SEX AND ETHNIC DIFFERENCES IN APTITUDE INDICATOR MEASUREMENT MODELS

by

Dianne W. Robertshaw

Dissertation submitted to the Graduate Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY
in
Educational Research

APPROVED:

Lee M. Wolfle, Chairman

Lawrence H. Cross

Dennis E. Hinkle

Klaus H. Hinkelmann

Margaret A. Eisenhart

July, 1982
Blacksburg, Virginia

ACKNOWLEDGMENTS

My deepest appreciation is extended to my committee chairman, Dr. Lee M. Wolfle, for his expert guidance and his constant confidence in my ability to meet his high standards. He never ran out of time, patience or encouragement. Dr. Wolfle possesses all the qualities one hopes for in a committee chairman. I would like to thank the remaining members of my committee, Dr. Lawrence Cross, Dr. Dennis Hinkle, Dr. Margaret Eisenhart and Dr. Klaus Hinkelmann, for their critical comments and suggestions, and for their willingness to read the final draft during their vacation time in June.

The Cunningham Foundation provided financial support during the past year, enabling me to devote full time to this project. I thank , Dean of Research and Graduate Studies, and the Cunningham Fellowship selection committee for this support, and for the honor of being chosen one of the first Cunningham Fellows.

Many people provided moral support during the past year. The understanding I received from my children, and , far exceeded what I expected. They realized the importance of the project to me and never complained about the time that was taken away from them. My friend let me share all the good moments and, more importantly, all the frustrating ones. I thank him sincerely for all the hours and for helping me keep things

in perspective. I thank my many good friends, especially , and , for their empathy and encouragement.

Finally, I sincerely thank my parents, and , for their support and encouragement over the years whenever I have undertaken to further my education. They have always considered my goals to be as important as I have considered them to be.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

INTRODUCTION TO SEX AND ETHNIC DIFFERENCES IN APTITUDE INDICATOR MEASUREMENT MODELS

STATEMENT OF THE PROBLEM

Problems of measurement are found in every branch of science. Measuring instruments used in physical sciences may yield erroneous readings because of imperfections, or the condition of their mechanisms. Even when machine readings are accurate, they may be observed or recorded with error. In the social sciences measurement is even more problematic because it is rarely a mechanical process. More commonly, measured variables merely index abstract theoretical constructs, which themselves cannot be measured directly. In this event, the measurement problem is to confirm the link between these abstract theoretical constructs and their empirical indicators (Zeller and Carmines, 1980). The usefulness of an indicator as representative of a construct is determined by its measurement error, defined by Blalock (1968) as the degree to which an indicator does not measure a theoretical construct.

Measurement theory identifies two sources of measurement error in construct indicators. The first is unsystematic, random measurement error, which determines the reliability of the indicator. The second is nonrandom measurement error, which determines the

validity of the indicator. Ideally a construct indicator should have only one underlying construct. When there is systematic variance in an indicator other than that related to the hypothesized construct, there is variation in the indicator that is related to other constructs. If indicators are not valid measures of one construct, several interpretations of statistical results are possible. If the indicators are not reliable, parameter estimates will be biased and conclusions may be incorrect.

Measurement properties of construct indicators have been largely ignored by social scientists. Until recently it was rare to see reliability and validity coefficients reported in research papers (Bohrnstedt and Borgatta, 1980). Furthermore, Bohrnstedt and Borgatta (1980) observe, even when reliability coefficients for valid indicators were reported, parameter estimates were almost never corrected for attenuation due to random measurement error. This attitude prevailed even though it had been shown that, in regression analysis, the violation of the assumption of perfect measurement is far more likely to affect the quality of regression coefficients than the violation of any other assumption (Bohrnstedt and Carter, 1971).

Least-squares regression coefficients have long been known to be biased because of random measurement error (Walker and Lev, 1953). Wolfle (1979), for example, has shown in the bivariate case that the regression slope corrected for random error is:

$$b'_{yx} = b_{yx} \left(\sigma_x^2 \; / \; [\, \sigma_x^2 \; - \; \sigma_e^2 \,] \right),$$

in which b'_{yx} is the corrected slope, σ^2_x is the variance of x and σ^2_e is the variance of the random measurement error in x. Therefore, in the presence of random measurement error, ordinary least-squares regression coefficients are attenuated. In complex models, with multiple independent variables, error may be large for some variables and small for others. As a result, the effect of errors on coefficients may be additive or offsetting (Mason et al., 1976). It is difficult to determine, therefore, the extent and direction of bias in regression coefficients a priori, even if reliabilities of the indicators are known (Mason et al., 1976; Kenny, 1979). In complex models, although some coefficients are underestimated and some overestimated, it has been shown that explained variance (R^2) in the dependent variable, is attenuated by the presence of random measurement error (Bielby et al., 1977; Cuttance, 1982).

Difficulties may be compounded in studies in which the major concern is the determination of differences between regression coefficients for two groups. If coefficients are equally biased for the groups (random measurement error in the independent variables is the same for both groups), the magnitude of the differences in coefficients will be unaffected. If coefficients are unequally biased, however, the magnitude of the differences will be affected by the differential random measurement error (Wolfle and Lichtman, 1981).

Blalock (1970) suggested the need for studies that correct parameter estimates for random measurement error, and also called for methodological studies on the effects of random measurement error on parameter estimates and hypothesis testing. He contended that methodological studies on the implications of random error can help researchers to see more clearly the steps that must be taken to become increasingly precise in the development of adequate social theory.

In the sociological literature on status attainment, one can find a number of studies in which regression coefficients in structural models have been corrected for random measurement error, and for systematic error resulting from correlated errors (Siegel and Hodge, 1968; Bowles, 1972; Jencks et al., 1972; Bowles and Nelson, 1974; Bowles and Gintis, 1976; Mason et al., 1976; Treiman and Hauser, 1976; Bielby et al., 1977; Corcoran, 1980). The variables in these studies were respondents' reports of information such as parental status characteristics, educational attainment, and economic success. When conclusions were made about the effect of measurement error on regression coefficients, opinions differed. Jencks et al. (1972) and Siegel and Hodge (1968) concluded that bias was negligible; Bowles (1972), Bielby et al. (1977) and Corcoran (1980) found it to be substantial enough to affect substantive interpretations of results.

Bielby et al. (1977) considered the conclusions of Jencks et al. (1972), Siegel and Hodge (1968) and Bowles (1972) to be based on

inappropriate data, inadequate specifications and crude estimation procedures. The later studies by Bielby et al. (1977) and Corcoran (1980) do not necessarily suffer from these problems; both studies concluded that measurement error seriously biases regression coefficients and differences between coefficients when groups are compared. Bielby et al. (1977) compared blacks and whites; Corcoran (1980) compared males and females.

In educational research, the effect of measurement error on parameter estimates and substantive conclusions has received less attention. Patteson and Wolfle (1981) compared regression coefficients uncorrected for random measurement error to those corrected for error for a model of educational attainment and found bias to be substantial for many coefficients. Cuttance (1982) investigated measurement error in questions asked in an educational survey of school dropouts in Scotland. The questions were categorized as factual and non-factual (opinions). Reliability differences were found between types of questions and the effects of these differences on correlation and regression coefficients were examined. Cuttance (1982) concluded that bias and differential bias in question types cannot be ignored.

Most of the studies in both sociology and education that have assessed the effects of measurement error have been interested in error in survey questions. No in-depth studies have been undertaken to investigate the effects and implications of measurement error in test data.

Perhaps the most important constructs in educational research are those which underlie ability, achievement or aptitude tests. (Henceforth, these tests and constructs will be referred to as aptitude. The distinction between ability, achievement and aptitude tests has always been unclear to most researchers, and the opinion of many psychometricians is that there are no real differences [see Humphreys, 1962].) Either a general aptitude construct, or separate constructs such as verbal fluency and numerical facility, are frequently represented by independent variables and covariates in educational research. This country's commitment to equal educational opportunity has inspired a great deal of research and evaluation designed to determine the extent to which specific educational performances and achievements are dependent upon aptitude, and which specific educational programs have effects net of aptitude.

One kind of research, which has become especially popular, is that which makes comparisons between the sexes or among whites, blacks and Mexican-Americans. Questions are posed about the equality of effects of aptitude on performance, and the equality of effects of other factors, controlling for aptitude. The aim of such research is to establish the degree to which our educational system truly provides equal opportunity for advancement to those with equal aptitudes, regardless of gender or ethnic group.

Studies of the effects of aptitude, in which regression coefficients have not been corrected for random measurement error,

may or may not contain faulty conclusions. It is certain that they contain biased results. Only their extent is uncertain for different sociocultural groups. It is also likely that studies which have compared the sexes, or whites with blacks or Mexican-Americans, have not only reported biased coefficients within groups, but have reported biased differences in regression coefficients between groups.

The empirical evidence indicates that aptitude construct indicators are not always valid or equally reliable across sex or ethnic groups. Even without empirical evidence, it should be expected that the sexes and different ethnic groups would be measured with differential reliability by aptitude indicators. Blacks and Mexican-Americans, for example, consistently score lower on aptitude tests than do whites. Males consistently score higher on tests of mathematical aptitude, and females score higher on tests of verbal aptitude. Level of aptitude is known to affect reliability coefficients because lower aptitude test takers are more likely to guess (Nunnally, 1978). Those groups scoring lower usually have restricted ranges of scores, as well, which may also lower reliability coefficients.

In summary, educational research has neglected the issue of measurement error in test data and, specifically, in aptitude tests which are frequently used in quantitative research studies in the field. The validity and reliability of aptitude indicators are often not investigated and, when they are, corrections for the presence of

random error in valid indicators are seldom made. Furthermore, little attention has been given to the possibility that aptitude indicators are not equally reliable for different groups, thereby differentially biasing within group parameter estimates and biasing differences in parameter estimates between groups. Evidence exists that suggests males and females, and whites, blacks and Mexican-Americans are measured with differential reliability by indicators of cognitive aptitudes, yet there are no studies that confirm this, or that indicate the extent to which failure to correct for random error yields biased within group parameter estimates and differences in parameter estimates between groups. It was the purpose of this research to investigate the validity and reliability of typical aptitude construct indicators for white, black and Mexican-American males and females, and to determine the extent to which random error biases regression coefficients and differences in coefficients in structural models.

PREVIOUS INVESTIGATIONS

As I discussed above, it is already known that random measurement error biases regression coefficients and attenuates explained variance in structural models. It is also known that it biases differences in regression coefficients, when two groups are measured with differential error. The present investigation sought only to determine the degree of bias in typical measures of aptitudes for different sociocultural groups.

It was hypothesized that different sociocultural groups are measured with differential reliability by valid indicators of aptitudes. This general hypothesis was based in part on the consistent finding that some of these groups score lower than others on aptitude tests and that those with lower aptitudes tend to guess more often, and, as a group, have smaller ranges of scores than groups with higher aptitudes. The literature reviewed below consists of empirical studies that indicate, first, whether aptitude indicators tend to be valid measures of the same construct across different sociocultural groups, and, second, whether the reliability of valid indicators tends to be the same across groups. Empirical studies most helpful in formulating specific hypotheses for this research were those that have compared factor structures and loadings of aptitude tests for males and females, or for the three ethnic groups of interest. Equality of factor structures and loadings for diverse sociocultural groups has been used as a means of establishing the fairness of the use of tests for culturally diverse persons (Reschuly, 1978).

The majority of studies of factorial invariance reviewed below used exploratory factor analysis procedures to extract factors, and various indices of similarity to compare structures and loadings. One or more of three types of comparisons were made. The first was the number of factors extracted. Coefficients of congruence between matching factors, or some similar index, were computed to determine whether the factors could be said to be the same across groups.

The next type of comparison was of factor relationships. Factor relationships were compared in one of several ways. Either the factor correlations, the loadings of all indicators on the first unrotated factor, or the results of a second order factor analysis were compared among groups. The third type of comparison was of the similarity in the magnitudes of factor loadings among groups.

The first two types of comparisons indicate whether indicators are valid for the same constructs across groups, and the third indicates, indirectly, whether indicator reliabilities are the same across groups. Because the results of these studies were based on exploratory factor analysis, there were no inferential tests of structures and loading magnitudes, or of equality of structures and loading magnitudes across groups. Furthermore, Alwin and Jackson (1981) point out that comparisons based on exploratory procedures mask some distinct issues in factorial invariance.

Alwin and Jackson (1981) demonstrate that the regression coefficients (loadings) in exploratory factor analysis, because they are standardized, are affected by the variances of both the factors and the indicators. This is in comparison to confirmatory factor analysis procedures, with which invariance of the loadings, the variance-covariance structure of the factors, and the variance-covariance structure of the error variances can be investigated separately, thereby allowing one to determine the source of differences in reliability.

These drawbacks to exploratory factor analysis limit the usefulness of studies using it to formulate hypotheses for studies of group differences in measurement models. The value to this research of studies of factorial invariance using exploratory factor analysis was primarily in the information they provided on the validity of construct indicators across groups. Relating findings, other than those about validity, was also restricted by the failure of some studies to report test means for groups, and because a variety of ages, types and sizes of samples, and types of tests were used in the studies.

Atkin et al. (1977) factor analyzed 16 cognitive measures represented by subtests of the Sequential Tests of Educational Progress, the School and College Ability Tests and the Tests of General Information for a sample of 1817 black and white males and females. The groups were followed longitudinally and separate analyses were done for grades 5, 7, 9 and 11. Findings for the 11th grade are most relevant to this research. The number of factors was found to differ for the four (sex by race) groups in 11th grade when principal axes were extracted using squared multiple correlations in the diagonal, and the parallel analysis criterion for number of factors was applied. Fewer factors were found for both black males and females than for white males and females. Factors were rotated to oblique simple structure. Factors were less well defined for white females and both black groups than for white males. Loadings on

one general factor extracted in a second order factor analysis, however, were quite similar for 11th grade males and females, collapsed across race. The difference in factor differentiation for males and females was attributed to differential growth of interests, knowledge and skills caused by pressures to assume sex roles. The difference between number of factors for whites and blacks was left unexplained and additional research using larger and equal numbers of blacks and whites was suggested.

Hennessy and Merrifield (1976) compared structures and loadings for 2985 high school seniors planning to enter a community college in the City University of New York. Factor structures and loadings for 10 subtests of the Comparitive Guidance and Placement Program battery were compared for blacks, Hispanics, Jewish and Caucasian-Gentile groups. A principal components analysis with squared multiple correlations in the diagonal, and Kaiser's criterion to determine number of factors, yielded three factors in each of the four groups. Rotation was to an oblique solution by the Oblimin procedure. To determine the degree of similarity in patterns of loadings, a target rotation was done. Each factor matrix was rotated against the matrices of the other groups until all six possible comparisons were made. Correlations between like factors approached unity for all comparisons; therefore it was concluded that the patterns of factor loadings were similar for the four groups. A deviation index (an algebraic summing of the residuals resulting from the least-squares

prediction of one group's matrix from the target matrix) indicated the degree of similarity in magnitudes of loadings. The magnitudes of loadings were concluded to be different, but not markedly so. Jewish and Caucasian-Gentile groups had the most similar loadings, while blacks and Hispanics had the most dissimilar. For their particular sample of community college bound high school seniors, Hennessy and Merrifield concluded that the test battery had a high degree of cross-ethnic validity.

A number of studies of factorial invariance among ethnic groups have investigated subtests of the Wechler Intelligence Scale for Children-Revised (WISC-R). Reschuly (1978) obtained WISC-R subtest scores for a stratified random sample of size 950 with approximately equal numbers of Anglos, blacks, Chicanos and Native-American Papagos. There were approximately even numbers of males and females and equal numbers of children in grades 1, 3, 5, 7 and 9 in each group. A preliminary principal components analysis was used with 1s in the diagonal, and the criterion of eigenvalues greater than 1, to determine number of factors for each of the groups. Rotation to an orthogonal solution was accomplished by the Varimax method. An unrestricted maximum likelihood analysis with a χ^2 goodness-of-fit test was then conducted for each group to provide a statistical test for models with different numbers of factors. A two factor solution was indicated for blacks and Papagos, and a three factor solution for Anglos and Chicanos. Coefficients of congruence

for a two factor solution were found to be high for all four groups (.86 to .99). Three methods were used to extract a general factor -- principal factor, principal components and restricted maximum likelihood analysis. Regardless of the method, proportions of variance attributable to a general factor were approximately the same for all groups, and similar to the standardization sample.

Dean (1980) conducted a study of the WISC-R subtests similar to Reschuly's (1978). Dean compared factor structures for 109 Anglo children and 123 bilingual Mexican-American children. A principal factors solution with iterated communalities was used to extract factors. Those with eigenvalues greater than 1 were rotated using the Varimax method. Three factors emerged for each group and were judged to be the same as in the Reschuly (1978) study. Coefficients of congruence among the three factors for the two groups indicated similar solutions and factor structures for the two groups. The coefficients ranged from .84 to .88 and were smaller than in the Reschuly (1978) study. Dean concluded that from a construct validity point of view, his investigation failed to indicate that the WISC-R was an unfair test for Mexican-American children.

A third study of the WISC-R by Gutkin and Reynolds (1980) compared factor structures for 142 Chicano and 78 Anglo children who were referred for psychological services. A principal factor analysis with multiple correlations as initial communality estimates and a criterion for number of factors of eigenvalues greater than 1,

indicated two factors for each group. These were rotated using the Varimax method. Coefficients of congruence for the first unrotated principal factor extracted as a measure of general intelligence was .99. Although the two factor solution differed from findings of other studies (Reschuly, 1978; Dean, 1980), Gutkin and Reynolds pointed out that in earlier studies of the WISC-R using varied samples, the two factor solution had always been the most stable. Gutkin and Reynolds concluded that their results support the conclusion that the WISC-R has construct validity for Anglos and Chicanos referred for psychological services.

The remaining two studies reviewed did not use exploratory procedures and provide more reliable information on differences in test reliabilities among groups. The only related study of factorial invariance using confirmatory factor analysis procedures was conducted by McGaw and Joreskog (1971). Simultaneous factor analysis in multiple populations (Joreskog, 1971b) was used to confirm similarities and differences in factor structures, relationships, loadings and error variances among groups. Although the study did not compare measurement models for sociocultural groups, it did compare models for groups of different ability and socioeconomic status. Since blacks and Mexican-Americans consistently score lower on ability tests than whites, and are generally of lower socioeconomic status, the findings of this study are relevant in formulating hypotheses in the present study. McGaw and Joreskog

(1971) used data from the Project Talent study of 11,743 high school students. Subjects were divided into four groups formed from the four combinations of high and low intelligence (IQ) and high and low socioeconomic status (SES). Twelve cognitive measures were used in the analysis. To determine number of factors for the entire sample, an unrestricted maximum likelihood analysis indicated which tests would load on which factors. The fit of the model with tests loading on four factors was satisfactory. The restricted model with correlated factors was then tested for fit to the data for the four subgroups simultaneously. The fit was judged to be reasonably good. A test of equality of the four factor dispersion matrices found them to be statistically different, thus indicating that the relationships between factors was not the same for the four samples. Error variances also differed for the four groups. The error variances reported were consistently larger for the low IQ groups.

Wolfle and Lichtman (1981) obtained measurement error-free regression coefficients for educational attainment models for whites, blacks and Mexican-Americans using respondents in the National Longitudinal Study of the High School Class of 1972. The measurement model used four ability subtests, which were assumed to load on one general ability factor; the same model was used for all three groups. No statistical test was attempted to determine if the measurement models differed across groups. It was assumed that they did differ, and, in the estimation of the structural models,

factor, or true score, variances for ability, and all other constructs, were allowed to differ across groups. Reliability coefficients for the ability subtests were consistently smaller for blacks and Mexican-Americans than for whites, even though scores corrected for guessing were used in the analyses. (Scores are corrected for guessing to decrease the effect on reliabilities of guessing by lower aptitude test takers.) Females were combined with males in the final model comparisons; however, unreported reliability coefficients for the ability tests for white females were lower than for males on the tests of mathematical ability and letter groups, and higher on the test of reading ability.

The studies reviewed above indicated that aptitude indicators are generally valid for the same constructs for different groups, although there was some evidence to suggest that differentiation of some aptitudes is less for blacks than whites or Mexican-Americans (Atkin et al., 1977; Reschuly, 1978); and less for females than males (Atkin et al., 1977). With regard to reliabilty, unreported findings of Wolfle and Lichtman (1981) suggested that females are measured less reliably than males on some tests (mathematics and letter groups) and more reliably on other tests (reading). In general, blacks and Mexican-Americans appear to be measured less reliably by aptitude indicators than whites. The unreported findings of Wolfle and Lichtman (1981) demonstrated this directly; those of McGaw and Joreskog (1971) demonstrated this indirectly since they classified

groups based on IQ and socioeconomic status rather than sociocultural characteristics. Other studies simply indicated that the magnitudes of factor loadings were different for whites, blacks and Mexican-Americans (Hennessy and Merrifield, 1976), or that the patterns of loadings were different (Reschuly, 1978; Dean, 1980).

These empirical findings lead to the expectation that females and males would be measured with differential reliability by certain aptitude tests and that blacks and Mexican-Americans would be measured less reliably than whites on aptitude tests in general. These findings were used in the formulation of specific hypotheses detailed in the following section.

NATURE OF THE STUDY

To eliminate the problems in the use of exploratory factor analysis to determine if aptitude tests are valid and equally reliable for different gender and ethnic groups, this research used maximum likelihood confirmatory factor analysis procedures to make these determinations. Using an aptitude test battery administered to a national sample of senior high school students from the High School and Beyond study (NORC, 1980), a hypothesized measurement model structure was tested for fit to the data for six groups -- white males and females, black males and females and Mexican-American males and females -- to determine if the tests were valid indicators of the same constructs across groups.

Using a generalization of confirmatory factor analysis called simultaneous factor, or covariance structure, analysis in multiple populations, a determination was then made as to whether measurement model parameter estimates (loadings, error variances, and factor variances) in the models for the six groups were equal. The effect of group differences in parameter estimates on indicator reliabilities was assessed.

Next, using a simple structural model, regression coefficients uncorrected for random measurement error and then corrected for random error were obtained and compared for each of the six groups to determine the extent of bias that results when random measurement error is not accounted for. The uncorrected (biased) coefficients were obtained by least-squares regression and the corrected (unbiased) coefficients were obtained by covariance structure analysis. Finally, regression coefficient differences uncorrected for differential random measurement error were compared to coefficient differences corrected for differential error. This comparison indicated the extent to which differences in regression coefficients are biased when two groups are known to be measured with differential random error by aptitude indicators. Differences obtained by least-squares regression were compared to differences obtained by simultaneous covariance structure analysis.

The structural equation model used to obtain regression coefficients included the aptitude constructs as independent variables

and a measure of high school grades as the dependent variable. The model was necessarily simple in order to make comparisons of regression coefficients corrected and uncorrected for random measurement error in aptitude indicators.

The simplicity of the structural model was a result of focusing primary attention on the measurement properties of the aptitude indicators. The structural simplicity, while necessary, thus has an affect upon the substantive conclusions that can be drawn from this part of the analysis. For example, substantive conclusions about the effects of aptitudes on grades for the subgroups are limited by failure to control for factors such as socioeconomic status, which may be differentially confounded in the relationship between aptitude and grades for the different groups.

Data on high school seniors were taken from the High School and Beyond study, a national, longitudinal study of high school seniors and sophomores sponsored by the National Center for Education Statistics. The High School and Beyond study was chosen for several reasons. Use of a national probability sample maximizes the external validity of the study. The data set provided substantial numbers of blacks and Mexican-Americans. Many studies are handicapped by small sample sizes for these groups. The data set also provided multiple indicators of aptitudes, which were needed to determine validity and reliability of the indicators.

Finally, the High School and Beyond study data have only recently become available to researchers. To date the only major study using these data has been Coleman's (1980) study of public versus private schooling. Since the data have been collected as the first wave in a longitudinal study of educational progress, there will no doubt be a rush to examine the progress of women, blacks and Mexican-Americans vis-a-vis that of white men. It is imperative that researchers using these data be aware of the extent to which measurement models are the same across groups, and of the biasing effects of random measurement error on regression coefficients and differences in coefficients across groups.

Educational policy decisions are often made based on findings from data of this scope. Substantive conclusions should not be made lightly. When they are based on models whose constructs have not been fully investigated, there is a high probability that they are faulty. Techniques are now available to aid the researcher in estimating and comparing measurement and structural models with more precision than ever before. To ignore these techniques when they can be used is to invite unnecessary controversy. By using these techniques to estimate and compare aptitude measurement models for the groups of interest, this research not only provides information on measurement models for the High School and Beyond study aptitude data, but also demonstrates the latest procedures for the examination of measurement models for other constructs, and for fitting and comparing structural models.

The following research hypotheses were investigated:

1. H_a: There is no difference in the structure of measurement models for the six groups.

This hypothesis concerns the construct validity of the aptitude indicators. Do the same indicators load on the same factor for all groups and is the relationship between factors the same for all groups? The empirical evidence suggested that construct validity would be the same for the different groups.

2. a) H_a: Within gender groups, blacks and Mexican-Americans will be measured less reliably than whites by aptitude indicators on which they have lower scores.

This hypothesis was based on empirical research and the common finding that blacks and Mexican-Americans score lower on aptitude tests, in general, than whites, and that lower scores are associated with lower reliability coefficients.

2. b) H_a: Within ethnic groups, females will score lower and will be measured less reliably than males on mathematical aptitude indicators, and will score higher and be measured more reliably on verbal aptitude indicators.

This hypothesis was based on empirical research and the common finding that the two groups score differently on different aptitude subtests, and that lower scores are associated with lower reliabilities.

3. H_a: Within groups, regression coefficients will be biased and explained variance (R^2) will be attenuated in structural

models in which the coefficients have not been corrected for random measurement error. The extent of bias produced by typical aptitude indicators for the different groups was of interest in testing this hypothesis.

In the bivariate case a parameter estimate biased because of measurement error is attenuated. With multiple independent variables estimates may be biased in different ways, but explained variance in the dependent variable will be attenuated.

4. H_a: For groups measured with differential reliability by aptitude construct indicators, differences in regression coefficients will be biased in structural models in which regression coefficients have not been corrected for differential error.

If was expected that some of the groups in the study would be measured with differential reliability by some of the aptitude indicators, and, therefore, that differences in regression coefficients would be biased. If groups are measured with equal reliability by aptitude indicators, differences will be unbiased, even though within group regression coefficients are themselves biased. The extent of bias in regression coefficient differences was the primary interest in testing this hypothesis.

CHAPTER II

THE HIGH SCHOOL AND BEYOND DATA FILE

THE SAMPLE

Data for this study were taken from a national longitudinal study of high school sophomores and seniors, sponsored by the National Center for Education Statistics. The initial survey was conducted in 1980 by the National Opinion Research Center (NORC, 1980), and has come to be called "High School and Beyond" (HSB). Follow-ups are in progress for 1982 and being planned for 1984. Students were selected by a stratified two-stage probability sample of 1,015 high schools; 36 seniors and 36 sophomores were then selected from each school. The samples represent 3,800,000 sophomores and 3,000,000 seniors in more than 21,000 schools.

Schools were stratified by type (public, Catholic private, non-Catholic private, and alternative), racial and ethnic composition, enrollment, region, and location (central-city, suburban and rural). Some strata were oversampled to allow a sufficient number of respondents for special analyses of subgroups of students or schools. School and student responses were weighted to correct for over-sampling and non-response. Nonetheless, some bias may remain due to unknown differences between respondents and non-respondents.

In the 1980 base-year survey, questionnaires and cognitive tests were group administered to the students in the sample. The student questionnaire covered school experiences, activities, attitudes, plans, background characteristics and language proficiency.

Only data for the senior sample were used for the present study. Data were collected for 28,240 seniors. Table 2.1 shows the number of cases remaining in each subgroup (sex by ethnicity) after a listwise deletion of all cases for which there was incomplete information on the variables used in the study.

VARIABLES

The aptitude tests used in the analyses were developed by the Educational Testing Service (ETS) and were described by Heyns and Hilton (1982). All tests were administered with warnings that scores would be corrected downward for guessing. Low scores represent low aptitude and high scores, high aptitude. All tests were determined by ETS to be unspeeded, with the exception of the mosaics tests.

The tests used were:

1. Vocabulary 1 (15 items, 5 minutes). A brief test using synonym format. Items were selected to avoid academic or collegiate bias and to be of an appropriate level of difficulty for the twelfth grade population.

Table 2.1 Number of Cases in Each Subgroup

	Ethnic Group		
Sex	White	Black	Mexican-American
Male	6463	647	280
Female	7026	786	321

2. Vocabulary 2 (12 items, 4 minutes). A test with the same format as Vocabulary 1. The purpose was to increase the range of measurement of the first test, i.e. easier and harder items were added.
3. Reading (20 items, 15 minutes). A test based on short passages (100-200 words) with several related questions concerning a variety of reading skills (analysis and interpretation), but focused on straightforward comprehension. In combination with the Vocabulary 1 test, it provides a means to derive a verbal score which can allow links to normative data available for the Standardized Achievement Test (SAT).
4. Mathematics 1 (25 items, 15 minutes). Quantitative comparisons in which the student indicates which of two quantities is greater, or asserts their equality or the lack of sufficient data to determine which quantity is greater. This type of item is relatively quickly answered and provides measurement of basic competence in mathematics.
5. Mathematics 2 (8 items, 4 minutes). A test similar to the Mathematics 1 test, but with more difficult items. This test was designed to assess mathematics achievement and, therefore, is more curriculum-specific.
6. Picture-Number (15 items, 5 minutes). A test measuring short-term memory for nonsense pairings.

7. Mosaics 1 (56 items, 3 minutes). A test which measures perceptual speed and accuracy through items that require that small differences be detected between pairs of otherwise identical mosaics or tile-like patterns. This test is deliberately speeded.
8. Mosaics 2 (33 items, 3 minutes). Identical to the Mosaics 1 test except more difficult.
9. Visualization in Three Dimensions (16 items, 9 minutes). A test measuring the ability to rotate objects in mental space.

The dependent variable, grades, was determined by asking seniors to respond to the question, "Which of the following best describes your grades so far in high school?" Possible responses and coding were as follows:

1. Mostly A's (or a numerical average of 90-100), coded 1.
2. About half A's and half B's (or 85-89), coded 2.
3. Mostly B's (or 80-84), coded 3.
4. About half B's and half C's (or 75-79), coded 4.
5. Mostly C's (or 70-74), coded 5.
6. About half C's and half D's (or 65-69), coded 6.
7. Mostly D's (or 60-64), coded 7.
8. Mostly below D (or below 60), coded 8.

Coding was reversed for statistical analyses so associations with aptitude would be positive.

CHAPTER III

METHODOLOGY

FITTING THEORETICAL MODELS AND ESTIMATING PARAMETERS

Techniques now available to fit measurement and structural models, and to estimate the associated parameters, have been largely ignored by social scientists, despite the fact that these techniques overcome major problems in the use of more familiar procedures. This section will trace the development of these techniques and illustrate their use to show their value to investigations such as were undertaken in this research.

Measurement Models

Before corrections for random measurement error may be considered, the measurement properties of construct indicators must first be known. The validity of a measure as an indicator of a hypothesized construct and the reliability of the measure must be established.

Construct validity is determined primarily by establishing that there is a relationship between indicator scores hypothesized to measure the same construct and a lack of relationship between indicator scores hypothesized to measure different constructs.

Ideally, a valid indicator should measure only one construct. Of validity and reliability, validity is by far the harder to determine. Although an indicator may be shown to correlate with other indicators hypothesized to be measuring the same construct, it is almost impossible to establish that it does not also measure some other construct. Only its lack of relationship with a few of the infinite number of possible constructs can be established. The validity of an indicator rests strongly on the intelligent choice of these few constructs.

The most common method for determining whether hypothesized constructs underlie a battery of tests is to subject the test scores to an exploratory factor analysis. When tests hypothesized to measure the same construct load highly on the same factor, the validity of the tests as measures of a construct is supported. If a test loads substantially on more than one factor, it cannot be used legitimately as an indicator of one construct. The meaning of the correlation between the indicator and other variables in a structural model, for example, would be ambiguous. Even when a test loads on only one factor, one cannot conclude that it is a valid measure of one construct only. There may still be systematic variance in the test that, while not explained by other factors underlying the particular test battery, may be explained by factors not revealed by the analysis. This possibly unexplained systematic variance must be combined with unsystematic random variance in an indicator and

together they are considered to be the random measurement error in the indicator. More thorough investigations of construct validity would seek further to confirm that this error does not contain systematic variance from sources outside the model.

Reliability is, theoretically, an assessment of the amount of systematic, valid variance and random error variance that are present in a construct indicator. Nevertheless, one can never be sure the error variance consists of random variation alone. Although a researcher must live with this reality in reliability estimation, traditional methods of assessing reliability require not only the assumption that measurement error is random, but also require unrealistic assumptions that can be proven to be untrue.

The dominant measurement model in use today is the parallel measurement model of classical test score theory (Lord and Novick, 1968). The classical model

$$Y = \tau + e$$

posits that an observed variable Y is due to an underlying true score, τ, and an error component, e. It is assumed that Y and e are uncorrelated and the mean of the errors is zero. In this discussion, Y corresponds to a subject's score on an aptitude indicator; τ to the indicator true score; and e to the random measurement error in the indicator. To determine reliability, the model requires that one have two parallel measures of the same

construct. Parallel measures must have the same metric and must have equal variances. The reliability coefficient is obtained by correlating the two parallel indicators. The correlation represents the amount of variance explained by a hypothesized construct in each of the indicators; therefore, this variance, called true score variance, and random error variance in the two indicators are assumed to be equal.

The parallel measurement model is too restrictive. It is not an estimate in the sense that it is recognized that different parallel measures will correlate differentially, and that each correlation is only an estimate of reliability. The model does not allow for unequal variances in indicators, and does not recognize that the indicators may have different true score variances because they are differentially related to the construct.

A slightly more realistic measurement model is that which defines tau-equivalent measures (Lord and Novick, 1968). Tau-equivalent indicators are assumed to have equal true score variances, but may themselves have unequal variances, and, therefore, unequal error variances. Cronbach's (1951) coefficient alpha provides an estimate of reliability that uses all the variance and covariance information of the separate indicators of the same construct. It is calculated using the following formula:

$$\alpha = \frac{N}{N-1}\left[1 - \frac{\Sigma\sigma^2(Y_i)}{\sigma_X^2}\right]$$

where N is the number of indicators, $\Sigma\sigma^2(Y_i)$ is the sum of indicator variances, and σ_X^2 is the variance of the total composite of indicators. Unlike the parallel measurement model reliability coefficient, the tau-equivalent reliability coefficient is considered an estimate in recognition of the fact that the use of different sets of indicators will produce different coefficients. Conceptually, coefficient alpha is the average correlation among the indicators. Like the reliability coefficient of the classical measurement model, however, coefficient alpha does not allow indicators to be differentially related to the construct.

To overcome this problem, researchers began to use exploratory factor analysis not only to aid in establishing validity, but also to estimate reliability without the restrictions imposed by the parallel or tau-equivalent measurement models (Zeller and Carmines, 1980). With factor analysis, the squared standardized loadings of tests on a factor are assumed to represent the variance in the tests explained by the factor. The reliability of the composite of tests loading on the same factor is obtained by calculating coefficient theta or coefficient omega which use the differential loadings in their calculation. The use of exploratory factor analysis to estimate reliability presents its own set of problems, however. Except for maximum likelihood exploratory factor analysis (Joreskog, 1967), the procedures commonly used provide no inferential tests to determine whether the number of factors extracted is adequate to reproduce the correlation

matrix obtained for the indicators. Second, there is no way to establish that some loadings are not the result of sampling variability and are in fact zero in the population. In other words, there is no way to test the fit of the measurement model to the observed data. For example, if some loadings are zero in the population, a more restrictive model in which these loadings were set to zero could provide an equally well-fitting, more plausible model.

With the development of maximum likelihood confirmatory factor analysis (Joreskog, 1969), and Joreskog's (1971a) development of the notion of congeneric measures, reliability may now be determined without the requirement that measures be parallel or tau-equivalent, and hypothesized measurement models may be tested for fit.

Joreskog (1971a) defined congeneric measures as those whose true scores are perfectly correlated, but unequal, and whose variances are unequal. Therefore, true score variances and random error variance may both be unequal for congeneric measures of the same construct. With this definition a third model was added to the category of true score measurement models. By then demonstrating that all three true score models represent restricted common-factor models estimated by confirmatory factor analysis (Joreskog, 1971a; Alwin and Jackson, 1979), Joreskog linked reliability estimation to confirmatory factor analysis.

Confirmatory factor analysis differs from exploratory factor analysis in several ways. With confirmatory factor analysis one begins

with an hypothesis about relationships between construct indicators and factors. The hypothesized model is then tested for fit to the variance-covariance matrix of the indicators and unique measurement model parameter estimates are obtained. To obtain unique parameter estimates and a X^2 value for the test of fit, confirmatory factor analysis requires that constraints be placed on the hypothesized model, i.e. not all indicators can be allowed to load on all factors, which is the case in exploratory factor analysis. By placing these constraints one "identifies" the model. The minimum condition for identification is that the number of correlations is equal to or greater than the number of parameters to be estimated in the model.

A model with such constraints is called a restricted common-factor model. The model obtained from an exploratory factor analysis is an unrestricted common-factor model.

Joreskog (1971a) linked reliability estimation to confirmatory factor analysis by recognizing that, when an indicator is constrained to load on one factor only, the assumption of true score models that only one construct underlies an indicator is satisfied. If the indicator is found to load on one factor only, the remaining variance may be assumed to be random, another assumption of true score models. Each of the three types of true score models -- parallel, tau-equivalent and congeneric -- requires different constraints on the model. The congeneric model requires only that indicators for a factor load only on that factor. The tau-equivalent model requires that all indicators

loading on a factor be constrained to have equal loadings (true score variances). The parallel model requires that all indicators loading on a factor have equal loadings and error variances. (Parallel measures must also have the same metric.) When measurement models for several constructs are investigated simultaneously, as is usually the case when a test battery is analyzed, further constraints may be placed on the model to indicate that orthogonal and oblique relationships between constructs are being hypothesized.

Once a model has been fit, the reliability of the separate indicators, and the composite of indicators for a construct, may be obtained. When a variance-covariance matrix is being analyzed, the reliability of each indicator is

$$\lambda_i^2\sigma_\tau^2 \;/\; (\lambda_i^2\sigma_\tau^2 + \sigma_{\varepsilon i}^2)$$

where λ_i is the loading of the indicator on the factor, σ_τ^2 is the variance of the factor, or true score variance, and $\sigma_{\varepsilon i}^2$ is the indicator random error variance. The reliability of the composite of indicators loading on the same factor is computed using a formula determined by the type of model fit (see Alwin and Jackson, 1979).

A typical model for a hypothetical five test battery with two correlated constructs hypothesized to underlie it will illustrate the parameters that can be estimated using confirmatory factor analysis. In Figure 1 the model specifies that the two construct factors, ξ_1 and

ξ_2, are correlated. The correlation is represented by ϕ. The first two X variables are hypothesized to measure, or load on, ξ_1 only. The next three X variables are hypothesized to measure ξ_2 only. The λs represent these loadings. The absence of arrows from X_1 and X_2 to ξ_2, and from X_3, X_4 and X_5 to ξ_1, represent constraints on the model based on the hypothesis that these loadings are zero. If variables X_1 and X_2 are hypothesized to be parallel measures, λ_1 and λ_2 are constrained to be equal. Variables X_3 - X_5 could be hypothesized to be congeneric; therefore, the associated λs are not constrained. The δ s represent random measurement error, and in the case of the parallel measures, are constrained to be equal. All other error variances are unconstrained and, because the errors are hypothesized to be random, the error covariances are constrained to be zero.

The remaining assumption of the model is that true scores are uncorrelated with the error terms. The hypothesized measurement model, in deviation form, can be described by the following structural equations:

$$x_1 = \lambda_1 \xi_1 + \delta_1$$

$$x_2 = \lambda_2 \xi_2 + \delta_2$$

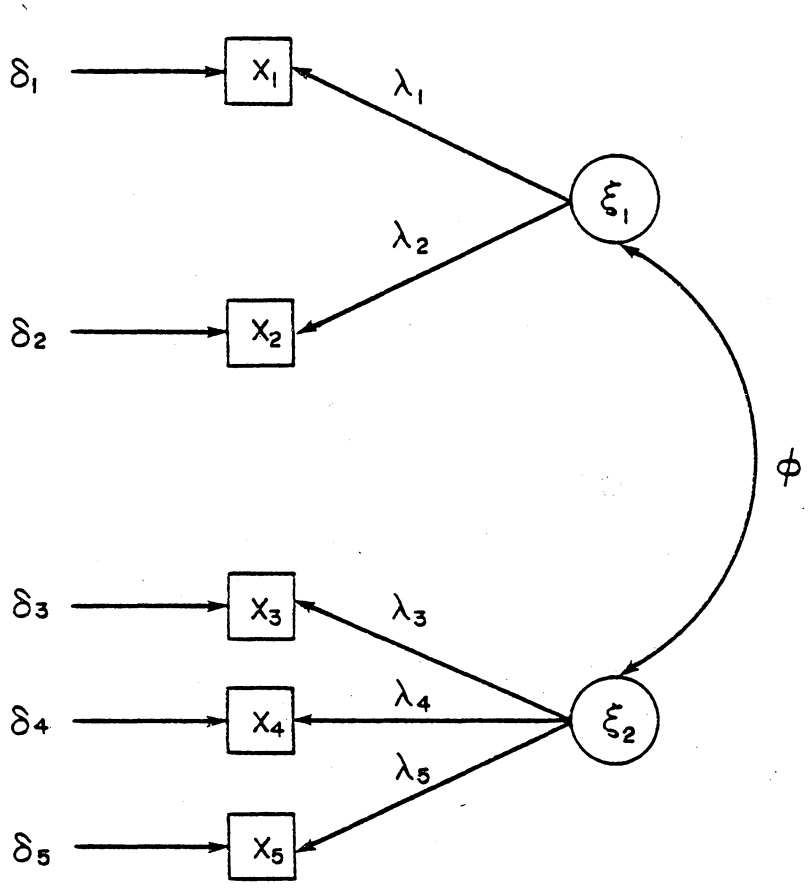

Figure 1

Hypothetical true score measurement model

where x_1 and x_2 have the same metric and variances, $\lambda_1 = \lambda_2$ and $\delta_1 = \delta_2$.

$$x_3 = \lambda_3 \xi_3 + \delta_3$$

$$x_4 = \lambda_4 \xi_4 + \delta_4$$

$$x_5 = \lambda_5 \xi_5 + \delta_5$$

To test the fit of the model, the estimated variance-covariance matrix, $\underset{\sim}{\Sigma}$, obtained by imposing the hypothesized model, is fit to the sample matrix $\underset{\sim}{S}$. Degrees of freedom for the X^2 goodness-of-fit test are $d=1/2(p)(p+1)-t$, where p is the number of X variables and t is the number of unknown parameters in the model.

Joreskog (1971b) has generalized the confirmatory factor analysis procedure to allow one to test whether a measurement model and its parameter estimates are the same in two or more populations. Simultaneous factor, or covariance structure, analysis of multiple populations (Joreskog, 1971b) provides a X^2 goodness-of-fit test for group differences in structure or parameter estimates, or both. One may constrain structures and all of the parameters to be equal across groups, and if the models are found to be different, one may relax various constraints to determine where the models are the same and where they are different.

Structural Models

Like a measurement model, a hypothesized structural model may also be fit to a variance-covariance matrix of variables in the model (Joreskog, 1970). Covariance structure analysis (Joreskog, 1970) estimates parameters for the structural model and provides a X^2 goodness-of-fit test of the model.

To estimate a structural model free of random measurement error, a measurement model may be estimated simultaneously with a structural model that hypothesizes relationships among the latent factors, rather than among the indicators which contain random error. Figure 2 illustrates a typical structural model to be estimated from factor, or true score, variances. The left side of the model is the measurement model for the exogenous, or independent, variables. In this model the X variables are hypothesized to be congeneric measures of their respective latent factors; therefore, the only constraints on this measurement model are that X_1 and X_2 have zero loadings on ξ_2, X_3 and X_4 have zero loadings on ξ_1, and all error covariances are zero. The Y_1 and Y_2 variables are indicators of the endogenous, or dependent factor, η_1; Y_3 and Y_4 are indicators of a second endogenous factor, η_2. The λ_y parameters are the loadings of the Y variables on these endogenous factors. Each indicator is hypothesized to be measuring the construct with random error represented by $\varepsilon_1 - \varepsilon_4$. The constraints on the measurement model for the Y variables are hypothesized to be the same as those for the

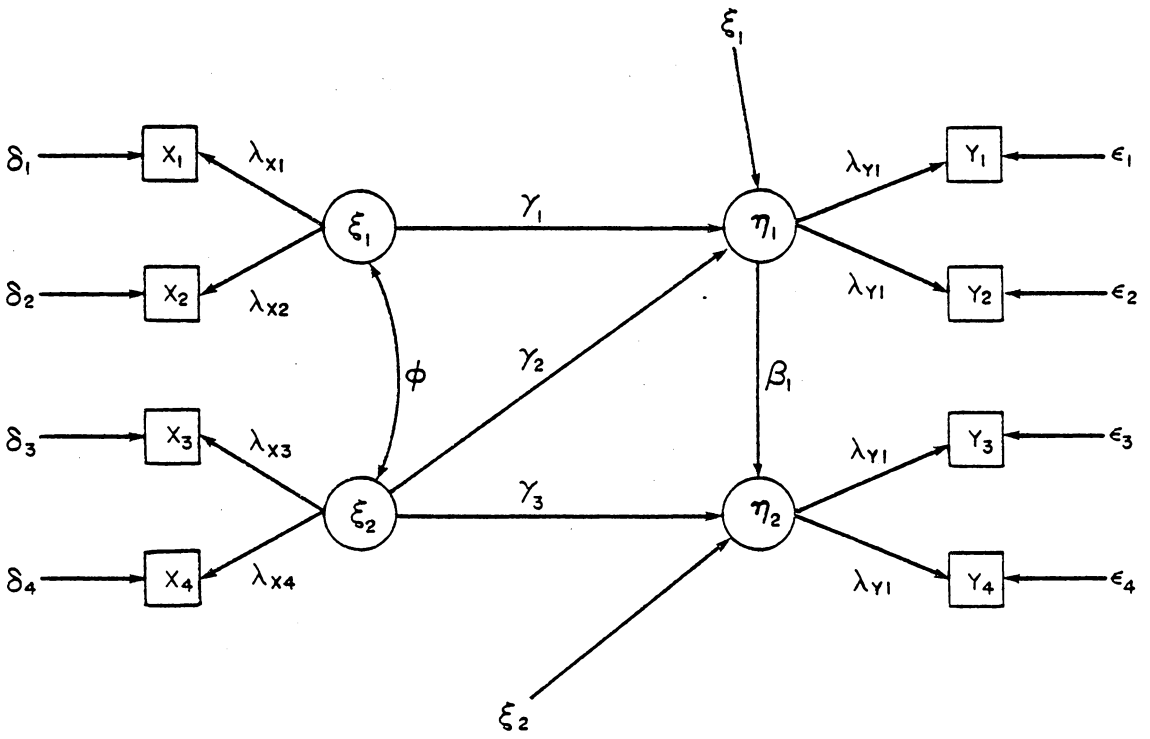

Figure 2

Hypothetical structural model estimated from true score variances

X variables The residuals for the dependent latent factors are represented by ζ_1 and ζ_2. The paths, or regression coefficients, of the exogeneous variables on the endogenous variables are represented by γ_1 to γ_3. The path from ξ_1 to η_2 is here constrained to be zero for substantive reasons. The parameter β_1 is the path expressing the effect of η_1 on η_2.

Other assumptions of the model are that the residuals are uncorrelated with true scores, and the errors of measurement are uncorrelated with true scores for the X variables (ξ), the Y variables (η) and the residuals (ζ). The measurement model in deviation form is described by the following linear structural equations:

$$x_1 = \lambda_{x_1}\xi_1 + \delta_1$$
$$x_2 = \lambda_{x_2}\xi_1 + \delta_2$$
$$x_3 = \lambda_{x_3}\xi_2 + \delta_3$$
$$x_4 = \lambda_{x_4}\xi_2 + \delta_4$$
$$y_1 = \lambda_{y_1}\eta_1 + \varepsilon_1$$
$$y_2 = \lambda_{y_2}\eta_1 + \varepsilon_2$$
$$y_3 = \lambda_{y_3}\eta_2 + \varepsilon_3$$
$$y_4 = \lambda_{y_4}\eta_2 + \varepsilon_4$$

The structural model is described by the following:

$$\eta_1 = \gamma_1\xi_1 + \gamma_2\xi_2 + \zeta_1$$
$$\eta_2 = \gamma_3\xi_2 + \beta_1\eta_1 + \zeta_2$$

The entire model is fit in the same way as described for the measurement model in Figure 1, with degrees of freedom for the X^2 goodness-of-fit test $d=1/2(p+q)(p+q+1)-t$, where p is the number of X variables, q is the number of Y variables, and t is the number of parameters to be estimated. Structural models to be compared among groups may have different measurement models or measurement models in which some parameters are constrained to be equal, not only within groups, but among groups.

The method for estimating structural models free of measurement error is superior to the commonly used method of correcting correlation matrices for attenuation. If the manifest variable correlation matrix has been corrected for attenuation, a structural model, or comparisons between structural models, cannot be tested for fit (Kenny, 1979). Using covariance structure analysis, the measurement and structural models represent one causal model and are tested for fit simultaneously to determine the extent to which they reproduce the original variance-covariance matrix.

A second advantage to estimating structural models from latent factors is that systematic variance due to correlated errors or method variance may be removed from error and factor, or true score, variance to more precisely estimate structural parameters. Also, when an indicator measures more than one construct in a model, it can be used meaningfully and unambiguously in the estimation of parameters.

Systematic variance other than that which explains a hypothesized construct is often found in indicators of psychological and sociological attributes. However, the calculation of reliability coefficients on which corrections for attenuation are based require that such systematic variance be retained with true score variance, if all indicators of a construct contain additional systematic variance from the same source; or be included with random error variance, if some but not all indicators contain additional systematic variance. In the first case true score variance is inflated, and in the latter, random measurement error is inflated. It has been noted that one can never be sure measurement error is truly random, as assumed in true score models. When it has been confirmed measurement error is not random, or that some true score variance is due to a source other than the underlying construct, separation of the sources of variance is called for.

A measurement model in which some indicators have more than one source of systematic variance is called a common-factor measurement model (Alwin and Jackson, 1979). This class of models does not constrain indicators to load on one factor only, as do all of the true score models. A common-factor model separates the different sources of systematic variance in the measurement model used to estimate structural models and provides estimates of the separate effects of the different sources of variation on dependent variables, as well as more precise estimates of random error variance. Figure 3

illustrates a typical common-factor model which may be used to estimate regression coefficients in a structural model. Variables X_3 and X_4 load on two factors. Assuming ξ_1 and ξ_3 are meaningful constructs, determined by the indicators that load on them, factor ξ_2 may represent a third construct, or variance attributable to method of testing. If the model fits, the separate effects of the constructs on the dependent variable may be determined as well as estimates of random error, free of the systematic variance caused by ξ_2.

The factor ξ_2 may also represent variance due to correlated errors; however, it is customary, if this is assumed to be the case, to allow the error variances of X_3 and X_4 to be correlated (Alwin and Jackson, 1979). In this case the model would be that seen in Figure 4. Rather than estimating paths to ξ_2, and the variance of ξ_2, this model would estimate the covariance between δ_3 and δ_4, δ_{34}.

The procedures developed by Joreskog have been available for some time; however, researchers have been slow to use them to investigate measurement models, or structural models hypothesized to answer substantive questions. Kenny (1979) attributes this to the complexity of the computer programs developed to estimate and fit models. The latest in a series of such programs is LISREL IV (Linear Structural Relationships by the Method of Maximum Likelihood) (Joreskog and Sorbom, 1978). As revisions of LISREL have occurred, the manuals have become easier to understand and simplified explanations have been provided by others (e.g., Long, 1976; Wolfle,

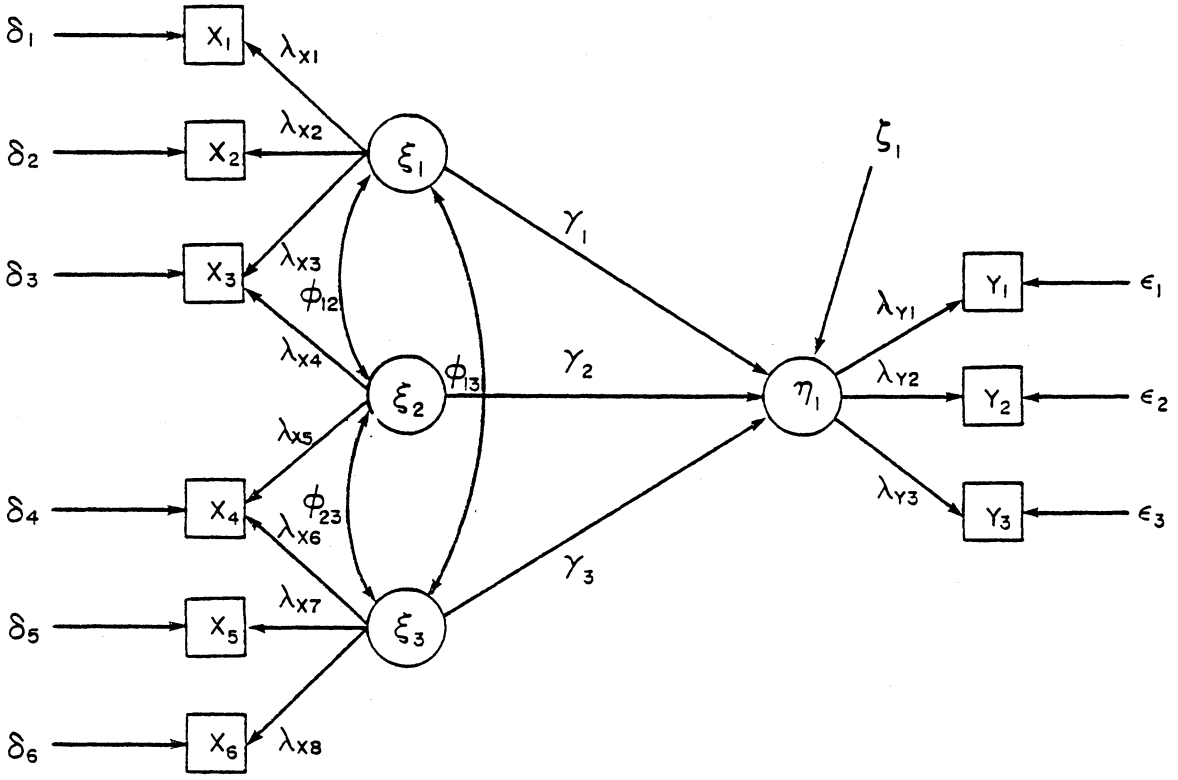

Figure 3

Hypothetical structural model estimated from a common-factor measurement model

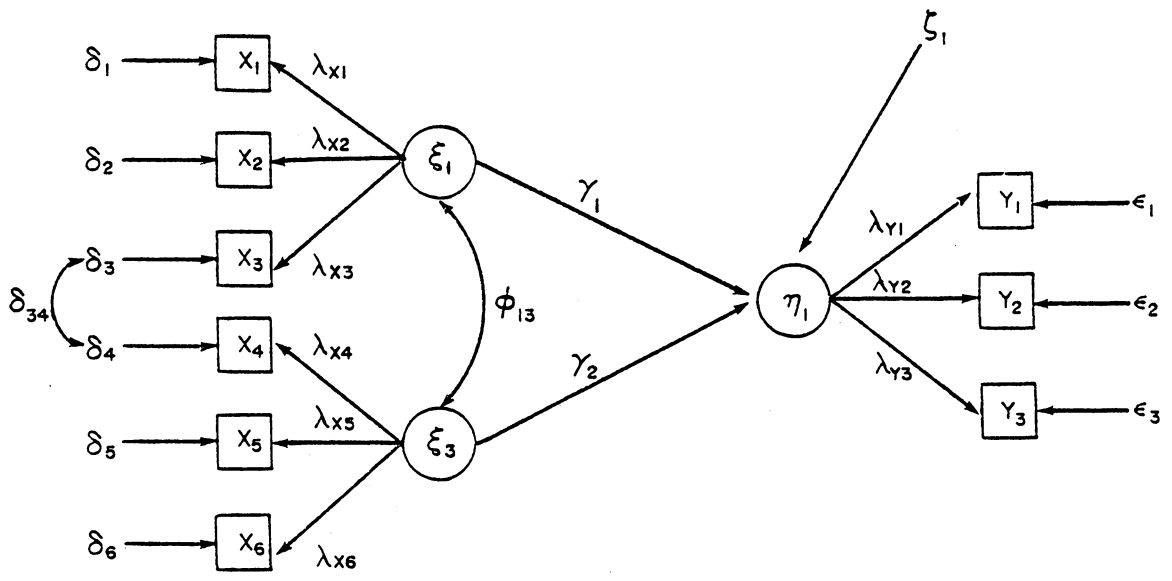

Figure 4

Hypothetical structural model estimated from a measurement model with correlated errors

1982). At present Joreskog's procedures for fitting and estimating parameters for measurement and structural models, and for fitting and estimating the two types of models simultaneously to provide error-free regression coefficients, represent the most up-to-date methods available. These procedures were used in the present research to estimate and compare measurement models in order to investigate validity and reliability questions and to estimate and compare structural models to answer questions about the biasing effect of random measurement error on regression coefficients.

CHAPTER IV

SEX AND ETHNIC DIFFERENCES IN HSB APTITUDE MEASUREMENT MODELS

DETERMINING MEASUREMENT MODEL STRUCTURES

The first hypothesis to be considered was that the six ethnic-gender groups would all possess the same aptitude indicator measurement structure. In other words, aptitude indicators would be valid measures of the same constructs across the six groups. Were an indicator not to be valid for the same construct, it would not be measuring the same trait, and would therefore not be useful for the same purpose across groups.

Before testing this hypothesis, it was first necessary to hypothesize a model to be tested. Educational Testing Service had already hypothesized and fit a model to the data for the entire HSB senior sample of 23,870 (Heyns and Hilton, 1982). The results of their analysis showed that a hypothesized three factor, true score model fit the data. The two vocabulary and the reading tests loaded only on the first factor (verbal aptitude); the two mathematics tests loaded only on the second factor (mathematics aptitude); and the two mosaics tests loaded only on the third factor (mosaics aptitude). The picture-number and visualization tests did not load highly on any of the three factors and the analysts concluded that these two tests

represented additional factors outside the model of interest. One goal of the research in the present study was to estimate error-free regression coefficients in structural models; therefore, multiple indicators of constructs were necessary to determine factor, or true score, variances. Because the picture-number and visualization tests were found to be indicators of single latent factors, they were not used in analyses for this study.

The ETS model was used as the hypothesized model in the present study and was fit to the variance-covariance matrices of the six groups separately using maximum likelihood confirmatory factor analysis. Surprisingly, the ETS model did not adequately fit the data for any of the six groups. The resulting X^2 values with 11 degrees of freedom ranged from 51.424 ($p<.029$) to 357.4323 ($p<.001$). Unfortunately, the Heyns and Hilton (1982) study did not report X^2 values for the fit of their models; as a result the adequacy of the fit of their final model cannot be assessed. All one knows is that the present factor analysis did not confirm it.

Since the ETS model did not fit the data for any of the groups, it seemed likely that the ETS model was misspecified in some way. Proceeding on this assumption, the reading test seemed most likely to be causing the lack of fit of the model since all the tests required reading skills. It seemed logical to expect that the reading test would load on all factors. The ETS model allowed the reading test to load on only one factor, verbal aptitude. It loaded highly on this factor

and the ETS analysts did not explore further. When the reading test was allowed to load on all three factors in the present analysis, the fit of the model was significantly improved for all groups. It was concluded that the reading test measured not only verbal aptitude, but the reading component of the mathematics and mosaics tests. The reading test was dropped from the model since it was found to be related to more than one construct.

When the reading test was dropped from the analysis, the model then consisted of six subtests, two vocabulary, two mathematics, and two mosaics, which were thought to be manifest indicators of three underlying factors. This three factor, true score model was tested for fit to the variance-covariance matrices of the six groups separately, and was found to fit adequately in all cases (see below). This model is shown in Figure 5.

The model shown in Figure 5 is a congeneric, true score measurement model. It is a true score model because all indicators are measures of one construct, or factor, only, and error variances for all indicators are assumed to be random. It is congeneric because no parameters for indicators of the same construct have been hypothesized to be equal. In this particular model, the three constructs are hypothesized to be correlated since all are measures of aptitudes.

Table 4.1 gives the X^2 values resulting from the goodness-of-fit tests of the three factor, true score measurement model in Figure 5 to the variance-covariance matrices of each of the six groups. A

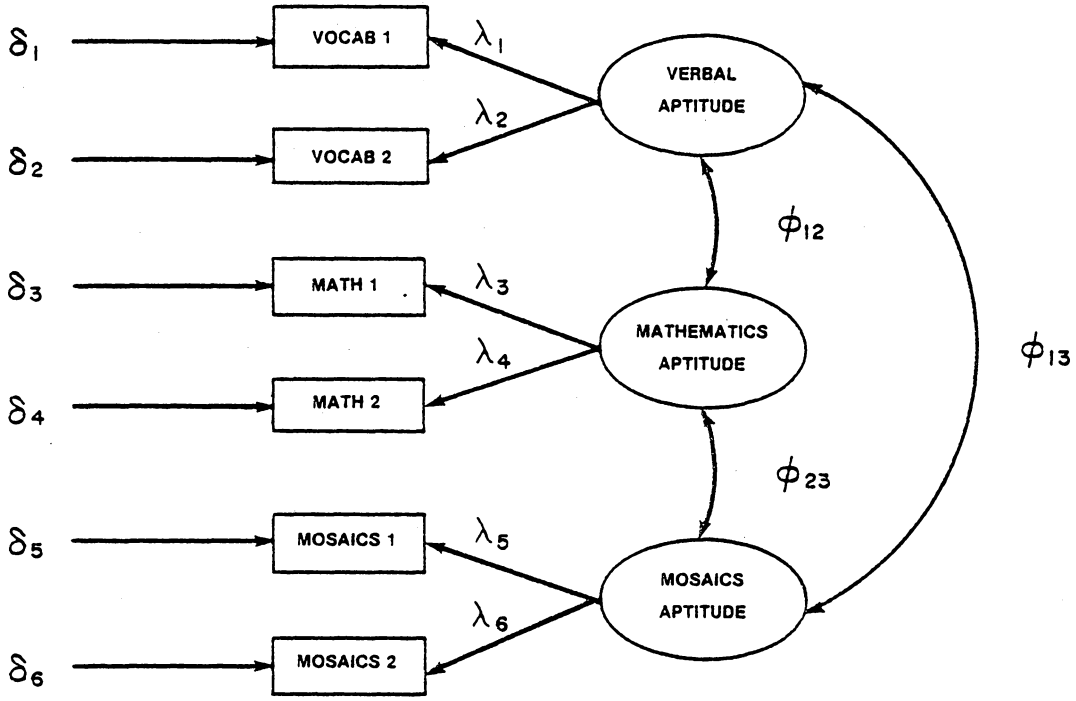

Figure 5

Congeneric, true score measurement model for six tests administered to HSB seniors

Table 4.1

Goodness-of-Fit of the Three Factor, Six Test Model

Group	X^2	df	prob.	Tucker-Lewis coefficient
White Males	3.058	6	.802	1.000
White Females	5.413	6	.492	1.000
Black Males	9.358	6	.154	.994
Black Females	14.670	6	.023	.984
Mexican-American Males	2.375	6	.882	1.016
Mexican-American Females	9.266	6	.159	.981

Type I probability error of .10 was chosen for rejection of the fit of the model since the hypothesis was that the model would fit the data. The model fit at $\alpha = .10$ for all groups except black females. Because X^2 values are sensitive to slight deviations from fit with large sample sizes, an alternate index of fit, the Tucker-Lewis coefficient (Tucker and Lewis, 1973), was used to determine if the three factor model was adequate from a practical viewpoint for the black female data. The index suggests whether the variance accounted for by a model is large or small in relation to the total variance in the data. The coefficient may be estimated by

$$\hat{\rho} = \frac{M_0 - M_k}{M_0 - 1}$$

in which $M_0 = X_0^2 / df_0$ and $M_k = X_k^2 / df_k$. The X^2s and degrees of freedom were obtained first for a maximum likelihood factor model with zero common factors; that is, each of the six manifest variables was hypothesized to have one latent factor and these latent factors were constrained to have zero covariance with any other factor. The fit of this model was then compared to the fit of a model with three common factors (shown in Figure 5). The Tucker-Lewis coefficient for black females was .984. The index is not normed necessarily to lie between zero and one, but experience has shown that values above .90

indicate that the fit of a model cannot be substantially improved (Bentler and Bonett, 1980). Based on the Tucker-Lewis coefficient calculated for the black female group, the three factor model adequately fit the data for this group, as well as the other five groups. Tucker-Lewis coefficients for all other groups were greater than .98 as well and are shown in Table 4.1.

The factor correlations for each group are shown in Table 4.2. These coefficients are similar in value for all groups except Mexican-American females. For this group, correlations have the same relationship to one another as they do for the other five groups, but the associations of mathematics aptitude with the other two aptitudes are smaller for this group than the others. This finding indicates that mathematics aptitude is not as highly related to verbal and mosaics aptitude for Mexican-American females as for the other groups. The six test means for all groups are shown in Table 4.3. From an examination of these means, it can be seen that Mexican-American females were the lowest scoring group on mathematics tests, but scored relatively high on mosaics tests in comparison to other groups. The substantially lower correlation between mathematics and mosaics constructs reflects this deviation. Why this occurs for Mexican-American females and not for any of the other five groups is an open question whose answer is beyond the scope of the present analysis.

Table 4.2

Factor Correlations

Group	Vocab w/Math	Vocab w/Mosaic	Math w/Mosiac
White Males	.726	.293	.472
White Females	.725	.238	.413
Black Males	.746	.267	.465
Black Females	.714	.294	.455
Mexican-American Males	.778	.283	.462
Mexican-American Females	.663	.268	.289

Table 4.3

Test Means and Standard Deviations

Group	Vocab1	Vocab2	Math1	Math2	Mosaics1	Mosaics2
WM	6.494 (3.984)	5.265 (3.103)	14.952 (6.415)	3.177 (2.226)	26.552 (9.931)	17.655 (6.748)
WF	6.540 (3.851)	4.689 (3.011)	13.285 (6.137)	2.617 (2.031)	28.070 (9.11)	19.645 (5.976)
BM	4.128 (3.576)	3.232 (2.792)	9.641 (6.298)	1.880 (1.968)	20.533 (10.937)	13.680 (7.466)
BF	3.697 (3.463)	2.532 (2.474)	8.046 (5.534)	1.638 (1.735)	22.160 (10.992)	14.929 (7.353)
MM	3.640 (3.561)	2.985 (2.881)	9.448 (6.673)	1.933 (2.124)	23.186 (11.028)	16.074 (8.028)
MF	3.572 (3.336)	2.687 (2.681)	8.010 (5.599)	1.698 (1.681)	25.313 (11.166)	17.858 (6.810)

Overall, these results support the first hypothesis. The aptitude indicators are valid measures of the same constructs for each of the six groups, and the constructs are related to one another for each of the groups in a similar fashion.

A further note is in order about the validity of the six tests in the three factor model. An adequate fit of a model in which each test loaded on only one factor would suggest that each test is a valid measure of one construct only. Nevertheless, it is possible that unexplained variance in the indicators attributed to random error may contain reliable variance which is explained by a construct not represented by factors in the model. One indication of whether this might be the case may be obtained by comparing reliability coefficients for the tests resulting from the factor analysis with those resulting from the calculation of KR-20 reliability coefficients.

Shown in Table 4.4 are test reliabilities, which will be discussed in detail below, and KR-20 reliability coefficients reported by ETS (Heyns and Hilton, 1982). These KR-20 coefficients were computed on a slightly different, but comparable, sample of students. KR-20 coefficients are meaningful only for unspeeded tests (see Nunnally, 1978), therefore they are not reported for the two mosaics tests. The values indicate the internal consistency of the tests (i.e., the average correlation of all items with one another) and are a measure of the amount of reliable variance in a test. The reliabilities calculated from the factor analysis solution indicate the amount of

Table 4.4

Test Reliabilities

Group	Vocab1	Vocab2	Math1	Math2	Mosaics1	Mosaics2
WM	.714	.711	.799	.566	.542	.826
WF	.660	.696	.798	.423	.477	.833
BM	.665	.640	.723	.323	.585	.937
BF	.630	.618	.699	.260	.518	.932
MM	.634	.606	.709	.431	.558	.792
MF	.565	.443	.862	.193	.430	.934
KR-20*	.75	.65	.84	.50	**	**

* KR-20 coefficients are for a different, but comparable sample of students.

** KR-20 coefficients are not reported for the mosaics tests as they are inappropriate for speeded tests (see Nunnally, 1978).

reliable variance in a test which is explained by a construct, or factor, that underlies two or more tests hypothesized to measure the construct. When the KR-20 coefficient is similar to the reliability coefficient calculated from the factor analysis, the reliable variance in a test can be considered to be explained by one construct only. If the KR-20 coefficient is larger, there is more reliable variance in a test than is explained by a single underlying construct.

The KR-20 coefficients for the four unspeeded tests are similar to the reliability coefficients obtained from the factor analysis for most of the groups. This indicates that these tests are valid measures of a single construct, and each test measures the same construct for each of the six groups.

The reliability coefficients calculated from the factor analysis are reasonably high for all tests except the Math1 and the Mosaics1 tests. The similarity between the KR-20 and factor analysis coefficients for the Math1 test suggests that the unexplained variance in this test is random error variance. No such comparison can be made for the Mosaics1 test because it was a speeded test. Therefore, it is possible that the reliability coefficients for this test are low because there is additional systematic variance in the test unexplained by the three identified constructs.

The adequate fit of the three factor, true score measurement model, which was obtained for all six groups, suggested that there were no other common factors underlying the six tests. However,

because the fit of the model was not perfect, it was still possible that another factor existed that, when defined, would improve the fit of the model further. It was decided to investigate whether there was any reliable variance in the Mosaics1 test that might be explained by a fourth factor underlying either the Mosaics1 and the mathematics tests, or the Mosaics1 and the vocabulary tests.

The search for a possible fourth factor was carried out using the data on white males and black females. These two groups were selected as representative. Given the results to be immediately reported, no further analyses along these lines were deemed necessary. First, errors for the Mosaics1 test and the Vocab1 test were allowed to be correlated. In neither of the two groups was there a significant improvement in the fit of the model. The same was found when errors between Mosaics1 and Math1 were allowed to be correlated. It was concluded that there was no reliable variance in the Mosaics1 test that could be explained by a fourth factor which was underlying any of the six tests in the model.

COMPARISONS OF MEASUREMENT MODEL PARAMETER ESTIMATES

Once it was decided that a common factor structure existed for the six ethnic-gender groups, it became possible to compare parameter estimates attached to the common measurement model to test hypotheses about group differences in indicator reliabilities.

Reliability differences not only produce varying degrees of accuracy with which a test measures a trait, they also produce differential bias in regression coefficients when indicators are used as independent variables in structural models.

The parameters estimated with LISREL include indicator loadings, error variances, and factor, or true score, variances and covariances. The indicator loadings, error variances, and factor variances are used in the calculation of indicator reliability coefficients with the formula

$$\lambda_i^2 \, \sigma_\tau^2 \; / \; (\; \lambda_i^2 \sigma_\tau^2 + \sigma_{\varepsilon i}^2 \;)$$

where λ_i is the loading of the i-th indicator on a factor, σ_τ^2 is the variance of the factor and $\sigma_{\varepsilon i}^2$ is the indicator error variance.

Hypothesis 2a stated that indicator reliabilities calculated from these coefficients would be lower for blacks and Mexican-Americans than for whites of the same sex. With regard to individual coefficients, it was expected that error variances would be consistently larger for blacks and Mexican-Americans because lower scoring groups have more opportunity for guessing, and that true score variances would be consistently smaller because of restricted ranges in scores for lower scoring groups. Loadings, however,

which are equivalent in this case to regression coefficients, were expected to be approximately the same for all groups. Similar loadings were expected because there was no reason or evidence leading to an expectation that they would vary systematically. Exploratory factor analysis studies traditionally compare only standardized loadings, not metric coefficients from which the standardized loadings are determined. The only study to compare non-standardized loadings for aptitude tests for different groups (of which I am aware) was that by McGaw and Joreskog (1971). Their study found loadings to be equal for groups that varied by socioeconomic status and level of aptitude.

Hypothesis 2b stated that females would obtain lower scores on mathematics tests and have lower reliabilities on these tests than males of the same ethnic group, and that females would score higher than males on the vocabulary tests and have higher reliabilities on these tests. It was expected, therefore, that error variances and factor variances would vary systematically on vocabulary and mathematics tests for these groups, but that loadings would not.

As hypothesized, blacks and Mexican-Americans scored lower on all tests than whites of the same sex (see Table 4.3). Females, however, scored generally lower on vocabulary tests, which was not predicted, and lower on mathematics tests, which was predicted.

Table 4.5 gives the measurement model parameter estimates from which the reliabilities were calculated. The test reliabilities for each

Table 4.5

Measurement Model Parameter Estimates

Loadings (λ)

	Tests					
Group	Vocab1	Vocab2	Math1	Math2	Mosaics1	Mosaics2
WM	1.0*	.777	1.0	.292	1.0	.839
WF	1.0	.803	1.0	.241	1.0	.867
BM	1.0	.766	1.0	.209	1.0	.864
BF	1.0	.634	1.0	.191	1.0	.897
MM	1.0	.791	1.0	.248	1.0	.867
MF	1.0	.712	1.0	.142	1.0	.899

Error Variances (δ)

	Tests					
Group	Vocab1	Vocab2	Math1	Math2	Mosaics1	Mosaics2
WM	4.542	2.775	8.287	2.151	45.187	7.943
WF	5.040	2.759	7.603	2.379	43.431	5.947
BM	4.283	2.810	10.995	2.620	49.660	3.498
BF	4.439	3.081	9.205	2.229	58.245	3.670
MM	4.460	3.269	12.937	2.569	53.700	13.368
MF	4.841	4.001	4.315	2.278	71.055	3.045

* The loading of one indicator of each factor is set to 1.0 to establish a metric for the factor.

(continued)

Table 4.5 (con't.)

Measurement Model Parameter Estimates

Factor Variances and Covariances** (ϕ)

Group	s_{11}	s_{22}	s_{33}	s_{12}	s_{13}	s_{23}
WM	11.333	32.864	53.439	14.012	7.201	19.785
WF	9.786	30.055	39.568	12.436	4.680	14.225
BM	8.507	28.670	69.693	11.646	6.524	20.814
BF	7.554	21.419	62.577	9.082	6.395	16.672
MM	8.043	31.584	67.905	12.404	6.607	21.386
MF	6.285	27.037	53.627	8.638	4.922	11.020

** Factor 1=Vocabulary, Factor 2=Mathematics, Factor 3=Mosaics.

group are shown in Table 4.4. To determine if there were any group differences in measurement model parameter estimates used to compute reliabilities, the loadings, error variances and factor variances were constrained to be equal across the six groups. A X^2 value for the fit of the model was obtained using simultaneous factor analysis in multiple populations. This X^2 value was 669.352 with 96 degrees of freedom. The X^2 obtained when all parameters were allowed to vary among groups was 44.140 with 36 degrees of freedom. A significant ΔX^2 indicates that constraining the parameter estimates to be equal across groups produces a significant deterioration in the fit of the model. The ΔX^2 was 625.212 with 60 degrees of freedom and was significant ($p<.001$) at the chosen Type I error probability of .05. This finding indicated that the parameter estimates in the measurement models were different for at least some of the groups.

To determine if different ethnic groups had equal loadings, error variances, and factor variances, measurement models were compared to each other in ethnic group pairs, within gender group. For each sex separately, the model for whites was compared to those of blacks and Mexican-Americans. The model for blacks was then compared to that of Mexican-Americans for each sex separately.

Using simultaneous factor analysis for multiple populations, parameter estimates were constrained to be equal across the two groups being compared and a X^2 value was obtained. This value was compared to one obtained when the parameter estimates were allowed

to vary. The ΔX^2 indicates whether there is a significant deterioration in the fit of the model when all estimates are constrained to be equal. The results of these comparisons are shown in Table 4.6. For a Type I error probability of .05, it was found that blacks and Mexican-Americans of the same sex had measurement models with parameter estimates which did not differ significantly. The ΔX^2 for males was 19.487 with 12 degrees of freedom ($p>.05$) and the ΔX^2 for females was 16.207 with 12 degrees of freedom ($p>.10$). For comparisons of whites to the other two groups within gender group, it was found that parameter estimates were significantly different. The ΔX^2 values for 12 degrees of freedom ranged from 54.952 ($p<.001$) to 191.543 ($p<.001$).

By constraining loadings, error variances and factor variances to be equal between groups, one determines whether test reliabilities may be considered to be equal. This is a strict and conservative test of equal reliabilities since it requires not simply that the ratio of explained and total variance for a test be equal between groups, but that the actual explained and total variance for a test be equal between groups. From the above comparisons it may be concluded that test reliabilities for blacks and Mexican-Americans of the same sex differ only because of sampling variability.

Since the parameter estimates used in the calculation of reliability coefficients were not the same for whites as compared to blacks, or for whites as compared to Mexican-Americans, within

Table 4.6

Ethnic by Sex Group Comparisons
of Measurement Model Parameter Estimates
Used in the Calculation of Reliability Coefficients

Model	X^2	df	prob.
White and black females			
Estimates equal	211.626	24	
Estimates unequal	20.083	12	
ΔX^2	191.543	12	<.001
White and Mexican-American females			
Estimates equal	105.168	24	
Estimates unequal	14.679	12	
ΔX^2	90.489	12	<.001
Black and Mexican-American females			
Estimates equal	40.142	24	
Estimates unequal	23.935	12	
ΔX^2	16.207	12	>.10

(continued)

Table 4.6 (con't.)

Ethnic by Sex Group Comparisons
of Measurement Model Parameter Estimates
Used in the Calculation of Reliability Coefficients

Model	X^2	df	prob.
White and black males			
Estimates equal	108.271	24	
Estimates unequal	12.416	12	
ΔX^2	95.855	12	<.001
White and Mexican-American males			
Estimates equal	60.385	24	
Estimates unequal	5.433	12	
ΔX^2	54.952	12	<.001
Black and Mexican-American males			
Estimates equal	31.220	24	
Estimates unequal	11.733	12	
ΔX^2	19.487	12	>.05

gender group, the test reliabilities and parameter estimates from which they were computed were examined to see if they differed in the expected directions. Test reliabilities (see Table 4.4) are generally lower for blacks and Mexican-Americans than for whites of the same sex on vocabulary and mathematics tests. An examination of the parameter estimates in Table 4.5 shows that for these tests, loadings are generally lower, error variances are generally higher, and factor variances are consistently lower for blacks and Mexican-Americans than whites of the same sex. The directions of differences in all these parameter estimates produce lower reliabilities for blacks and Mexican-Americans as compared to whites. The loadings and error variances, however, are not consistently different, but the factor variances are. This finding suggests that the lower reliabilities on vocabulary and mathematics tests for blacks and Mexican-Americans are primarily a function of smaller factor variances. Factor variances are themselves a function of test variances. An examination of the standard deviations for vocabulary and mathematics tests (see Table 4.3) shows that the standard deviations for blacks and Mexican-Americans are generally smaller than those of whites of the same sex, meaning that score ranges for blacks and Mexican-Americans are more restricted than for whites. Score ranges appear to have affected the size of factor variances, which, in turn, contributed significantly to producing lower reliabilities for blacks and Mexican-Americans on vocabulary and mathematics tests.

Reliabilities for mosaics tests are unexpectedly slightly higher for blacks and Mexican-Americans as compared to whites of the same sex, although whites scored higher on these tests. Loadings and factor variances are higher for blacks and Mexican-Americans on these tests, but error variances are higher on one mosaics test and generally lower on the other. The major contributor in producing higher reliabilities for blacks and Mexican-Americans appears, again, to be the size of factor variances. In this case, the factor variances are unexpectedly higher for the lower scoring groups, indicating that shortened ranges of scores occurred for the higher scoring instead of the lower scoring groups.

In summary, for ethnic group comparisons, within gender group, it was found that blacks and Mexican-Americans scored more similarly to each other than to whites, and had measurement model parameter estimates, and therefore indicator reliabilities, that did not differ. Whites scored higher than blacks and Mexican-Americans on all aptitude tests, and parameter estimates used in the calculation of reliability coefficients differed between whites and blacks, and between whites and Mexican-Americans. The differences in the parameter estimates produced lower reliabilities for blacks and Mexican-Americans on vocabulary and mathematics tests, and higher reliabilities on mosaics tests. Factor variances were consistently different for whites as compared to the other two groups and appear to be a major factor in explaining why blacks and Mexican-Americans

exhibit smaller reliabilities on vocabulary and mathematics tests, and higher reliabilities on mosaics tests.

Hypothesis 2a was therefore supported by the findings for vocabulary and mathematics tests. Blacks and Mexican-Americans scored lower than whites and had lower test reliabilities. For the mosaics tests, whites achieved higher mean scores, but had lower reliabilities. Differences in reliabilities appear to be primarily the result of smaller factor variances resulting from shortened score ranges -- for blacks and Mexican-Americans on vocabulary and mathematics tests, and for whites on mosaics tests.

Hypothesis 2b was that females would score lower than males of the same ethnic background on mathematics tests and higher on vocabulary tests. Reliabilites for females were hypothesized to be lower than those of males on mathematics tests and higher on vocabulary tests. It has already been noted that females scored lower on the average than males on the vocabulary tests rather than higher, but did score lower on the mathematics tests as predicted. When all parameter estimates were constrained to be equal in comparisons between males and females of the same ethnic group, the ΔX^2 values with 12 degrees of freedom, resulting from the difference in the fit of this model and the model with all parameter estimates free to vary, were found to be 212.727 ($p<.001$) for white males and females, 33.621 ($p<.001$) for black males and females, and 36.348 ($p<.001$) for Mexican-American males and females (see Table 4.7).

Table 4.7

Sex by Ethnic Group Comparisons
of Measurement Model Parameter Estimates
Used in the Calculation of Reliability Coefficients

Model	X^2	df	prob.
White males and females			
Estimates equal	221.198	24	
Estimates unequal	8.471	12	
ΔX^2	212.727	12	<.001
Black males and females			
Estimates equal	57.648	24	
Estimates unequal	24.027	12	
ΔX^2	33.621	12	<.001
Mexican-American males and females			
Estimates equal	47.989	24	
Estimates unequal	11.641	12	
ΔX^2	36.348	12	<.001

The effect of these differences are reflected in test reliabilities (see Table 4.4). Reliabilities for vocabulary and mathematics tests are generally lower for females than males of the same ethnic background. Again, as in the case of within gender ethnic group comparisons, differences in parameter estimates are the most pronounced for factor variances. The lower scoring groups, females, have slightly smaller factor variances than males. Differences in test loadings also lead to lower reliabilities for females, as the loadings are generally smaller for females than males. Error variances, however, are not consistently larger for females, as would be expected of lower scoring groups.

No prediction was made about differences in scoring or reliabilities between males and females on the mosaics tests. Females did score higher than males of the same ethnic group. Of the parameter estimates, only factor variances differed consistently between males and females. Females had lower factor variances than males. This finding is consistent with findings from ethnic group comparisons. Higher scoring groups on mosaics tests have lower factor variances resulting from shortened score ranges on these tests. Smaller factor variances, however, did not produce consistently lower reliabilities for females than males as they did for whites as compared to blacks and Mexican-Americans. Differences in the measurement model parameter estimates produced reliabilities for mosaics tests which were close in value for males and females of the

same ethnic group. The reliabilities were more similar than those for vocabulary and mathematics tests.

Hypothesis 2b was supported by these findings. Females generally scored lower and had lower reliabilities on both the vocabulary and mathematics tests than males of the same ethnic background. While females unexpectedly scored lower on vocabulary tests, as expected of a lower scoring group, they had lower reliabilities on these tests. These findings support the expectation that lower scoring groups have lower test reliabilities and suggest that, as in the case of ethnic group comparisons, the lower reliabilities result primarily because of group differences in the size of factor variances, which result from shortened score ranges for the lower scoring groups.

The finding that a construct indicator is not equally reliable for two groups suggests that the indicator is differentially related to the construct for the groups. The construct is better measured by the indicator for one group than the other. The result is that the indicator will measure one group more accurately than another. Furthermore, a measure of the relationship between the indicator and an indicator of another construct will be differentially biased because of group differences in indicator random measurement error. Although reliability differences have serious implications for the measurement of traits, and the assessment of relationships between traits, Alwin and Jackson (1981) point out that comparisons of

reliability coefficients do not answer the question of whether an indicator is similarly related to a construct for different groups. The size of a reliability coefficient is affected by the variance of the indicator. Indicator error variance may be the same for two groups, but if total variance is not the same, the reliability coefficient will be smaller for one group than the other.

The real test of whether a construct indicator is similarly related to a construct for different groups is a test of the equality of the unstandardized loadings of the indicator on the factor representing the construct. An unstandardized loading is the regression slope of a construct indicator regressed on a latent factor. If these slopes are equal across groups, it would mean that a one-unit increase in the latent true score produced an equal change in the construct indicator score for all groups; but if the slopes are not equal across groups, this would mean the indicator is differentially related to the construct for some of the groups.

Although some consistent differences were found in test reliabilities between some ethnic and gender groups, it was considered important to know whether aptitude tests are similarly related to constructs for groups with different reliabilities. We already know that blacks and Mexican-Americans of the same sex can be considered to have equal test loadings and equal reliabilities.

When measurement models with loadings constrained to be equal were compared between whites and blacks and between whites and

Mexican-Americans of the same sex (four comparisons), only in the comparison between white and Mexican-American males were the loadings found to be equal. The ΔX^2 value resulting from the difference in the model with loadings constrained to be equal and the model with all parameter estimates free to vary was 2.504 with 3 degrees of freedom ($p<.10$). The ΔX^2 values for the other comparisons were 10.262 ($p<.02$) for the comparison between white and Mexican-American females, 17.283 ($p<.001$) for the comparison between white and black males, and 14.398 ($p<.001$) for the comparison between white and black females.

For male-female comparisons, within ethnic group, the loadings were found not to differ only in the comparison between black males and females. The ΔX^2 value with 3 degrees of freedom was 4.132 ($p>.10$). The ΔX^2 values for the other comparisons were 51.558 ($p<.001$) for the comparison between white males and females, and 7.376 ($p<.05$) for the comparison between Mexican-American males and females.

These findings indicate that reliability differences between certain ethnic and gender groups truly reflect the fact that aptitude indicators are not similarly related to constructs. For groups found to have equal loadings, reliability differences are a function of different error and factor variances.

BIAS IN REGRESSION COEFFICIENTS

Hypothesis 3 stated that structural model regression coefficients uncorrected for random measurement error would be biased and that the coefficient of determination (R^2) would be attenuated. The primary interest in testing this hypothesis was to determine the extent of bias in coefficients and the degree of attenuation in R^2 for the different groups when certain aptitude indicators from the High School and Beyond study were used as independent variables. In the bivariate case, the regression coefficient uncorrected for error is attenuated. In more complex models with multiple independent variables, some coefficients may be underestimated and some overestimated, thereby leading to different substantive conclusions than when the coefficients are corrected for their differential random error.

Regression coefficients uncorrected for random measurement error were obtained by ordinary least-squares regression (OLS). The dependent variable, grades, was regressed on three aptitude tests -- Vocab1, Math1 and Mosaics1. Regression coefficients corrected for measurement error were obtained by LISREL. In LISREL regressions the dependent variable, grades, was regressed on the verbal, mathematics and mosaics aptitude factors. The factors contain only reliable variance and, therefore, the LISREL regression coefficients are not biased by random measurement error.

Table 4.8 shows the results of the two types of regressions for the six groups separately. In all six groups the coefficients of determination (R^2) were smaller in the least-squares regressions than in the LISREL regressions. The differences in R^2 ranged from .03 to .07 with an average of .05.

With regard to the regression coefficients themselves, the least-squares regression estimates underestimated the LISREL regression coefficient for mathematics aptitude and overestimated the coefficients for vocabulary and mosaics aptitude. In some cases, coefficients that were significant in least-squares regressions were not significant in the LISREL regressions. Although the sizes of the coefficients differed in the two types of regressions, for most groups, the relative importance of the different aptitudes in the explanation of variance in grades did not change. Mathematics was the best predictor of grades in all groups for both types of regressions.

Underestimation and overestimation of coefficients and attenuation of R^2 were greatest for blacks and Mexican-Americans compared to whites. For example, the mathematics coefficient was attenuated by 27% and 21% for white males and females, respectively, and was attenuated from 37% to 50% for the other four groups. The R^2 for the white groups was attenuated 17% and 14% and from 23% to 35% for the other four groups. Since reliabilities were lower for blacks and Mexican-Americans on two of the three tests used in the

Table 4.8

Within Group Comparisons of
LISREL and OLS Regression Coefficients

Procedure	Vocab	Math	Mosaics	R^2	R^2 Diff.	R^2 Atten.
White males (N=6463)						
LISREL	.043* (.100)	.129* (.511)	-.002 (-.010)	.34		
OLS	.056* (.156)	.093* (.414)	.004* (.028)	.28	.06	17%
White females (N=7026)						
LISREL	.062* (.143)	.113* (.456)	.005 (.024)	.34		
OLS	.061* (.183)	.089* (.404)	.005* (.035)	.29	.05	14%
Black males (N=647)						
LISREL	.014 (.032)	.075* (.310)	.003 (.022)	.12		
OLS	.030 (.085)	.045* (.223)	.007 (.057)	.09	.03	25%

Tests used to obtain OLS coefficients were Vocab1, Math1 and Mosaics1. Standardized coefficients are in parentheses under unstandardized coefficients.

* Coefficient significantly different from zero at α=.05.

(continued)

Table 4.8 (con't.)

Within Group Comparisons of
LISREL and OLS Regression Coefficients

Procedure	Vocab	Math	Mosaics	R^2	$\%R^2$ Diff.	R^2 Atten.
Black females (N=786)						
LISREL	-.001 (-.001)	.134* (.456)	-.004 (-.025)	.20		
OLS	.046* (.121)	.067* (.279)	.004 (.037)	.13	.07	35%
Mexican-American males (N=280)						
LISREL	-.064 (-.128)	.161* (.659)	-.009 (-.052)	.30		
OLS	.022 (.056)	.096* (.452)	.0001 (.001)	.23	.07	23%
Mexican-American females (N=321)						
LISREL	.013 (.025)	.091* (.347)	.009 (.049)	.15		
OLS	.042 (.111)	.057* (.253)	.002 (.021)	.11	.04	26%
			Average		.05	23%

least-squares regressions, greater attenuation in R^2 would be expected. Although females were measured less reliably than males on at least two of the three tests, attenuation of R^2 was not consistently greater for females than males of the same ethnic group. Differences in reliabilities between females and males were not as pronounced as between whites and blacks and between whites and Mexican-Americans.

Hypothesis 3 was therefore supported by these findings. Regression coefficients were biased in least-squares regressions and R^2 was attenuated. The extent of attenuation was, of course, dependent upon reliabilities of the aptitude tests. Coefficients were biased to a greater extent and R^2s were more attenuated for groups with lower test reliabilities. Nevertheless, bias in coefficients and attenuation of R^2 were generally modest and the relative importance of the three aptitudes did not change when coefficients were corrected for measurement error.

Hypothesis 4 stated that, when two groups are compared, <u>differences</u> between biased regression coefficients would not be the same as <u>differences</u> between unbiased coefficients if the two groups were measured with differential random error by independent variables. Discrepancies in the sizes of regression coefficient <u>differences</u> reflect the effect of differential random measurement error in aptitude indicators. If random measurement error in indicators used as independent variables is the same for two groups,

differences in biased coefficients will not be themselves biased. If error is not the same, coefficient differences will be biased. Biased differences were expected. The extent of this bias was, however, unknown.

In previous analyses in this research it was found that differences in measurement model parameter estimates produced different test reliabilities between some groups. The next analysis compared differences in biased OLS and unbiased LISREL regression coefficients for groups who were found to have different test reliabilities.

In each two group comparison, biased regression coefficients obtained by least-squares regression were subtracted one from another to obtain a set of differences. For the same two groups, unbiased regression coefficients obtained by LISREL were subtracted one from another to obtain the second set of regression coefficient differences. Measurement model parameter estimates in the regressions of grades on factors were constrained to be equal for the two groups where it had previously been found to be appropriate. For example, since white and Mexican-American males were found to have equal test loadings, these were constrained to be equal in the regressions of grades on factors used to obtain unbiased regression coefficients.

Table 4.9 gives the regression coefficients and differences for all comparisons between groups with different test reliabilities. In

Table 4.9

Differences in OLS and LISREL
Unstandardized Regression Coefficients
for Groups with Different Test Reliabilities

	OLS			LISREL		
Group	Vocab	Math	Mosaics	Vocab	Math	Mosaics
WM	.056*	.093*	.004*	.043*	.129*	-.002
BM	.030	.045*	.007	.014	.075*	.003
Diff.	.026	.048	-.003	.029	.054	-.005
WM	.056*	.093*	.004*	.043*	.129*	-.002
MM	.022	.096*	.0001	-.089	.193*	-.014
Diff.	.034	-.003	.0039	.132	-.065	.012
WF	.065*	.089*	.005*	.062*	.113*	.005
BF	.046*	.067*	.004	-.001	.134*	-.004
Diff.	.019	.022	.001	.061	-.021	.009
WF	.065*	.089*	.005*	.062*	.113*	.005
MF	.042	.057*	.002	.013	.091*	.009
Diff.	.023	.032	.003	.049	.022	-.004
WM	.056*	.093*	.004*	.043*	.129*	-.002
WF	.065*	.089*	.005*	.062*	.113*	.005
Diff.	-.009	.004	-.001	-.019	.016	.003

* Coefficient significantly different from zero at α =.05.

(continued)

Table 4.9 (con't.)

Differences in OLS and LISREL Regression Coefficients for Groups with Different Test Reliabilities

	OLS			LISREL		
Group	Vocab	Math	Mosaics	Vocab	Math	Mosaics
BM	.030	.045*	.007	.013	.074*	.004
BF	.046*	.067*	.004	-.009	.142*	-.005
Diff.	-.016	-.022	.003	.022	-.068	.009
MM	.022	.096*	.0001	-.059	.154*	-.006
MF	.042	.057*	.002	-.020	.120*	.008
Diff.	-.018	.039	-.0019	-.039	.034	-.014

most of the two group comparisons, differences were larger between the unbiased LISREL coefficients than between the biased OLS coefficients. In other words, slope differences were generally underestimated when random measurement error was not accounted for. The seriousness of underestimation can best be seen by examining differences in coefficients for mathematics aptitude. Vocabulary and mosaics coefficients were not significant for one or both groups in most comparisons, but the mathematics coefficient was significant for all groups in all regressions. The most serious case of underestimation of the differences in mathematics coefficients occurred for the comparison between white males and Mexican-American males. The coefficient for Mexican-American males was 3% higher than that of white males in the least-squares regressions and was 50% higher in the factor regressions. A similar degree of underestimation occurred in the comparison between black males and black females, but for all other comparisons underestimation of the difference in mathematics coefficients was negligible or there was a negligible overestimation when equal test reliabilities were assumed.

Hypothesis 4 was therefore supported by the findings. Differences between biased regression coefficients for two groups with differential test reliabilities are themselves biased as a result of differential reliabilities. Overall, the differential reliabilities produced modest to slight biases in regression coefficient differences.

CONCLUSIONS

This research was stimulated by the neglect of social scientists to investigate the effect of measurement error on the quality of their results. Of specific concern was the nature and extent of measurement error in typical aptitude tests for different sociocultural groups. Aptitude tests are essential to educational research (primarily as control variables), yet until now no thorough investigation had been undertaken to confirm suggestions that measurement properties of these tests differ between males and females and among whites, blacks and Mexican-Americans.

If tests are not valid indicators of the same construct for different groups, test results may not be interpreted in the same manner. If they are valid for the same construct, but not equally reliable, the groups will be measured with differential accuracy (standard errors of measurement will be different) and parameter estimates in structural models will be differentially biased. It has been uncommon for researchers to investigate differential measurement error, much less to correct parameter estimates for the known biasing effects of random error.

This study sought to determine whether typical measures of aptitudes are indeed valid for the same constructs across groups, and, if so, whether they measure the construct with equal reliability. Furthermore, to indicate the seriousness of different test reliabilities, the biasing effects of differential reliability upon structural model regression coefficients was assessed.

Data for senior high school students from the High School and Beyond (HSB) national longitudinal study were used in this research to maximize the external validity of the results and to provide users of HSB with information on the specific tests administered. Maximum likelihood confirmatory factor analysis procedures were used to test hypotheses in order to overcome problems with more traditional techniques, and to illustrate their application to common problems in educational research.

The findings indicated, as predicted, that the vocabulary, mathematics and mosaics tests administered to High School and Beyond senior respondents were valid for the same constructs for the six ethnic-sex groups of interest. This finding is consistent with most exploratory factor analysis studies concerned with the same issue. The three aptitudes investigated in this research were clearly differentiated cognitive aptitudes. In contrast to these findings, Atkin et al. (1977) analyzed a variety of ability and information tests, and found that information subtests had different factor structures for males and females and for whites and blacks. In general, they found fewer factors for females than males and fewer for blacks than whites. Factors underlying information tests were found to be a function of differential interests, knowledge and skills. Therefore, the findings from the present research should not be generalized to all tests that may be called aptitude tests, especially those designed to measure knowledge of highly specialized information.

Reliabilities were expected to be a function of scoring differences, with lower scoring groups having lower test reliabilities. Lower reliabilities were expected to result from larger error variances and smaller factor, or true score, variances for the lower scoring groups. Although reliability differences appeared to occur in part for the reasons hypothesized, they were also influenced by differential relationships of tests to constructs, reflected in factor loadings, as well.

For ethnic group comparisons, whites scored higher than blacks and Mexican-Americans on all tests, but had generally higher reliabilities only on vocabulary and mathematics tests. Reliabilities were lower for whites on the mosaics tests. Reliability differences appeared to be a function of true score variances, as expected, but smaller true score variances were not uniquely associated with lower scoring groups. True score variances were smaller for whites for mosaics tests.

Test error variances were, in general, larger for lower scoring ethnic groups, as expected, but not markedly so for most tests. Larger error variances were expected because lower scoring groups have more opportunity for guessing. Tests from High School and Beyond were administered with instructions not to guess as scores would be corrected downward as a penalty. Error variances may not have differed dramatically because students followed instructions.

With regard to ethnic group comparisons, it is concluded that there are differences in aptitude test reliabilities between whites and blacks and between whites and Mexican-Americans, within gender groups. However, especially if guessing has been controlled, aptitude test reliability differences cannot always be assumed to be lower for lower scoring groups.

Similar conclusions are drawn for the results of reliability comparisons between males and females. Differences were found in loadings, error variances and true score variances between the two groups, and these differences produced lower reliabilities for the lower scoring females on vocabulary and mathematics tests. However, on the mosaics tests, on which females scored higher, reliabilities were not consistently higher or lower. Once again, the size of true score variances appeared most influential in producing reliability differences.

IMPLICATIONS

These results provide important new information, both for those whose primary interest is the interpretation of aptitude test results, and for educational researchers who use such tests as control variables. Test developers often provide reliability information for test administrators and researchers. Unfortunately, this information is not usually obtained by ethnic group. The findings of this research indicate that such information may be grossly inaccurate for

studies limited to specific groups. For example, reliability coefficients for some of the tests in the High School and Beyond study vary considerably from one group to another. The Vocab1 test, which is the longer and more likely test to be used as an indicator of verbal aptitude, had a reliability of .714 for white males and .565 for Mexican-American females. The Math2 test, which overall was the least reliable test, had what might be considered an acceptable reliability for white males, but a clearly unacceptable one for most other groups. Such findings stress the need to obtain test reliabilities for specific ethnic groups being assessed or studied. Differences in reliabilities between males and females are not as pronounced as among blacks, whites and Mexican-Americans; however, the findings suggest that reliabilty coefficients for aptitude tests should be obtained separately for males and females as well as ethnic groups, especially if standard errors of measurement or corrections for attenuation are of interest.

The consequence of ignoring random measurement error was assessed by comparing ordinary least-squares (OLS) and LISREL regression coefficients obtained by regressing a dependent variable (self-reported high school grades) on three of the aptitude tests (OLS) and on the aptitude factors (LISREL) for each group separately. Underestimation and overestimation of coefficients and attenuation of R^2 occurred for all groups, but most seriously for blacks and Mexican-Americans. Some coefficients were significant in

the biased OLS case and not significant in the unbiased LISREL case for some groups. The relative magnitudes of the coefficients, however, were not noticeably different in the two types of regressions.

Reliabilities for the vocabulary and math tests used to obtain biased OLS coefficients were relatively high as reliabilities go. For most groups they were greater than .60 and .70 respectively. The reliabilities for the Mosaics1 test were slightly lower at approximately .50. High reliabilities and the weak association of two of the tests (Vocab1 and Mosaics1) with grades lead to generally modest biases in the regression coefficients and R^2, and yielded little change in substantive conclusions about the importance of the tests as predictors. Had the vocabulary and mosaics tests had stronger effects, even a modest degree of bias might have changed the relative importance of the predictors.

It appears that ordinary least-squares regression, in which variables are assumed to be error-free, produces good approximations of unbiased coefficients if reliabilities are greater than .50. If, however, effects of more than one variable are strong, the relative importance of predictors may be affected even by a modest degree of bias. If some or all independent variables have reliabilities less than .50, bias may substantially affect the size of regression coefficients, R^2, and the relative importance of predictors.

Reliabilities for aptitude tests used to obtain regression coefficients in this research were greater than .50 for most ethnic and gender groups. This will not always be the case with other aptitude measures, or with other types of variables. We know that some ethnic and gender groups have differential reliabilities not only on aptitude tests, but on measures of socioeconomic background (Bielby et al., 1977; Corcoran, 1980; Wolfle and Robertshaw, 1982). It is possible that these groups have differential reliabilities on other types of measures as well. Therefore, it is important to obtain reliabilities for all variables of interest for specific ethnic or gender groups before assessing the degree of bias in regression coefficients that might be expected. To assure that the precision and meaning of coefficients are not affected by any amount random measurement error, it is recommended that corrections for the presence of error be undertaken before estimating structural models.

When regression coefficients are compared between groups measured with differential reliability by independent variables, differences in the coefficients will be biased. When the extent of bias was assessed for group comparisons of regresssion coefficients obtained with aptitude constructs as independent variables, it was found that bias was slight to modest for significant coefficients. Although test reliabilities differed among the groups compared, the differences were not great enough to substantially bias regression coefficient differences.

For the tests used as independent variables to obtain biased coefficients, the six groups in this study did not have substantially different test reliabilities. The range for the Vocab1 test was .565 to .714; for the Math1 test, .699 to .862; and for the Mosaics1 test, .430 to .585. Therefore, when independent variable reliabilities differ by less than .20, regression coefficient differences are biased, but not substantially so. In group comparisons in which reliabilities are more discrepant, it should be expected that coefficient differences will be seriously biased. Again, in the interest of precision, it is recommended that differential bias be accounted for when regression coefficients are compared between groups.

In summary, the six aptitude tests used in this study may be considered to be measuring the same trait for different ethnic and gender groups. They do not, however, measure traits with equal reliabilities across groups. This finding has implications for assessing test performances for different groups and for the biasing effects of random measurement error. Regression coefficients for groups measured less reliably will be more seriously biased, and when regression coefficients are compared between groups, the more discrepant the reliabilities between groups, the more biased will be regression coefficient differences.

The aptitude tests included in this analysis were chosen because of their substantive importance as control variables in educational research. For example, Page and Keith's (1981) major criticism of

Coleman et al.'s (1981) study of the cognitive effects of public and private schools was that Coleman failed to include a proper control for ability in his analysis. It has been found here that the aptitude tests in the High School and Beyond Study are fairly reliable indicators of the latent traits they purport to measure. These reliabilities do differ from group to group, primarily as a result of differences in true score and error variances, but the magnitude of these differences is not large. When cross-group comparisons are made of OLS regression coefficients, it has been found here that these comparisons are fairly robust in the presence of modest degrees of measurement error.

Finally, a word of caution. These conclusions about the robustness of OLS regression estimates in the face of differential measurement error are limited to situations in which the extent of measurement error is not severe. These results do not automatically generalize to other variables often used as controls in educational research. While measures of socioeconomic background enjoy reliability estimates that approach the magnitude of those observed here (Wolfle and Robertshaw, 1982a), other control variables do not. Wolfle and Robertshaw (1982b), for example, found reliability estimates for an abreviated version of the locus-of-control instrument to be as low as .13. It is therefore advised that educational researchers undertake cross-ethnic or cross-gender group comparisons of OLS regression estimates only after they have assured themselves that measurement error will not bias their substantive conclusions.

REFERENCES

Alwin, Duane F. and David J. Jackson
1979 "Measurement models for response errors in surveys: Issues and applications." Pp. 68-119 in Karl F. Schuessler (ed.), Sociological Methodology 1980. San Francisco: Jossey-Bass.

Alwin, Duane F. and David J. Jackson
1981 "Applications of simultaneous factor analysis to issues of factorial invariance." Pp. 249-279 in David J. Jackson and Edgar F. Borgatta, Factor Analysis and Measurement in Sociological Research. London: Sage.

Atkin, Robert, R. Bray, M. Davison, S. Herzberger, L. Humphreys and U. Selzer
1977 "Ability factors differentiation, grades 5 through 11," Applied Psychological Measurement 1(19): 65-76.

Bentler, Peter M. and Douglas G. Bonett
1980 "Significance tests and goodness-of-fit in the analysis of covariance structures," Psychological Bulletin 88: 588-606.

Bielby, W. T., R. M. Hauser and D. L. Featherman
1977 "Response errors of black and nonblack males in models of the intergenerational transmission of socioeconomic status," American Journal of Sociology 82(6): 1242-1288.

Blalock, H. M.
1968 "The measurement problem." Pp. 5-27 in H. M. Blalock and A. Blalock (eds.), Methodology in Social Research. New York: McGraw-Hill.

Blalock, H. M.
1970 "Estimating measurement error using multiple indicators and several points in time," American Sociological Review 35: 101-111.

Bohrnstedt G. W. and E. F. Borgatta
1980 "Forward: Special issue on measurement," Sociological Methods and Research 9(2): 139-146.

Bohrnstedt, G. W. and T. W. Carter
1971 "Robustness in regression analysis." Pp. 118-146 in H. L. Costner (ed.), Sociological Methodology: 1971. San Francisco: Jossey-Bass.

Bowles, Samuel
1972 "Schooling and inequality from generation to generation," Journal of Political Economy 80 (May-June): S219-S251.

Bowles, Samuel and Herbert Gintis
1976 Schooling in Capitalistic America. New York: Basic.

Bowles, Samuel and Valerie I. Nelson
1974 "The inheritance of IQ and the intergenerational reproduction of economic inequality," Review of Economics and Statistics 5 (February): 39-51.

Coleman, James, T. Hoffer and S. Kilgore
1981 Public and Private Schools: A Report to the National Center for Education Statistics by the National Opinion Research Center. University of Chicago.

Corcoran, Mary
1980 "Sex differences in measurement error in status attainment models," Sociological Methods and Research 9(2): 199-217.

Cronbach, L. J.
1951 "Coefficient alpha and the internal structure of tests," Psychometrica 16: 297-334.

Cuttance, Peter F.
1982 "Covariance structure modelling of reliability and differential response in educational survey data." Paper presented at the American Educational Research Association annual meeting, New York.

Dean, Raymond S.
1980 "Factor structure of the WISC-R with Anglos and Mexican-Americans," The Journal of School Psychology 18(3): 234-239.

Gutkin, Terry B. and Cecil R. Reynolds
1980 "Factorial similarity of the WISC-R for Anglos and Chicanos referred for psychological services," Journal of School Psychology 18(1): 34-39.

Hennessy, James and Philip Merrifield
1976 "A comparison of the factor structures of mental abilities in four ethnic groups," Journal of Educational Psychology 68: 754-759.

Heyns, Barbara and Thomas L. Hilton
1982 "The cognitive tests for High School and Beyond: An assessment," Sociology of Education 55: 89-102.

Humphreys, L. G.
1962 "The nature and organization of human abilities," 19th Yearbook of the National Council on Measurement in Education: 39-45.

Jencks, Christopher
1972 Inequality: A Reassessment of the Effect of Family and Schooling in America. New York: Basic.

Joreskog, K. G.
1967 "Some contributions to maximum likelihood factor analysis," Psychometrika 32: 443-482.

Joreskog, K. G.
1969 "A general approach to confirmatory maximum likelihood factor analysis," Psychometrika 34: 183-202.

Joreskog, K. G.
1970 "A general method for analysis of covariance structures," Biometrika 57: 239-251.

Joreskog, K. G.
1971a "Statistical analysis of sets of congeneric tests," Psychometrika 36(2): 109-133.

Joreskog, K. G.
1971b "Simultaneous factor analysis in several populations," Psychometrika 36(4): 409-426.

Joreskog, K. G. and Dag Sorbom
1978 LISREL: Analysis of Linear Structural Relationships by the Method of Maximum Likelihood User's Guide. Chicago: National Educational Resources.

Kenny, David A.
1979 Correlation and Causality. New York: John Wiley & Sons.

Levinsohn, Jay R., Louise B. Henderson, John A. Riccobono and R. Paul Moore
1978 National Longitudinal Study: Base Year, First, Second and Third Follow-Up Data File Users Manual, Volumes I and II. Washington, D. C.: National Center for Education Statistics.

Long, J. S.
1976 "Estimation and hypothesis testing in linear models containing measurement error: A review of Joreskog's model for the analysis of covariance structures," Sociological Methods & Research 5: 157-206.

Lord, F. M. and M. R. Novick
1968 Statistical Theories of Mental Test Scores. Reading, Mass.: Addison-Wesley.

Mason, W. M., R. M. Hauser, A. C. Kerckhoff, S. S. Poff and K. Manton
1976 "Models of response error in student reports of parental socioeconomic characteristics." Pp. 443-494 in W. H. Sewell, R. M. Hauser and D. L. Featherman (eds.), Schooling and Achievement in American Society. New York: Academic Press.

McGaw, Barry and K. G. Joreskog
1971 "Factorial invariance of ability measures in groups differing in intelligence and socioeconomic status," British Journal of Mathematical and Statistical Psychology 24: 154-168.

National Opinion Research Center
1980 High School and Beyond Information for Users Base Year (1980) Data. Washington, D. C.: National Center for Educational Statistics.

Nunnally, Jum C.
1978 Psychometric Theory. New York: McGraw-Hill.

Page, Ellis B. and Timothy Z. Keith
1981 "Effects of U.S. private schools: A technical analysis of two recent claims," Educational Researcher 10(7): 7-17.

Patteson, B. J. and Lee M. Wolfle
1981 "Specification bias in causal models with fallible indicators," Multiple Linear Regression Viewpoints 11(1): 75-89.

Reschuly, Daniel J.
1978 "WISC-R factor structures among Anglos, Blacks, Chicanos and Native-American Papagos," Journal of Consulting and Clinical Psychology 46(3): 417-422.

Siegel, Paul M. and Robert W. Hodge
1968 "A causal approach to the study of measurement error." Pp. 28-59 in H. M. Blalock and A. B. Blalock (eds.), Methodology in Social Research. New York: McGraw-Hill.

Treiman, Donald J. and Robert M. Hauser
1976 "Intergenerational transmission of income: An exercise in theory construction." Pp. 271-302 in Robert M. Hauser and David L. Featherman (eds.), The Process of Stratification. New York: Academic Press.

Tucker, L. R. and C. Lewis
1973 "A reliability coefficient for maximum likelihood factor analysis," Psychometrika 38: 1-10.

Walker, H. M. and J. Lev
1953 Statistical Inference. New York: Holt, Rinehart and Winston.

Wolfle, Lee M.
1979 "Unmeasured variables in path analysis," Multiple Linear Regression Viewpoints 9: 20-56.

Wolfle, Lee M.
1982 "Causal models with unmeasured variables: An introduction to LISREL," Multiple Linear Regression Viewpoints 11: 9-54.

Wolfle, Lee M. and Marilyn Lichtman
1981 "Educational attainment among whites, blacks and Mexican-Americans". Paper presented at the American Educational Research Association annual meeting, Los Angeles.

Wolfle, Lee M. and Dianne Robertshaw
1982a "Racial differences in measurement error in educational achievement models." Paper presented at the American Educational Research Association annual meeting, New York.

Wolfle, Lee M. and Dianne Robertshaw
1982b "Effects of college attendance on locus of control," Journal of Personality and Social Psychology (in press).

Zeller, Richard A. and Edward G. Carmines
1980 Measurement in the Social Sciences. London: Cambridge University Press.

APPENDIX A

Correlation Matrices

Table A

Correlation Matrix for All Variables
White Males
N=6463

Grades	Vocab1	Vocab2	Math1	Math2	Mosaics1	Mosaics2
1.00						
.38850	1.00					
.40427	.71277	1.00				
.50838	.54915	.54699	1.00			
.45114	.45913	.46195	.67240	1.00		
.18458	.18751	.18254	.30823	.26729	1.00	
.24013	.22808	.21889	.38337	.32286	.66883	1.00

Table B

Correlation Matrix for All Variables
White Females
N=7026

Grades	Vocab1	Vocab2	Math1	Math2	Mosaics1	Mosaics2
1.00						
.40129	1.00					
.38777	.67768	1.00				
.50981	.52790	.53986	1.00			
.36756	.37761	.39498	.58109	1.00		
.16186	.13238	.12951	.25283	.19098	1.00	
.22627	.18561	.17509	.33551	.24871	.63035	1.00

Table C

Correlation Matrix for All Variables
Black Males
N=647

Grades	Vocab1	Vocab2	Math1	Math2	Mosaics1	Mosaics2
1.00						
.20820	1.00					
.22943	.65223	1.00				
.28349	.51011	.49783	1.00			
.22021	.37157	.37207	.48336	1.00		
.14124	.17556	.19661	.31065	.16931	1.00	
.16861	.19876	.21550	.39951	.19396	.74038	1.00

Table D

Correlation Matrix for All Variables
Black Females
N=786

Grades	Vocab1	Vocab2	Math1	Math2	Mosaics1	Mosaics2
1.00						
.26179	1.00					
.22095	.55929	1.00				
.34733	.47641	.40680	1.00			
.27411	.27182	.32340	.42597	1.00		
.13841	.20124	.15030	.27504	.14300	1.00	
.17895	.23633	.17778	.37600	.19243	.69482	1.00

Table E

Correlation Matrix for All Variables
Mexican-American Males
N=280

Grades	Vocab1	Vocab2	Math1	Math2	Mosaics1	Mosaics2
1.00						
.29573	1.00					
.27259	.61993	1.00				
.48162	.52986	.51433	1.00			
.29715	.38606	.38847	.55270	1.00		
.13871	.18948	.15124	.28132	.24944	1.00	
.19117	.21886	.17163	.33099	.30834	.66529	1.00

Table F

Correlation Matrix for All Variables
Mexican-American Females
N=321

Grades	Vocab1	Vocab2	Math1	Math2	Mosaics1	Mosaics2
1.00						
.23193	1.00					
.16234	.50037	1.00				
.30872	.46431	.40433	1.00			
.26583	.21832	.22830	.40897	1.00		
.08608	.15096	.06790	.18938	-.00796	1.00	
.15760	.18330	.19081	.26264	.08275	.63394	1.00

The vita has been removed from the scanned document

SEX AND ETHNIC DIFFERENCES IN APTITUDE INDICATOR MEASUREMENT MODELS

by

Dianne W. Robertshaw

(ABSTRACT)

Measurement error in construct indicators is known to bias structural model regression coefficients, and differences in regression coefficients when two groups are measured with differential error. The validity and reliability of six aptitude tests administered to high school seniors in the High School and Beyond (HSB) national longitudinal study were investigated for white, black and Mexican-American males and females. The tests were found to be valid measures of the same constructs across groups, but test reliability coefficients were found to differ between males and females, between whites and blacks, and between whites and Mexican-Americans. For unspeeded tests, reliability coefficients were consistently lower for females than males, and lower for blacks and Mexican-Americans than for whites.

The effect of different test reliabilities on structural model regression coefficients, and differences in coefficients when two groups are compared, was assessed. The coefficient of determination (R) was attenuated to a greater extent and regression coefficients were more biased for blacks and Mexican-Americans than for whites. Coefficient differences were modestly biased when two groups with

different test reliabilities were compared.

Maximum likelihood covariance structure analysis procedures were used to estimate, fit and compare measurement and structural models. The computer program used was LISREL IV. Ordinary least-squares regression coefficients were compared to LISREL regression coefficients in assessing the extent of bias in regression coefficients.