknown Objective	1	0.009	0	0.000	[0.000, 0.067]
Total Number of Items	112		53		

Appendix R

Item-Total Correlations and Corresponding Values of Fisher's Z, for Items of Original and Synthetic Tests

This appendix lists the item-total correlation and corresponding Fisher Z for the items of each original and synthetic test. The values for the original tests are presented first, followed by those for the GRP- synthetic tests, then the GRS- synthetic tests.

Item-Total Correlations and Corresponding Values of Fisher's Z, for Items of the Eighth-Grade Reading Test

Item	r_{pb}	Fisher's Z	Item	r_{pb}	Fisher's Z
1	0.11441	0.114913	26	0.38188	0.402259
2	0.11742	0.117964	27	0.27261	0.279681
3	0.42188	0.449977	28	0.33531	0.348799
4	0.28979	0.298337	29	0.32616	0.338525
5	0.35467	0.370776	30	0.31020	0.320767
6	0.22338	0.227211	31	0.25211	0.257665
7	0.26584	0.272382	32	0.48137	0.524766
8	0.28801	0.296395	33	0.32953	0.342301
9	0.30354	0.313414	34	0.48560	0.530286
10	0.33204	0.345119	35	0.45599	0.492237
11	0.30767	0.317970	36	0.46675	0.505907
12	0.20338	0.206256	37	0.32114	0.332918
13	0.25322	0.258850	38	0.41348	0.439802
14	0.36491	0.382539	39	0.40132	0.425221
15	0.33111	0.344074	40	0.33293	0.346120
16	0.37923	0.399160	41	0.26257	0.268867
17	0.25368	0.259342	42	0.32306	0.335060
18	0.35038	0.365877	43	0.26292	0.269243
19	0.38169	0.402036	44	0.29039	0.298992
20	0.25021	0.255637	45	0.34660	0.361574
21	0.23802	0.242674	46	0.32363	0.335696
22	0.29342	0.302304	47	0.42299	0.451328
23	0.29987	0.309377	48	0.33790	0.351720
24	0.30933	0.319804	49	0.38716	0.408455
25	0.17717	0.179059	50	0.41747	0.444624

Appendix R, continued

Item-Total Correlations and Corresponding Values of Fisher's Z, for Items of the Eighth-Grade Reading Test continued

Item	r_{pb}	Fisher's Z
51	0.36965	0.388018
52	0.36462	0.382204
53	0.39742	0.420581
54	0.37668	0.396185
55	0.35089	0.366458
56	0.35944	0.376243
57	0.49904	0.548027
58	0.50393	0.554560
59	0.39833	0.421662
60	0.30120	0.310839
61	0.41307	0.439307
62	0.37218	0.390951
63	0.35760	0.374131
64	0.53907	0.602844
65	0.37610	0.395509
66	0.40105	0.424900
67	0.50704	0.558737
68	0.42090	0.448785
69	0.41862	0.446018
70	0.46459	0.503149
71	0.41623	0.443123
72	0.46604	0.505000
73	0.56985	0.647301
74	0.34753	0.362632
75	0.43465	0.465616
76	0.28636	0.294597
77	0.52899	0.588742
78	0.36552	0.383242
79	0.34702	0.362052
80	0.42361	0.452083
81	0.44368	0.476804
82	0.44191	0.474602
83	0.46060	0.498073
84	0.37574	0.395090
85	0.49560	0.543457
86	0.39975	0.423351
87	0.28890	0.297366
88	0.40132	0.425221
89	0.28818	0.296580
90	0.43638	0.467751
91	0.39745	0.420617
92	0.17950	0.181466
93	0.35084	0.366401
94	0.40735	0.432430
95	0.54470	0.610814
96	0.52895	0.588686
97	0.47075	0.511033
98	0.28268	0.290592
99	0.26659	0.273189
100	0.45375	0.489413
101	0.37045	0.388945
102	0.51789	0.573452
103	0.46224	0.500156
104	0.42862	0.458205
105	0.38478	0.405658
106	0.47037	0.510545
107	0.33102	0.343973
108	0.42260	0.450853
109	0.49562	0.543483
110	0.42223	0.450403
111	0.36778	0.385853
112	0.48488	0.529345
113	0.41103	0.436850
114	0.36424	0.381766
115	0.47621	0.518071
116	0.51724	0.572564
117	0.41846	0.445824
118	0.23643	0.240989
119	0.39272	0.415012
120	0.41247	0.438584
121	0.51393	0.568056
122	0.40022	0.423911

Appendix R, continued

Item-Total Correlations and Corresponding Values of Fisher's Z, for Items of the Eighth-Grade Mathematics Test

Item	r_{pb}	Fisher's Z	Item	r_{pb}	Fisher's Z
1	0.21971	0.223351	38	0.28771	0.296068
2	0.29425	0.303213	39	0.47472	0.516146
3	0.33233	0.345445	40	0.34257	0.357001
4	0.53999	0.604142	41	0.19081	0.193178
5	0.31828	0.329732	42	0.38414	0.404907
6	0.40029	0.423994	43	0.22211	0.225874
7	0.37748	0.397118	44	0.31870	0.330199
8	0.47619	0.518045	45	0.06447	0.064560
9	0.21565	0.219089	46	0.27973	0.287389
10	0.36085	0.377863	47	0.29998	0.309498
11	0.44585	0.479509	48	0.40328	0.427560
12	0.44962	0.484224	49	0.32475	0.336948
13	0.45961	0.496817	50	0.26383	0.270221
14	0.28456	0.292637	51	0.35158	0.367245
15	0.40674	0.431699	52	0.38315	0.403746
16	0.32201	0.333888	53	0.25561	0.261406
17	0.51857	0.574382	54	0.26376	0.270145
18	0.36638	0.384235	55	0.10249	0.102851
19	0.29976	0.309256	56	0.23899	0.243703
20	0.44055	0.472913	57	0.31342	0.324334
21	0.39438	0.416976	58	0.36035	0.377288
22	0.25201	0.257558	59	0.42370	0.452193
23	0.29285	0.301681	60	0.16225	0.163697
24	0.30581	0.315917	61	0.18442	0.186554
25	0.54684	0.613862	62	0.33681	0.350490
26	0.33861	0.352522	63	0.51805	0.573671
27	0.42919	0.458903	64	0.35680	0.373214
28	0.37105	0.389640	65	0.36002	0.376909
29	0.41232	0.438403	66	0.51581	0.570614
30	0.33756	0.351336	67	0.35856	0.375232
31	0.16926	0.170905	68	0.51575	0.570532
32	0.51241	0.565992	69	0.37943	0.399394
33	0.44808	0.482295	70	0.38882	0.410409
34	0.49571	0.543602	71	0.38797	0.409408
35	0.31526	0.326375	72	0.21732	0.220842
36	0.41454	0.441081	73	0.35193	0.367645
37	0.41924	0.446770	74	0.24381	0.248821

Appendix R, continued

Item-Total Correlations and Corresponding Values of Fisher's Z, for Items of the Eighth-Grade Mathematics Test continued

Item	r_{pb}	Fisher's Z
75	0.47784	0.520181
76	0.45804	0.494828
77	0.36736	0.385368
78	0.38653	0.407714
79	0.31296	0.323823
80	0.21832	0.221891
81	0.36904	0.387311
82	0.32811	0.340709
83	0.48223	0.525886
84	0.33462	0.348022
85	0.43649	0.467886
86	0.45315	0.488657
87	0.16404	0.165536
88	0.30464	0.314626
89	0.43461	0.465566
90	0.53328	0.594718
91	0.10402	0.104398
92	0.47605	0.517864
93	0.50615	0.557540
94	0.48830	0.533826
95	0.31775	0.329142
96	0.10135	0.101699
97	0.50872	0.561001
98	0.36626	0.384097
99	0.33934	0.353346
100	0.36166	0.378794
101	0.38245	0.402926
102	0.13611	0.136960
103	0.40768	0.432826
104	0.38076	0.400948
105	0.09674	0.097044
106	0.47323	0.514224
107	0.25436	0.260069
108	0.34724	0.362302
109	0.19357	0.196044
110	0.46694	0.506150
111	0.53170	0.592512
112	0.42310	0.451462
113	0.53996	0.604099
114	0.18639	0.188595
115	0.37978	0.399803
116	0.41249	0.438608
117	0.44912	0.483597
118	0.52112	0.577876
119	0.42222	0.450391
120	0.52037	0.576847
121	0.43841	0.470261
122	0.45032	0.485102
123	0.37884	0.398705
124	0.40132	0.425221
125	0.19672	0.199318
126	0.37774	0.397421
127	0.36716	0.385137

Appendix R, continued

Item-Total Correlations and Corresponding Values of Fisher's Z, for Items of the BST Reading Test

Item	r_{pb}	Fisher's Z	Item	r_{pb}	Fisher's Z
1	0.34150	0.355790	38	0.32542	0.337697
2	0.28660	0.294858	39	0.38446	0.405283
3	0.33687	0.350558	40	0.42536	0.454218
4	0.36902	0.387288	41	0.37040	0.388887
5	0.33148	0.344490	42	0.43000	0.459897
6	0.41257	0.438704	43	0.26607	0.272630
7	0.30098	0.310597	44	0.38424	0.405025
8	0.18339	0.185488	45	0.40452	0.429042
9	0.27822	0.285752	46	0.35752	0.374040
10	0.44040	0.472727	47	0.30365	0.313535
11	0.35461	0.370707	48	0.30622	0.316369
12	0.31952	0.331112	49	0.43355	0.464260
13	0.49582	0.543748	50	0.35400	0.370009
14	0.32691	0.339365	51	0.37544	0.394741
15	0.34502	0.359780	52	0.42927	0.459001
16	0.29344	0.302326	53	0.37531	0.394589
17	0.32211	0.334000	54	0.31574	0.326908
18	0.37506	0.394298	55	0.31058	0.321187
19	0.34018	0.354296	56	0.30301	0.312831
20	0.31685	0.328142	57	0.30452	0.314494
21	0.34150	0.355790	58	0.37609	0.395498
22	0.32562	0.337921	59	0.32437	0.336523
23	0.27546	0.282763	60	0.33985	0.353923
24	0.42995	0.459835	61	0.39530	0.418066
25	0.39986	0.423482	62	0.28126	0.289050
26	0.35820	0.374819	63	0.35510	0.371268
27	0.30514	0.315178	64	0.48552	0.530182
28	0.39542	0.418208	65	0.36216	0.379370
29	0.27405	0.281237	66	0.46814	0.507686
30	0.30211	0.311840	67	0.32293	0.334915
31	0.38159	0.401919	68	0.39637	0.419335
32	0.28721	0.295523	69	0.47395	0.515152
33	0.41854	0.445921	70	0.41679	0.443801
34	0.35942	0.376220	71	0.46252	0.500512
35	0.36751	0.385541	72	0.45075	0.485641
36	0.43252	0.462992	73	0.42227	0.450451
37	0.29087	0.299516	74	0.38891	0.410515

Appendix R, continued

Item-Total Correlations and Corresponding Values of Fisher's Z, for Items of the BST Reading Test

Item	r_{pb}	Fisher's Z
75	0.41620	0.443087
76	0.30278	0.312577
77	0.25644	0.262294
78	0.40454	0.429065
79	0.31045	0.321043
80	0.30799	0.318323
81	0.47672	0.518731
82	0.30097	0.310586
83	0.39545	0.418244
84	0.36012	0.377024
85	0.24417	0.249204
86	0.31839	0.329854
87	0.42375	0.452254
88	0.35564	0.371886
89	0.23870	0.243395
90	0.41903	0.446515
91	0.31750	0.328864
92	0.24221	0.247121
93	0.38635	0.407502
94	0.27380	0.280967
95	0.29550	0.304582
96	0.27635	0.283726
97	0.41819	0.445496
98	0.41401	0.440441
99	0.41990	0.447571
100	0.46093	0.498492
101	0.45496	0.490937
102	0.38047	0.400609
103	0.32679	0.339230
104	0.31989	0.331525
105	0.36822	0.386362

Appendix R, continued

Item-Total Correlations and Corresponding Values of Fisher's Z, for Items of the BST Mathematics Test

Item	r_{pb}	Fisher's Z	Item	r_{pb}	Fisher's Z
1	0.18343	0.185530	38	0.33848	0.352375
2	0.31221	0.322992	39	0.28125	0.289039
3	0.36405	0.381547	40	0.24550	0.250618
4	0.39823	0.421544	41	0.45581	0.492010
5	0.44749	0.481557	42	0.51149	0.564746
6	0.50758	0.559465	43	0.41141	0.437307
7	0.33606	0.349644	44	0.25636	0.262208
8	0.42134	0.449320	45	0.34234	0.356741
9	0.47349	0.514559	46	0.54493	0.611141
10	0.36926	0.387566	47	0.28995	0.298512
11	0.32015	0.331814	48	0.45226	0.487538
12	0.20388	0.206777	49	0.22675	0.230761
13	0.40507	0.429699	50	0.48988	0.535902
14	0.38572	0.406762	51	0.46091	0.498466
15	0.36407	0.381570	52	0.42756	0.456907
16	0.32233	0.334245	53	0.35244	0.368227
17	0.35298	0.368844	54	0.50767	0.559586
18	0.16890	0.170534	55	0.48390	0.528064
19	0.46574	0.504616	56	0.25065	0.256106
20	0.49547	0.543284	57	0.29123	0.299910
21	0.22615	0.230128	58	0.44014	0.472404
22	0.15346	0.154682	59	0.44221	0.474975
23	0.44079	0.473211	60	0.25793	0.263890
24	0.46582	0.504719	61	0.47407	0.515307
25	0.34489	0.359632	62	0.13459	0.135412
26	0.47770	0.520000	63	0.50912	0.561541
27	0.27508	0.282351	64	0.32446	0.336624
28	0.20961	0.212763	65	0.39254	0.414799
29	0.48352	0.527568	66	0.36007	0.376966
30	0.27841	0.285958	67	0.45495	0.490925
31	0.47708	0.519197	68	0.45953	0.496715
32	0.26888	0.275656	69	0.37824	0.398004
33	0.40260	0.426748	70	0.48376	0.527881
34	0.35530	0.371497	71	0.29749	0.306764
35	0.32397	0.336076	72	0.51034	0.563189
36	0.36637	0.384224	73	0.41253	0.438656
37	0.43806	0.469828	74	0.41271	0.438873

Appendix R, continued

Item-Total Correlations and Corresponding Values of Fisher's Z, for Items of the BST Mathematics Test continued

Item	r_{pb}	Fisher's Z
75	0.42777	0.457164
76	0.36006	0.376955
77	0.46246	0.500436
78	0.38034	0.400457
79	0.37605	0.395451
80	0.46532	0.504080
81	0.46307	0.501212
82	0.44463	0.477987
83	0.40116	0.425031
84	0.39747	0.420641
85	0.46649	0.505575
86	0.46434	0.502830
87	0.47205	0.512705
88	0.45243	0.487752
89	0.37605	0.395451
90	0.30986	0.320391
91	0.39989	0.423518
92	0.54636	0.613178
93	0.15459	0.155839
94	0.37552	0.394834
95	0.47945	0.522270
96	0.39395	0.416467
97	0.42010	0.447813
98	0.35769	0.374234
99	0.28997	0.298533
100	0.43917	0.471202
101	0.34018	0.354296
102	0.35831	0.374946
103	0.48423	0.528495
104	0.36720	0.385183
105	0.43146	0.461689
106	0.33985	0.353923
107	0.40215	0.426211
108	0.44762	0.481720
109	0.33911	0.353086
110	0.22359	0.227432
111	0.28596	0.294161
112	0.30984	0.320368

Appendix R, Continued

Item-Total Correlations and Corresponding Values of Fisher's Z, for Items of the GRP-Eighth-Grade Reading Test

Item	r_{pb}	Fisher's Z
1	0.11722	0.117761
2	0.11998	0.120561
4	0.29923	0.308674
5	0.36082	0.377828
6	0.23575	0.240269
7	0.28078	0.288529
9	0.29755	0.306829
11	0.31019	0.320756
12	0.19877	0.201452
13	0.26474	0.271199
14	0.37504	0.394275
16	0.38733	0.408655
17	0.26576	0.272296
18	0.34923	0.364567
19	0.40159	0.425543
20	0.24952	0.254901
21	0.26666	0.273265
22	0.29786	0.307170
23	0.31358	0.324511
24	0.32698	0.339443
25	0.19499	0.197519
26	0.40909	0.434518
27	0.27906	0.286662
28	0.35759	0.374120
29	0.34078	0.354975
30	0.32007	0.331725
31	0.25538	0.261160
32	0.48511	0.529645
33	0.33037	0.343244
34	0.48577	0.530509
35	0.45678	0.493235
36	0.48164	0.525117
37	0.33297	0.346165
39	0.41565	0.442422
40	0.35427	0.370318
41	0.26416	0.270575
42	0.32454	0.336713
43	0.26919	0.275990
44	0.29732	0.306577
46	0.32722	0.339712
47	0.42638	0.455464
48	0.34190	0.356242
49	0.40875	0.434110
50	0.40199	0.426020
51	0.40362	0.427966
52	0.39352	0.415958
55	0.35157	0.367234
56	0.36467	0.382262
57	0.50190	0.551843
58	0.52602	0.584627
59	0.41183	0.437813
62	0.40131	0.425209
63	0.37655	0.396033
64	0.54847	0.616190
65	0.39333	0.415733
66	0.40645	0.431351
67	0.51219	0.565694
68	0.44688	0.480795
69	0.44404	0.477252
74	0.34613	0.361040
75	0.44042	0.472752
76	0.29645	0.305623
77	0.55525	0.625940
79	0.35034	0.365831
81	0.46135	0.499025
84	0.36160	0.378725
85	0.49343	0.540584
86	0.40352	0.427846
89	0.28080	0.288550
90	0.45707	0.493601
91	0.40737	0.432454
92	0.19889	0.201577
93	0.37057	0.389084

Appendix R, Continued

Item-Total Correlations and Corresponding Values of Fisher's Z, for Items of the GRP-Eighth-Grade Reading Test continued

Item	r_{pb}	Fisher's Z
94	0.41077	0.436537
98	0.29446	0.303443
99	0.28138	0.289180
100	0.47592	0.517696
101	0.37234	0.391137
102	0.54583	0.612422
103	0.48414	0.528378
104	0.43925	0.471301
105	0.40516	0.429807
106	0.48814	0.533616
107	0.35644	0.372802
108	0.43364	0.464371
111	0.36693	0.384871
115	0.49039	0.536574
116	0.52158	0.578508
117	0.43784	0.469555
118	0.24571	0.250842
119	0.38074	0.400925
121	0.51179	0.565152
122	0.39784	0.421080

Appendix R, continued

Item-Total Correlations and Corresponding Values of Fisher's Z, for Items of the GRP-Eighth-Grade Mathematics Test

Item	r_{pb}	Fisher's Z
7	0.38976	0.411517
9	0.24140	0.246260
14	0.30428	0.314230
16	0.36345	0.380855
19	0.32936	0.342110
22	0.28539	0.293540
23	0.30420	0.314141
24	0.32116	0.332940
29	0.40463	0.429173
30	0.38679	0.408020
31	0.20128	0.204066
35	0.32288	0.334859
38	0.31135	0.322040
40	0.36761	0.385657
41	0.20684	0.209868
42	0.41850	0.445872
43	0.26273	0.269039
44	0.34374	0.358328
45	0.10560	0.105995
46	0.29029	0.298883
47	0.29783	0.307137
62	0.37451	0.393659
70	0.40693	0.431926
72	0.25265	0.258241
74	0.31458	0.325620
75	0.50892	0.561271
77	0.40842	0.433713
78	0.38355	0.404215
79	0.36762	0.385668
80	0.25924	0.265293
82	0.32292	0.334904
87	0.18669	0.188905
88	0.32682	0.339264
89	0.46854	0.508198
91	0.16317	0.164642
96	0.14653	0.147592
98	0.36167	0.378806
99	0.36683	0.384755
100	0.38901	0.410633
102	0.19944	0.202149
104	0.39981	0.423423
105	0.18816	0.190429
107	0.27325	0.280373
108	0.36854	0.386733
109	0.24244	0.247365
114	0.25264	0.258231
116	0.44685	0.480757
121	0.45563	0.491782
123	0.42272	0.450999
126	0.39424	0.416810
127	0.38028	0.400387

Appendix R, continued

Item-Total Correlations and Corresponding Values of Fisher's Z, for Items of the GRP-BST Reading Test

Item	r_{pb}	Fisher's Z
1	0.21225	0.215527
2	0.28764	0.295992
5	0.33002	0.342850
7	0.29901	0.308433
8	0.20329	0.206161
9	0.29904	0.308465
12	0.33493	0.348370
14	0.32236	0.334279
16	0.31742	0.328776
17	0.31825	0.329700
20	0.32873	0.341404
21	0.33861	0.352521
22	0.35379	0.369770
23	0.27923	0.286847
25	0.40332	0.427608
26	0.34302	0.357511
27	0.30255	0.312324
28	0.39064	0.412555
29	0.28848	0.296906
30	0.31114	0.321808
31	0.38397	0.404709
32	0.28574	0.293921
34	0.37584	0.395207
35	0.39120	0.413216
37	0.29229	0.301069
38	0.32855	0.341202
39	0.39335	0.415756
40	0.43242	0.462869
43	0.27894	0.286531
44	0.38536	0.406340
45	0.40842	0.433714
46	0.36138	0.378472
47	0.31148	0.322183
50	0.35904	0.375784
52	0.45450	0.490357
53	0.39142	0.413476
54	0.33388	0.347188
55	0.33949	0.353515
56	0.32568	0.337988
57	0.33159	0.344615
58	0.39016	0.411989
59	0.34995	0.365388
61	0.41654	0.443499
62	0.29878	0.308179
63	0.35834	0.374981
65	0.38701	0.408279
66	0.48861	0.534233
67	0.33610	0.349689
68	0.40680	0.431770
69	0.48132	0.524701
71	0.45679	0.493248
75	0.42148	0.449491
76	0.31821	0.329653
77	0.27293	0.280028
78	0.40259	0.426735
79	0.33336	0.346604
80	0.33430	0.347662
82	0.32320	0.335216
83	0.41047	0.436176
85	0.24302	0.247982
86	0.32075	0.332482
87	0.43141	0.461627
89	0.24823	0.253526
90	0.42985	0.459713
91	0.31862	0.330111
92	0.25040	0.255840
94	0.28547	0.293627
95	0.33244	0.345570
96	0.28725	0.295566
99	0.43846	0.470322
102	0.38590	0.406974
103	0.33588	0.349440
104	0.32671	0.339141

Appendix R, continued

Item-Total Correlations and Corresponding Values of Fisher's Z, for Items of the GRP-BST Mathematics Test

Item	r_{pb}	Fisher's Z
1	0.21122	0.214448
2	0.30718	0.317429
3	0.40227	0.426354
4	0.44302	0.475982
5	0.49796	0.546590
7	0.33830	0.352172
9	0.51900	0.574970
12	0.22015	0.223814
13	0.42551	0.454401
15	0.39659	0.419596
17	0.34475	0.359473
18	0.17754	0.179442
20	0.55442	0.624740
21	0.28501	0.293127
22	0.20170	0.204504
23	0.43928	0.471338
27	0.34326	0.357783
28	0.23294	0.237296
30	0.28948	0.297999
32	0.26580	0.272339
34	0.38619	0.407314
35	0.32984	0.342649
38	0.39417	0.416728
39	0.28399	0.292017
40	0.26667	0.273275
44	0.29600	0.305130
45	0.33006	0.342896
47	0.29951	0.308981
48	0.47097	0.511316
49	0.27731	0.284766
52	0.50974	0.562378
53	0.36067	0.377656
56	0.24230	0.247216
57	0.31040	0.320988
58	0.48572	0.530443
60	0.29112	0.299790
62	0.16652	0.168085
64	0.33296	0.346154
66	0.40305	0.427285
67	0.47499	0.516495
69	0.40631	0.431184
71	0.29629	0.305448
74	0.44031	0.472615
75	0.45518	0.491215
82	0.52779	0.587077
83	0.39162	0.413712
88	0.45322	0.488745
93	0.19924	0.201941
98	0.39024	0.412083
99	0.29653	0.305711
110	0.26903	0.275818
111	0.30141	0.311070
112	0.31981	0.331435

Appendix R, continued

Item-Total Correlations and Corresponding Values of Fisher's Z, for Items of the GRS-Eighth-Grade Reading Test

Item	r_{pb}	Fisher's Z	Item	r_{pb}	Fisher's Z
1	0.12037	0.120956	37	0.32686	0.339309
2	0.12028	0.120865	38	0.41391	0.440320
3	0.42256	0.450804	39	0.41104	0.436862
4	0.29671	0.305908	40	0.34463	0.359337
5	0.35729	0.373776	41	0.26096	0.267138
6	0.23432	0.238756	42	0.32380	0.335886
7	0.27592	0.283260	43	0.26687	0.273491
8	0.27822	0.285752	44	0.29673	0.305930
9	0.30042	0.309981	46	0.32728	0.339779
10	0.33274	0.345906	47	0.42353	0.451986
11	0.30756	0.317848	48	0.33907	0.353041
12	0.19772	0.200359	49	0.39129	0.413322
13	0.26119	0.267385	50	0.41179	0.437765
14	0.37440	0.393531	51	0.38864	0.410197
15	0.33176	0.344805	52	0.37703	0.396593
16	0.38493	0.405834	54	0.37705	0.396616
17	0.25937	0.265433	55	0.35360	0.369552
18	0.35245	0.368239	56	0.36177	0.378921
19	0.39018	0.412012	57	0.49710	0.545447
20	0.24517	0.250267	58	0.51161	0.564908
21	0.25358	0.259235	59	0.40206	0.426104
22	0.29163	0.300347	60	0.29900	0.308421
23	0.30762	0.317915	62	0.38379	0.404497
24	0.31765	0.329031	63	0.36404	0.381535
25	0.18536	0.187528	64	0.54087	0.605385
26	0.39534	0.418114	65	0.38612	0.407232
27	0.27471	0.281951	66	0.40465	0.429197
28	0.34406	0.358690	67	0.50885	0.561177
29	0.33615	0.349746	68	0.43463	0.465591
30	0.31537	0.326497	69	0.42809	0.457556
31	0.25663	0.262497	74	0.35090	0.366470
32	0.48543	0.530064	75	0.43790	0.469630
33	0.33083	0.343760	76	0.29009	0.298664
34	0.48667	0.531687	77	0.54036	0.604664
35	0.45502	0.491013	78	0.35932	0.376105
36	0.47530	0.516895	79	0.35013	0.365592

Appendix R, continued

Item-Total Correlations and Corresponding Values of Fisher's Z, for Items of the GRS-Eighth-Grade Reading Test continued

Item	r_{pb}	Fisher's Z
81	0.45297	0.488431
84	0.37162	0.390301
85	0.49501	0.542675
86	0.40435	0.428838
87	0.27882	0.286402
88	0.40230	0.426390
89	0.28939	0.297900
90	0.44633	0.480108
91	0.40427	0.428743
92	0.18892	0.191217
93	0.36177	0.378921
94	0.40870	0.434050
95	0.53776	0.600999
98	0.29003	0.298599
99	0.27687	0.284289
100	0.46644	0.505511
101	0.37170	0.390394
102	0.52846	0.588006
103	0.47396	0.515165
104	0.43396	0.464765
105	0.39264	0.414917
106	0.47795	0.520324
107	0.34279	0.357251
108	0.42286	0.451170
111	0.36385	0.381316
112	0.47370	0.514830
113	0.39991	0.423542
115	0.48262	0.526394
116	0.51969	0.575915
117	0.42500	0.453779
118	0.24482	0.249895
119	0.39338	0.415793
120	0.40958	0.435106
121	0.51497	0.569470
122	0.39906	0.422530

Appendix R, continued

Item-Total Correlations and Corresponding Values of Fisher's Z, for Items of the GRS-Eighth-Grade Mathematics Test

Item	r_{pb}	Fisher's Z	Item	r_{pb}	Fisher's Z
1	0.21795	0.221503	54	0.30105	0.310674
2	0.31270	0.323535	55	0.12598	0.126653
7	0.39074	0.412673	56	0.26257	0.268867
8	0.46490	0.503544	57	0.31700	0.328308
9	0.23771	0.242346	58	0.40192	0.425937
14	0.29077	0.299407	60	0.19964	0.202358
15	0.39740	0.420557	61	0.19801	0.200660
16	0.34559	0.360427	62	0.36392	0.381397
19	0.30821	0.318566	64	0.36004	0.376932
21	0.38675	0.407973	67	0.35487	0.371004
22	0.27546	0.282763	69	0.37990	0.399943
23	0.30423	0.314174	70	0.41023	0.435888
24	0.31841	0.329877	71	0.37430	0.393414
27	0.44049	0.472839	72	0.23713	0.241731
28	0.36472	0.382319	74	0.29219	0.300959
29	0.41104	0.436862	75	0.50398	0.554627
30	0.36905	0.387323	77	0.38703	0.408302
31	0.17995	0.181931	78	0.38796	0.409396
32	0.50813	0.560206	79	0.35195	0.367668
33	0.45287	0.488305	80	0.24388	0.248895
35	0.31973	0.331346	82	0.31620	0.327419
37	0.42123	0.449186	83	0.49339	0.540531
38	0.30302	0.312842	84	0.31481	0.325876
40	0.36614	0.383958	87	0.18034	0.182334
41	0.20128	0.204066	88	0.31356	0.324489
42	0.40334	0.427631	89	0.46515	0.503863
43	0.23891	0.243618	91	0.13587	0.136715
44	0.33320	0.346424	92	0.47379	0.514946
45	0.08627	0.086485	95	0.31555	0.326697
46	0.28995	0.298512	96	0.13333	0.134129
47	0.29688	0.306095	98	0.36177	0.378921
49	0.33571	0.349250	99	0.35165	0.367325
50	0.26158	0.267804	100	0.38085	0.401053
51	0.35679	0.373203	102	0.17477	0.176583
52	0.38458	0.405424	103	0.41146	0.437367
53	0.27406	0.281248	104	0.39248	0.414728

Appendix R, continued

Item-Total Correlations and Corresponding Values of Fisher's Z, for Items of the GRS-Eighth-Grade Mathematics Test continued

Item	r_{pb}	Fisher's Z
105	0.15426	0.155501
107	0.26527	0.271769
108	0.36004	0.376932
109	0.22129	0.225012
112	0.42389	0.452425
114	0.22749	0.231541
116	0.43363	0.464359
117	0.46156	0.499292
119	0.43409	0.464925
121	0.45285	0.488280
123	0.41356	0.439898
124	0.39837	0.421710
126	0.38983	0.411600
127	0.37443	0.393566

Appendix R, continued

Item-Total Correlations and Corresponding Values of Fisher's Z, for Items of the GRS-BST Reading Test

Item	r_{pb}	Fisher's Z
1	0.20194	0.204754
2	0.28603	0.294237
5	0.32968	0.342469
6	0.40619	0.431040
7	0.30308	0.312908
8	0.19467	0.197187
9	0.28793	0.296308
11	0.34667	0.361654
12	0.32832	0.340944
14	0.32245	0.334379
15	0.33447	0.347853
16	0.30149	0.311158
17	0.31750	0.328864
18	0.36428	0.381812
20	0.32079	0.332527
21	0.33906	0.353030
22	0.34385	0.358452
23	0.28199	0.289843
24	0.42659	0.455721
25	0.39906	0.422530
26	0.34680	0.361802
27	0.30195	0.311664
28	0.39176	0.413877
29	0.28118	0.288963
30	0.30253	0.312302
31	0.38213	0.402551
32	0.28532	0.293464
34	0.36583	0.383600
35	0.38006	0.400130
37	0.28506	0.293181
38	0.32774	0.340294
39	0.38817	0.409644
40	0.42906	0.458744
43	0.27155	0.278536
44	0.38414	0.404907
45	0.40621	0.431064
46	0.35586	0.372138
47	0.30738	0.317649
48	0.30086	0.310465
50	0.35385	0.369838
51	0.37499	0.394217
52	0.44256	0.475410
53	0.38951	0.411222
54	0.32152	0.333341
55	0.32932	0.342065
56	0.31198	0.322737
57	0.32104	0.332806
58	0.38593	0.407009
59	0.33338	0.346626
61	0.40585	0.430633
62	0.28843	0.296853
63	0.35433	0.370387
65	0.37589	0.395265
66	0.48123	0.524584
67	0.32937	0.342121
68	0.40475	0.429317
69	0.48047	0.523595
70	0.42448	0.453144
71	0.46062	0.498098
72	0.44845	0.482758
74	0.39032	0.412178
75	0.42157	0.449600
76	0.30956	0.320059
77	0.26062	0.266773
78	0.40088	0.424697
79	0.31597	0.327164
80	0.32005	0.331703
82	0.31181	0.322549
83	0.40677	0.431735
84	0.35685	0.373272
85	0.24160	0.246473
86	0.31987	0.331502
87	0.42939	0.459149

Appendix R, continued

Item-Total Correlations and Corresponding Values of Fisher's Z, for Items of the GRS-BST Reading Test continued

Item	r_{pb}	Fisher's Z
88	0.35870	0.375393
89	0.24697	0.252183
90	0.42622	0.455268
91	0.32235	0.334267
92	0.24660	0.251789
93	0.38898	0.410598
94	0.28437	0.292430
95	0.31769	0.329076
96	0.28229	0.290169
97	0.42179	0.449867
98	0.41782	0.445048
99	0.43486	0.465875
100	0.46191	0.499737
102	0.38624	0.407373
103	0.33104	0.343996
104	0.32519	0.337440
105	0.38000	0.400060

Appendix R, continued

Item-Total Correlations and Corresponding Values of Fisher's Z, for Items of the GRS-BST Mathematics Test

Item	r_{pb}	Fisher's Z	Item	r_{pb}	Fisher's Z
2	0.30502	0.315045	56	0.24149	0.246356
3	0.37843	0.398226	57	0.29871	0.308103
4	0.41941	0.446976	58	0.47104	0.511406
5	0.48501	0.529515	60	0.27977	0.287432
7	0.32989	0.342705	61	0.49436	0.541814
8	0.40357	0.427906	62	0.15324	0.154457
9	0.50144	0.551228	64	0.32565	0.337954
11	0.33828	0.352149	66	0.38635	0.407502
12	0.20277	0.205620	67	0.46162	0.499368
13	0.41610	0.442966	69	0.39382	0.416313
15	0.39094	0.412909	71	0.29010	0.298675
17	0.34843	0.363656	73	0.42534	0.454194
18	0.16672	0.168291	74	0.41993	0.447607
19	0.50857	0.560799	75	0.44879	0.483184
20	0.54623	0.612992	77	0.46332	0.501530
21	0.26908	0.275872	82	0.51184	0.565220
22	0.17997	0.181952	83	0.39242	0.414657
23	0.46690	0.506099	88	0.47715	0.519288
27	0.32447	0.336636	89	0.41036	0.436044
28	0.22674	0.230750	90	0.31597	0.327164
30	0.27611	0.283466	93	0.18134	0.183368
32	0.25720	0.263108	96	0.38615	0.407267
33	0.38155	0.401873	98	0.39119	0.413204
34	0.37213	0.390893	99	0.28418	0.292223
35	0.32090	0.332650	100	0.46521	0.503940
36	0.37264	0.391485	101	0.33791	0.351731
38	0.37938	0.399335	102	0.34115	0.355393
39	0.27755	0.285026	105	0.44031	0.472615
40	0.25438	0.260090	106	0.32476	0.336959
41	0.46672	0.505869	108	0.47451	0.515875
44	0.27428	0.281486	110	0.25807	0.264040
45	0.32290	0.334881	111	0.29524	0.304297
47	0.29386	0.302786	112	0.30845	0.318832
48	0.46514	0.503851			
49	0.26694	0.273566			
52	0.48280	0.526629			
53	0.34928	0.364623			

APPENDIX S

Format and Content of Type I and Type II Items of the Original Tests

This appendix shows the assignment of items of the original tests to the cells of an item content-by-format matrix. Item content is shown as rows, and item format as columns. A six by five matrix is used for the reading tests. A seven by four matrix is used for the mathematics tests. Within each cell, three numbers appear in the arrangement *a/b/c*. The number *a* is the number of items of the original test assigned to that cell of the matrix. The number *b* is the number of those items which were excluded only from the GRP- synthetic test. The number *c* is the number of items excluded from both the GRP- and GRS- synthetic tests.

Content and Format of Items of the Eighth-Grade Reading Test

	Format:					
Content:	Long Passage	Short Passage	Poem	Figure	Short Answer	Row Total
Literal Comprehension	31/3/4	7/1/0	0/0/0	21/3/2	9/1/0	68/8/6
Main Ideas	10/2/1	7/1/1	0/0/0	0/0/0	0/0/0	17/3/2
Inference	10/2/1	3/0/0	0/0/0	0/0/0	0/0/0	13/2/1
Figures of Speech	4/1/1	5/0/1	0/0/0	0/0/0	3/1/0	12/2/2
Language Structure & Rules	0/0/0	0/0/0	3/0/1	0/0/0	0/0/0	3/0/1
Fact vs. Opinion	6/0/0	3/0/1	0/0/0	0/0/0	0/0/0	9/0/1
Column Total	6/8/7	25/2/3	3/0/1	21/3/2	12/2/0	n= 122

Appendix S, continued

Content and Format of Items of the Eighth-Grade Mathematics Test

	Format:				
	Story Problem	Figure	Short Answer	Math Problem	Row Total
Content:					
Units of Measure	3/1/1	3/1/1	11/2/4	0/0/0	17/4/6
Money	7/2/2	2/0/0	4/0/1	0/0/0	13/2/5
Geometry	1/0/1	16/2/7	1/0/0	0/0/0	18/2/8
Fractions	6/1/2	1/0/0	5/0/4	14/2/4	26/3/10
Basic Facts & Principles	5/1/1	7/1/1	11/3/1	0/0/0	23/5/3
Arithmetic Operations	10/4/3	3/0/3	2/0/0	6/2/0	21/6/6
Probability & Statistics	1/0/1	8/0/2	0/0/0	0/0/0	9/0/3
Column Total	33/9/11	40/4/14	34/5/0	20/4/4	n=127

Appendix S, continued

Content and Format of Items of the BST Reading Test

	Format:					
	Long Passage	Short Passage	Poem	Figure	Short Answer	Row Total
Content:						
Literal Comprehension	36/7/5	10/1/0	0/0/0	7/0/2	2/0/1	55/8/8
Main Ideas	6/0/0	0/0/0	0/0/0	0/0/0	0/0/0	6/0/0
Inference	7/2/0	4/0/2	0/0/0	0/0/0	0/0/0	11/2/2
Figures of Speech	6/1/1	7/1/0	0/0/0	0/0/0	1/0/0	14/2/1
Language Structure & Rules	2/2/0	2/0/0	0/0/0	0/0/0	9/3/1	13/5/1
Fact vs. Opinion	2/0/0	0/0/0	0/0/0	0/0/0	4/0/2	4/0/2
Column Total	59/12/6	23/2/2	0/0/0	7/0/2	16/3/3	n=105

Appendix S, continued

Content and Format of Items of the BST Mathematics Test

	Format:				
	Story Problem	Figure	Short Answer	Math Problem	Row Total
Content:					
Units of Measure	4/2/0	0/0/0	10/1/6	0/0/0	14/3/7
Money	6/3/0	1/0/1	0/0/0	0/0/0	7/3/1
Geometry	3/0/0	10/2/4	1/0/0	0/0/0	14/2/5
Fractions	0/0/0	6/1/2	11/4/1	8/2/3	25/7/6
Basic Facts & Principles	2/0/1	9/1/2	9/3/1	0/0/0	20/4/5
Arithmetic Operations	10/2/6	5/0/3	7/1/4	4/0/0	26/3/15
Probability & Statistics	2/0/0	4/0/0	0/0/0	0/0/0	6/0/0
Column Total	27/7/7	35/4/12	38/9/12	12/2/3	n=112