Ethnic Group Membership as a Moderator of Job Performance

Brian S. O'Leary
James L. Farr
C. J. Bartlett

TECHNICAL REPORT NUMBER 1

April 1970

Prepared under contract from
Personnel and Training Research Programs
Psychological Sciences Division
Office of Naval Research
Department of the Navy

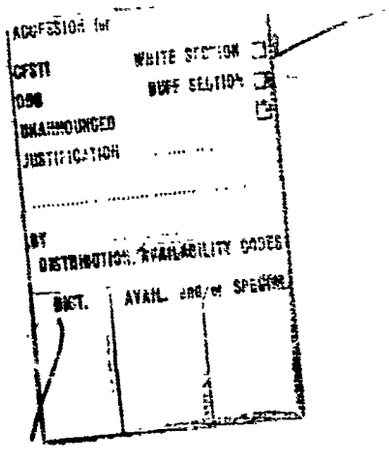
Contract No. N00014-68-C-0341

NR No. 151-277

Principal Investigator: Albert S. Glickman

AMERICAN INSTITUTES FOR RESEARCH
WASHINGTON OFFICE
Address: 8555 Sixteenth Street, Silver Spring, Maryland 20910
Telephone: (301) 587-8201

The contents of this oublication do not necessarily represent the officiai position or policy of the Department of the Navy.



The computer time for this research was supported in part through the facilities of the Computer Science Center of the University of Maryland.

AIR-753-4/70-TR-1

ETHNIC GROUP MEMBERSHIP AS A MODERATOR OF JOB PERFORMANCE

Brian S. O'Leary

James L. Farr

C. J. Bartlett

TECHNICAL REPORT NUMBER I

Prepared under Contract from
Personnel and Training Research Programs
Psychological Sciences Division
Office of Naval Research
Department of the Navy

Principal Investigator: Albert S. Glickman

Contract No. N00014-68-C-0341

NR No. 151-277

American Institutes for Research
Washington Office
Institute for Research in Organizational Behavior

April 1970

ABSTRACT

This report presents the findings of the first phase of a research project to investigate the problems which exist regarding subcultural differences in the prediction of job performance. Phase I of the project was an attempt to obtain an adequate picture of the effects of cultural factors on existing selection procedures. Seven independent studies were conducted in which the validity of commercial and industrially developed selection tests was examined separately for white and Negro subgroups of the population using the eleven different relationships presented in the Bartlett and O'Leary (1969) model. Occupational groups which were studied included toll collectors, correctional officers, toll facility officers, various clerical workers, and keypunch operators. A sample of inmates in a federal correctional institution was also studied.

The results of Phase I indicated that test bias is clearly present in a large number of cases where heterogeneous groups are combined in making predictions of job performance. However, it is erroneous to conclude that all inadvertent test bias denies opportunities to minority group members. The present study has demonstrated the need to validate tests separately for minority and majority group members. The traditional validation model which assumes homogeneous populations is clearly inappropriate.

The second phase of the project will involve the evaluation of procedures to control or eliminate bias. Differential prediction models, culture-equivalent tests, learning measures, as well as some non-cognitive measures will be examined.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Equal opportunity for minority group members in industrial and educational institutions has become an area of national concern. Both professionals and laymen have claimed that many of the current methods of assessing abilities may systematically deny opportunities for minority groups.

Although there is considerable agreement that a problem exists regarding subcultural differences in the prediction of job performance (see APA Task Force on Employment Testing of Minority Groups, 1969), there is a need to learn more about the nature of the problem. Bartlett and O'Leary (1969) have developed a model which demonstrates possible relationships which may exist when heterogeneous groups are combined in making predictions. Viewing this model in terms of subcultural bias, it becomes apparent that there are a number of different situations where inadvertent test discrimination may be found. More important, however, is the realization that solutions to the problem of test bias are dependent upon the nature of the existing relationship between the tests and the criterion.

No single technique, such as culture-free tests or test-taking training, will solve all problems, but each may be useful in certain situations. However, until a basic parametric study is conducted to determine the nature of the problem, haphazard applications of the various techniques which have been suggested as solutions may lead to the elimination of some potentially useful techniques. For example, one may be using test-taking training to eliminate unfair discrimination in situations which call for differential prediction, as in the example where one test has positive validity for one subgroup and negative validity for another.

Guion (1966) has alluded to the need for a basic parametric study, stating that there is no evidence now available to indicate which models will be most useful for eliminating unfair discrimination in testing. The present project was a response to this need.

Phase I of this two part project, essentially exploratory in nature, was an attempt to obtain an adequate picture of the effects of cultural factors on existing selection procedures. More specifically, an attempt was made to

determine the frequency of occurrence of the eleven different relationships presented in the Bartlett and O'Leary (1969) model, as well as how pervasive these relationships are across a number of different types of tests and criteria. Phase II activities, currently in progress, are directed toward the development and experimental evaluation of procedures to control or eliminate test bias.

The present technical report describes the results of Phase I research efforts. Over 30 different organizations were contacted in an effort to obtain test validation data.[1] Data were obtained from approximately 20 percent of those contacted. Many of the organizations contacted did not have enough minority group members in similar job classifications to obtain a separate validation sample. In addition, many agencies were reluctant to release data because of the controversial nature of the topic.

Test validation research for minority groups presents a number of unique methodological problems. First, since often only a few minority group members are employed in a specific job classification, it is virtually impossible to divide the groups for purposes of cross-validation. Secondly, because of the rather large differential in sample size, validity coefficients of equal magnitude are often not statistically significant for the minority sample but significant for the white sample.

The Bartlett-O'Leary model, which was being evaluated in this investigation, assumes that subgroup differences on the criterion measures are a function of actual differences in job performance. Although a few of the studies reported contain objective criteria, the most frequently used criterion was supervisory ratings of job performance. In most of the studies, meetings were held with supervisors to familiarize them with the rating scales and to stress the experimental nature of the ratings. Moreover, racial identification was obtained for each employee after the ratings had been collected. Despite these precautionary steps, no estimate was available concerning the nature and extent of bias affecting these ratings for the two racial groups.

---

[1] Academic and governmental institutions, as well as, industrial organizations were contacted.

Section I: Historical Background

In recent years there has been an increasing awareness of the need for socially responsible behavior on the part of all kinds of organizations. The passage of the Civil Rights Acts of 1964 has made the issue of discrimination in personnel selection a legal as well as a moral one. In particular, doubts have been raised about psychological tests used in personnel selection (Amrine, 1965). These tests have come under attack on many fronts for alleged bias against minority and culturally disadvantaged groups. The purpose of this investigation is to determine if this bias actually does exist by examining the relation between selection tests and job performance in a variety of occupational groups in which both majority and minority group members are employed.

Concepts of Bias

The definition of test bias used in the present study was that of Cleary (1966, p.1): "A test is biased for members of a subgroup of the population if, in the prediction of a criterion for which the test was designed, consistent nonzero errors of prediction are made for members of the subgroup. In other words, the test is biased if too high or too low a criterion score is consistently predicted for members of the subgroup when the common regression line is used."

This definition of test bias has several implications. First, a test, in and of itself, is not discriminatory. The use to which a test is put, however, can be discriminatory (Tenopyr, 1967). Unless an outside criterion is applied, a significant difference in mean test scores for different cultural or ethnic groups cannot be presumed to be bias against one or more of the subgroups. It is certainly not unreasonable to assume that the test is measuring a true difference between subgroups on the test dimension or dimensions. Thus, to label a test as discriminatory solely on the basis of difference in test performance between the different subgroups indicates a misunderstanding or a definition of the concept of test bias that differs from that used in the present investigation.

It should always be remembered that the purpose of a selection test is to differentiate between those job applicants who will be good performers on the job and those who will be poor (Guion, 1966). Only if an outside criterion, a measure of job performance, is applied can one determine whether a given selection test is biased or unbiased with respect to the different subgroups

comprising the applicant popula· n. If differences in the test performance of two groups are associated with group differences in the same direction on a job performance measure, then the test is doing its job; i.e., it is differentiating between good and poor performers on the job. A test in this particular situation is unbiased with respect to different groups within the job applicant population (Arvey, 1967). However, if group test performance differences are not associated with group differences in job performance or are associated with group differences in the opposite direction on the performance criterion, then the test is discriminating in an unfair manner and can properly be labeled is biased.

Aside from the legal aspects of test bias in selection procedures, the existence of such bias will usually result in a selection procedure which over- or under-predicts the job performance of certain subgroup members. Thus, the elimination of test bias is desired because it will increase the practical efficiency of the selection procedure in screening out those job applicants who will not be successful on the job and in accepting those job applicants who will be successful.

Bias Reduction

Several alternatives for the elimination of test bias are possible. First, psychological tests could be eliminated from the selection procedure. However, this alternative would perhaps lead to increased discrimination in the selection procedure because such devices as the interview and application blanks used in place of tests may be even more subject to bias. These are potentially more discriminating in an unfair manner than tests and, with these less sophisticated measures bias may be even more difficult to detect or eliminate. If alternative predictors which can be demonstrated to be superior to tests and free from bias are developed, then tests may be replaced by these measures in the selection procedure.

A second alternative is the development of "culture-free" tests. Krug (1966) states that a truly culture-free test must meet one of two conditions: a) all people of all cultures must have had equal oppprtunity and equal motive to learn all items on the test, or b) all items possess complete novelty for all people of all cultures. It is extremely unlikely that any test will ever be constructed so as to m· t either of these conditions. More promising are several variants of c· re-free tests, specifically culture-fair and culture-equivalent tests. The assumption of a culture-fair test is that there exists

a set of test stimuli which are equally appropriate to at least two cultural groups. In a culture-equivalent test, cultural counterparts of various test items are developed (Krug, 1966).

However, until the various subcultures within the major culture are fully investigated and criteria established as to what denotes cultural "fairness" or "equivalence" for these subcultures, it is doubtful that meaningful contributions to the problem of test bias will be made with this approach (Lockwood, 1966). Guion (1966) stated that culture-free tests might be useful as an indication of the degree of cultural deprivation of an individual. He proposed to do this by comparing test scores on a traditional measure of intelligence and on a culture-free test. The difference between the scores (expressed in standard score units) would be a measure of the cultural deprivation.

Tenopyr (1967) stated that the evidence suggests that the Negro job applicant may be at a greater disadvantage when so-called "culture-fair spatial tests are used in selection than when verbal tests are utilized. Kirkpatrick, Ewen, Barrett and Katzell (1967) found that non-verbal predictors were in general not valid for the prediction of job performance of Negro female clerical workers although they were valid for white female clerical workers. The evidence seems to indicate that, although culture-free tests or their variants may be useful in some situations or as supplementary instruments, they cannot be viewed as a panacea for all problems associated with personnel selection from culturally heterogeneous job applicant populations.

A third, perhaps more promising, approach to the elimination of test bias is to investigate the relationship of the predictor and criterion measures separately for each subgroup, i.e., to use subgroup membership as a moderator variable. The term moderator variable was introduced by Saunders (1956) and the concept has had many labels and many definitions (Banas, 1965). The definition of moderator variable used in the present investigation, as suggested by Banas (1965), is any variable, quantitative or qualitative, which improves the usefulness of a predictor by isolating subgroups of individuals for whom a predictor or set of regression weights are especially appropriate.

Moderator Variables and Validation

The moderator variable approach has been advocated by many investigators in this area. Arvey (1967) has stated that businesses wishing to see that Negroes get the jobs for which they are qualified should undertake sophisticated validation procedures for their existing tests and establish different norm groups and validity coefficients for Negroes and whites. Wallace, Kissinger and Reynolds (1966) have recommended that all tests be validated in the setting where they will be used and validation should be for as many separate groups as possible in preference to one large heterogeneous group. Mitchell, Albright and McMurray (1968), after failing to find either total sample or subgroup validity for the Wonderlic Personnel Test with a supervisory rating as the criterion measure, emphasized the need for subgroup validation research in all job situations.

Guion (1965, 1966) has also advocated the investigation of race as a moderator variable and has suggested that different expectancy tables be developed for Negroes and whites in the job applicant population. Kirkpatrick, et al. (1967), in their conclusions based upon a series of studies of differential selection among applicants from different socio-economic or ethnic backgrounds, stated that tests should be validated separately for each ethnic group and that either different standards of selection or different selection instruments should be used with different ethnic groups in most instances.

The Equal Employment Opportunity Commission (1966) has also stressed the importance of validating a selection test for each minority group in the population. Anastasi (1966), also advocating the use of moderator variables, stated that moderator variables are of particular interest because of the widespread concern regarding the use of tests with various subgroups of the general population, especially culturally disadvantaged subgroups. She believes that the empirical investigation of moderator variables in the interpretation of test scores is a more constructive approach than the evasive procedures of so-called culture-free tests.

Bartlett and O'Leary (1969) have developed a differential prediction model to moderate the effects of heterogeneous groups in personnel selection and classification. Several situations have been described in which subgroup test bias has been or could be found. These situations have been labeled 1) equal validity and unequal means; 2) differential validity; 3) opposite validity;

and 4) no validity in subgroups. Each of these general categories can be further divided into subcategories describing the specific relationship between the predictor and criterion measures for each subgroup. A survey of the literature in the area of personnel selection from a heterogeneous applicant population reveals the need for the use of such a differential prediction model in a selection procedure.

Literature Review

The following literature review has been organized by following the terminology suggested by Bartlett and O'Leary (1969).

1. Equal validity and unequal means. In this situation the predictor test yields equal validity for the subgroups but differential mean performance on the test or criterion exists. This typically results in a lower validity if the subgroups are combined. Conversely, separate prediction for the subgroups would lead to increased validity. An exception to this would be where the predictor and criterion mean differentials are in the same direction; i.e., group X is superior to group Y on both the predictor and criterion measures. In this particular situation the test is not biased since it reflects a real difference in predicted performance. (See Figure 1 in Appendix A for an illustration of this relationship.[1])

Cleary (1966) has reported a study in which equal validity but unequal means on both the predictor and criterion were found. Attempting to predict first year college grade point average at a state supported institution in the Southwest, Cleary found that the non-white group had lower mean scores on both the predictor (Scholastic Aptitude Test) and the criterion (grade point average) but that the separate validities of the white and non-white groups were approximately equal. Combining these two groups for purposes of prediction would probably lead to increased validity due to the increased heterogeneity.

Although the Cleary (1966) study is a case in which validity of prediction could be increased by combining groups, most other situations would result in reduced validity.

[1] Figures 1 through 11 in Appendix A are offered as illustrative models. They are not intended to literally represent the bivariate distributions or correlations cited.

Kirkpatrick, Ewen, Barrett and Katzell (1967), studying white and non-white groups (both from culturally deprived backgrounds) who were participating in a heavy vehicle driver training program for the unemployed in New York City, found a significant difference in favor of the white group on the mean predictor scores, yet no significant difference on the criterion measures. Predictors used were the Gates Reading Survey and the Numerical Ability Test of the Differential Aptitude Test; the criteria were graduation vs. termination in the training program and scores on verbal proficiency tests in the training program. If these two groups were combined, not only would a lower validity result but the non-white group would not be as likely to be selected if a cutting score based on the combined group were used. Since both groups had essentially equal chances of success, test discrimination would result if the groups were combined. However, by including race as a moderator, better prediction of criterion performance would be possible as well as the elimination of racial discrimination. (See Figure 2 in Appendix A.)

Kirkpatrick, et al. (1967) also report such a relationship between predictor and criterion measures with a sample of 493 white and 98 Negro female clerical workers in several insurance companies. In this concurrent validation study, the Negro group performed more poorly than the white group on all but one part of a clerical selection test battery, but no differences existed on either criterion measure, salary and supervisory ratings. The validities obtained were essentially the same for both groups. Although methodological problems prevented any conclusive statements about test bias in this situation, the data suggested that bias in the predictor test battery might exist.

Other situations are possible where equal validities but unequal means would lead to both poorer selection decisions and test bias if the subgroups were combined. First is the case where there is a difference between groups on criterion performance, yet no difference in test performance. (See Figure 3 in Appendix A.) This would result in overestimation of the change of success for one group and underestimation for the other. The existence of differences in mean performance on both the criterion and predictor, but in opposite directions, is another possible situation (See Figure 4 in Appendix A). If the two groups were combined, although positive validities existed for each group separately, an overall negative correlation would result. If personnel decisions were made on the basis of a regression equation for the combined groups, the worst from each group would be selected!

2. Differential validity. A selection test may be valid for one subgroup in an applicant population and not valid for another, or the validities may be of different magnitude or even different direction of relationship.

In a study of female toll collectors, Lopez (1966) found differential validity for the subgroups but no differences in mean performance on either the criterion (absences) or the predictor (Clerical Aptitude Test of the Differential Aptitude Tests). (See Figure 5 in Appendix A for an illustration of this relationship.) Lopez found no validity ($r = +.01$) for the white group, a significant correlation ($r = -.18$, $p < .01$) for the Negro group, and no validity for the combined group ($r = -.03$). With the same sample Lopez (1966) also found both differential validity and differential mean predictor performance (with an interview check list as the predictor) but no significant differences in mean criterion performance (see Figure 7 in Appendix A). Again Lopez reported no validity for the white sample ($r = +.02$), low but significant validity for the Negro group ($r = -.14$, $p < .01$). and no validity for the combined group ($r = -.07$). It should be noted that the correlations reported have been corrected for restriction of range. Whether the uncorrected correlations were significant was not reported.

Cleary (1966), investigating academic prediction, reported significant mean differences favoring the white group on both the predictor (Scholastic Aptitude Test - Mathematics) and the criterion (first year grade point average) but she also found differential validity. Cleary reported a significant validity coefficient ($r = .25$, $p < .05$) for the white group but no significant correlation ($r = .01$, n.s.) for the non-white group. Thus, this predictor would be appropriate for the white group but not for the non-white or the combined group. Although a valid prediction could be made from this test for the combined group, this was possible only because the test identified the lower performing group of non-whites (see Figure 8 in Appendix A).

Kirkpatrick, Ewen, Barrett and Katzell (1967) studied several job situations involving many different selection tests and criteria in an attempt to provide evidence in an industrial setting concerning possible test bias in selection procedures. They found differential validity in a number of different job situations.

With a sample of 102 white and 34 Negro female clerical workers, Kirkpatrick, et al., (1967) reported a validity coefficient of .21 ($p < .05$) for the combined group, using as a predictor the Numerical Test of the Short Employment Test and a merit rating criterion. For the white group the validity coefficient was .25 ($p < .05$), but for the Negro group it was .02. In another study reported by Kirkpatrick, et al. of 137 males in a General Maintenance Training program (31 white, 53 Negro and 53 Spanish), differential validity was also found. Using the Gates Reading Survey as the predictor and proficiency task scores as the criteria they obtained a significant validity coefficient ($r = .29$, $p < .01$) for the combined group, a significant coefficient ($r = .42$, $p < .01$) for the Negro group, yet no validity for either the white group ($r = .02$) or the Spanish group ($r = .07$). The correlations reported between the same predictor (Gates Reading Survey) and a termination criterion were .19 (n.s.) for the combined group, .08 (n.s.) for the white group, .31 ($p < .05$) for the Negro group, and .30 ($p < .05$) for the Spanish group. The mean performance on the Gates was significantly ($p < .01$) lower for both the Negro and Spanish groups than the white group. There were no significant differences on the termination criterion but the Spanish group performed significantly lower than the white group on the proficiency tasks ($p < .01$).

Kirkpatrick, et al. (1967) also reported a study using nursing students as the sample and validating a test battery (Pre-Nursing and Guidance Examination developed by the National League for Nursing) against a criterion consisting of a set of state licensing examinations. There were five examinations: medical nursing, surgical nursing, obstetrical nursing, pediatric nursing, and psychiatric nursing. The criterion examination appeared to be unbiased as no consistent pattern of mean performance scores emerged; i.e., whites were superior on two of the exams, Negroes were superior on one and there were no differences on two of the examinations. Inspection of the correlation matrix of the nine subscores on the PNG test battery and the five state examinations revealed 34 instances where validity existed for the combined and white groups but not for the Negro group; five instances where validity existed only for the white group but not for the combined or Negro groups; and six instances in which validity for all groups existed. This large percentage of cases in which differential validity

was found indicates that this situation is perhaps all too common in selection situations.

Ruda and Albright (1968) found that the correlations between the Wonderlic and a turnover criterion were -.26 (r_{bis}, $p<.01$) for the combined group, -.34 (r_{bis}, $p<.01$) for the white group, and + .10 (r_{bis}, n.s.) for the Negro group. The sample consisted of 147 white and 51 Negro clerical workers. Since there are questions about the appropriateness of testing a biserial correlation for significance, the present authors calculated the point biserial correlations for each of the above relationships and tested them for significance. The total group and white sample correlations were again found significant and the Negro sample correlation was not significant.

In all of the above-mentioned studies, it is apparent that the predictors used were not appropriate for all of the subgroups within the population. This points to the need for the development and use of valid predictors for each of the subgroups within a heterogeneous job applicant population.

3. Opposite validity. Lopez (1966) has reported a case where a test had significant positive validity for one group and significant negative validity for another. There were no significant differences in mean test performance (see Figure 9 in Appendix A). With a sample of toll collectors, Lopez reported a validity coefficient of .19 ($p < .01$) for the white group between the Clerical Aptitude Test of the Differential Aptitude Tests and a criterion of tolls accuracy, yet a corresponding correlation of -.23 ($p<.01$) for the non-white group. Thus, the use of this test for selection purposes with a combined group would have no validity. Only through the use of subgroup analyses could the proper interpretation of test performance be made; i.e., one should hire whites who have a high score but non-whites who have a low score on the test. Lopez (1966) also reported a similar situation where a mental ability test correlated in opposite directions for two racial groups but in this case the white group was superior in test performance (see Figure 10 in Appendix A). There was no significant difference in criterion (tolls accuracy) scores. The correlation for the white group was .16 ($p<.01$), but -.18 ($p<.01$) for the non-white group.

Either differential or non-linear prediction would result in validity in this situation, but with the combined group no linear prediction would be possible. Again it should be noted that the correlations reported by Lopez were corrected for restriction of range and the signficance of the uncorrected correlations is not known.

4. No validity in subgroups. It is possible that a test which is valid for a combined group is not valid for any of the subgroups within the population. This could occur if significant differences exist in the same direction on both the predictor and criterion measures (see Figure 11 in Appendix A). This effectively means that the selection procedure is based upon the use of a variable, for example race or socio-economic class, that is not related to job performance. Since the test is valid for neither Group X or Group Y, it should not be used in any way to influence personnel decisions. The validity of the combined group would be based only upon the fact that the two groups differed in performance. The test in this case is actually only a crude measure of the dimension on which the groups differ; for example, race. Failure to consider through appropriate analyses the validity in the subgroups would result in inadvertent racial discrimination through the personnel testing program.

Kirkpatrick, Ewen, Barrett and Katzell (1967) reported several cases in which no validity in subgroups was found. However, none of the data exactly fits the above model. In particular, in none of the cases reported do the groups differ on both the criterion and predictor variables. All of the reported cases involved a sample of 39 white and 33 Negro clerical workers. Using a vocabulary test as the predictor, a correlation of .25 ($p<.05$) was found for the combined group with a rating of quality of work as the criterion. However, the equivalent correlation for the white group was .25 (n.s.) and .19 (n.s.) for the Negro group. With the same predictor, correlations with a rating of overall performance were .27 ($p<.05$) for the combined group, .24 (n.s.) for the white group and .30 (n.s.) for the Negro group. The vocabulary test correlated with a rating of overall effectiveness .30 ($p<.05$) for the combined group, .28 (n.s.) for the white group and .26 (n.s.) for the Negro group. A significant difference ($p<.01$) in the mean rating of overall effectiveness was

the only significant difference found in any of the predictor or criterion measures. The significance of the combined group correlation appears to be only a function of the sample size. With larger samples, it would be likely that validity in the subgroups as well as in the combined group would be found.

This survey of the literature concerned with the problem of prediction of job success for heterogeneous job applicant populations indicates that discrimination in personnel selection tests has been found in a variety of occupational situations. One can only conclude that the proper consideration of this problem is a necessity for an adequate test validation procedure.

It should not be implied from the preceding literature survey that all personnel tests are biased against or for minority group members. Studies have been reported in which no test discrimination was found (Tenopyr, 1967; see also Kirkpatrick, et al., 1967). A report of the APA Task Force on Employment Testing of Minority Groups (1969) states that no clear trends have been established concerning the existence of bias in predicting job performance and that no firm conclusions are possible. Thus the present investigation is an attempt to provide more evidence as to the degree of pervasiveness of test bias in personnel selection procedures.

Section II: General Method

Seven independent studies are reported which employed similar methodology. This section provides an overview of the research effort to limit the amount of redundancy that would occur if all phases of each study were separately described in detail.

Subjects

As the purpose of this phase of the research project was to investigate existing predictor-criterion relationships in job situations, the subjects in all studies were current on-the-job employees or members of existing situational groups in the case of correctional institution inmates. Thus, the samples all consisted of pre-selected groups of individuals. The sample consisted of those persons who had been members of the group under study for at least three months. To assure as large a sample size as possible, a maximum tenure length was not used as a restrictive criterion for inclusion in the sample. i.e., no attempt was made to develop a relatively homongeneous sample with respect to tenure by setting a maximum length-of-service cutting point. The effects of tenure upon the predictor-criterion relationships were statistically controlled when deemed necessary.

Predictors

All predictors were psychological tests which were a part of the existing selection procedure. Most of the tests were used as explicit selection devices though some had been included only for experimental purposes. All of the actual test administration was conducted by the personnel of the organization furnishing the data. In most instances, the subjects in a given sample were not tested at the same time and by the same administrators due to tenure differences.

Criteria

A number of criterion measures were used in each study. Most criteria were already existing measures of job performance but in some cases the measures were developed by the investigators. In all studies an attempt was made to have criteria which measured a wide sample of job performance behaviors. This was limited in certain situations by the record systems of the organizations and other practical considerations.

Statistical Analyses

Means and standard deviations of all predictors and criterion variables were computed for the total sample, the white subgroup and the Negro subgroup. The significance of the difference between the mean predictor performance of the two subgroups was tested by means of the t test. Similar tests were computed for the mean criterion performance of the two subgroups. It should be noted that the distributions of some variables are rather skewed. A basic assumption of the t test is normality of the underlying distribution of the populations. However, Boneau (1960) has shown that the t test is relatively insensitive to violations of its assumptions. Hays (1963) states that the assumption of normality may be violated "almost with impunity provided that sample size is not extremely small," (p.322). A more serious problem is the interaction of the effects of unequal sample sizes and heterogeneity of the two sample variances. If an F test of the ratio of the sample variances revealed heterogeneity, the correction suggested by Welch (1947) was applied.

The validity of each predictor for each criterion measure was estimated by computing zero-order correlations for all possible predictor-criterion pairs for each sample. Validity coefficients were computed for the total sample, the white subgroup and the Negro subgroup. In those samples in which more than one predictor had been used, multiple correlations were not computed because of the instability of such statistics with samples of the relatively small size (in relation to the number of predictors) that existed in the present investigation. Furthermore, the subgroup sizes, especially of the Negro subgroup, were not large enough to permit the use of cross-validation procedures.

Comparisons of each predictor-criterion relationship for the white and Negro subgroups were made by three methods of analysis. First, the significance of the validity coefficients for both subgroups was examined. Tests of the significance of the difference between the two subgroup validity coefficients were computed. Also, the regression tests of the analysis of covariance (Potthoff, 1966) were computed to test the equality of the regression slopes and intercepts for the two subgroups for each predictor-criterion pair. This procedure results in three separate F ratios. F_1 simultaneously tests

the hypothesis that both the regression slopes and the intercepts are equal for the two groups. If F_1 is significant one may conclude that bias exists. F_2 tests the hypothesis that the regression slopes are equal for the two groups. F_3 tests the hypothesis that a common intercept is appropriate for the two groups. F_3 is an appropriate test only when F_2 is not significant.

These three methods of analysis actually constitute two different approaches to the comparison of the validity of a test in two different ethnic subgroups (Kirkpatrick, Ewen, Barrett, and Katzell, 1968). The first approach involves testing the null hypothesis that the validity coefficient for a given test and criterion is equal to .00 (for one or both of the subgroups). Three possible results exist with this approach. The test may be found to be valid for neither, both, or one of the subgroups. If the test is found to be valid for neither subgroup, nothing can really be said about differences in validity since the test is inappropriate in this situation. If the test is found to be valid for both of the subgroups, then it can be appropriately used with both subgroups to predict job performances. If the test is found to be valid for one subgroup but not for the other, there exists a difference in utility in that one may have more confidence that the test is validly useful in one ethnic subgroup than in the other.

The alternate approach to the comparison of the validity of a test in two different ethnic subgroups is to test the significance of the difference between the validity coefficients of the two subgroups. This approach tests the hypothesis that the two subgroups are drawn from the same population with respect to the degree of validity. Rejection of the null hypothesis would denote differential validity, while failure to reject would denote uniform validity for the two subgroups. It is possible that the second approach may fail to show a difference at a given level of confidence while the first does. This can occur because of differences between the two approaches with respect to both degrees of freedom and the sampling error associated with the test of significance. Kirkpatrick, et al. (1968) have indicated that the useful conclusion in this situation is one of a difference in significant validity, in that one might use the test with some confidence to select members of one ethnic subgroup but not of the other.

It is also possible for the first approach to show no validity in either subgroup but the second to show a significant difference between the validity coefficients. This can occur if one of the coefficients is positive and the other is negative. Again, the practical interpretation is to use the test with neither subgroup. In this series of studies, both methods of comparing validity in different ethnic subgroups have been employed and reported, but primary attention has been paid to the outcomes of the first because of its practical implications. The analysis of covariance for homogeneity of regression essentially may be categorized as utilizing the second approach but was also employed as a further means of analysis because of its ability to detect regression intercept differences.

Model Identification

Predictor-criterion relationships were analyzed using the Bartlett and O'Leary differential prediction model in an attempt to determine the relative frequency of the different models.

In accord with the above mentioned methods of analysis, two separate methods of model identification were utilized in those situations where differential validity was demonstrated for the two racial groups (Models 5-10). All predictor-criterion relationships in which a validity coefficient was significant for one racial group, but not significant for the other were identified as illustrations of models when the first method of model identification was used. Because of the rather large difference in sample size between the two racial groups, this procedure identified as models those relationships in which the absolute magnitude of the nonsignificant correlation for Negro sample was larger than the corresponding signficant correlation for the white sample. These cases have been identified as illustrations of models since it is difficult to justify the use of the test for the Negro sample. However, there is some justification in using the test for the white sample even though the absolute magnitude of the validity coefficient is smaller than for the Negro sample.

The second method used to identify illustrations of models imposed the additional criteria of a statistically significant difference between the

validity coefficients for the two racial groups. This method tends to identify clear illustrations of the various models. In each study reported a distinction is made between the models which meet only the first criterion and those models which meet both criteria.

Section III: Studies of Existing Selection Procedures

Study 1: Toll Collectors

Sample

The subjects were 159 female toll collectors (115 white and 44 Negro) employed at the five toll facilities of a state highway department. All employees held state civil service classified positions. The major duties of these toll collectors are to determine the appropriate toll category for each vehicle; to collect cash or toll tickets in the appropriate amount from each vehicle; and to make change when necessary. Table 1 presents biographical information on these employees.

Table 1: Biographical Data - Toll Collectors

	Group	$\bar{X}$	s	N(1)	t(2)
Age	Total	33.86	10.32	152	
	White	34.48	11.18	108	
	Negro	32.32	7.70	44	1.35
Education	Total	11.68	1.02	152	
(in years)	White	11.56	1.03	108	
	Negro	11.98	0.95	44	2.32*
Tenure	Total	34.90	38.26	156	
(in months)	White	36.69	41.16	112	
	Negro	30.34	29.54	44	1.06

(1) Total N is less than 159 because of incomplete data for some subjects.

(2) t ratios are between the means of the white and Negro groups.

* $p < .05$

It can be seen from the data in Table 1 that the white and Negro groups differed significantly only in education, the Negro group having attained a higher educational level.

Predictor Comparisons

Two tests, both developed by the state personnel department, have been used as selection devices for the position of toll collector. Specifically, these tests were a Clerical Checking Test and an Arithmetic Reasoning Test. Because of the recent application of these tests, the number of subjects for whom data was available was considerably diminished. Table 2 presents the predictor means, standard deviations and tests of significance of mean differences for the white and Negro samples.

Table 2: Predictors - Means, Standard Deviations, N's and Tests of Significance of Mean Differences - Toll Collectors

Predictor	Group	$\bar{X}$	s	N	t(1)
Clerical	Total	75.36	4.29	128	
Checking	White	75.71	4.45	89	
	Negro	74.56	3.84	39	1.39
Arithmetic	Total	94.03	5.08	143	
Reasoning	White	94.88	4.54	101	
	Negro	91.99	5.74	42	3.18**

(1) t ratios are between mean test performance for the white and Negro groups.

** $p < .01$

The white and Negro groups did not differ significantly in performance on the Clerical Checking Test. However, the white group scored significantly higher on the Arithmetic Reasoning Test than the Negro sample. The intercorrelations of the two tests were .04 for the total sample, .09 for the whites and -.17 for the Negro sample.

Criterion Comparisons

Several measures of job performance were utilized in this study. Attendance records for three months were obtained from the records of the state highway department. This attendance data was treated in two ways. First, the raw number of days absent from the job was used in the analyses. Also the number of periods of absence was used, e.g., three consecutive days absent counted as one period of absence, but three nonconsecutive days absent counted as three periods of absence.

Extension of the required probationary period and job termination were also used as criteria. Every state civil service employee has a mandatory six month probationary period during which he may be dismissed for almost any reason his supervisor deems sufficient. This probationary period may be extended for one more six month period if the supervisor desires more time to decide if the employee should be permanently hired. Only one such extension is allowed. This criterion was dichotomously scored, a "0" representing extension of the probationary period and a "1" representing no extension of the probationary period. The termination criterion was also dichotomously scored, a "0" representing termination and a "1" representing an employee still employed.

Two objective criterion measures were obtained for this sample, dollar accuracy and axle accuracy. Dollar accuracy for a given toll collector was measured in terms of the ratio of the total number of transactions in a

Table 3: Criteria - Means, Standard Deviations, N's, and Tests of Significance of Mean Differences - Toll Collectors

Criterion	Group	$\bar{X}$	s	N	t(1)
Attendance -	Total	3.27	5.43	153	
Days	White	3.31	5.58	111	
Absent	Negro	3.16	5.08	42	0.14
Attendance -	Total	1.61	1.96	153	
Periods Absent	White	1.51	1.88	111	
in 3 months	Negro	1.85	2.17	42	0.97
Termination	Total	0.85	0.35	157	
	White	0.87	0.34	114	
	Negro	0.81	0.39	43	0.94
Extension of	Total	0.82	0.39	147	
Probation	White	0.81	0.39	106	
	Negro	0.83	0.38	41	0.28
Dollar	Total	150.40	22.71	129	
Accuracy	White	151.85	22.03	94	
	Negro	146.51	24.35	35	1.18
Axle	Total	150.23	23.07	129	
Accuracy	White	150.73	22.58	94	
	Negro	148.90	24.64	35	0.40

(1) t ratios are between means of white and Negro samples.

month that the toll collector completed to the amount of error (in dollars) in the toll receipts turned in during that month. Axle accuracy was measured by the ratio of the total number of transactions in a month to the number of errors in axle count in that month. The toll collector must count the number of axles to determine the proper toll category for trucks; the number of axles is also automatically recorded by a treadle-type counter for each toll booth. Because toll collectors from several facilities were included in the sample, the accuracy measures were converted to T-scores with a mean of 50 and a standard deviation of 10 before being grouped for the analyses. The T-score for each collector was based on the distribution of the accuracy measures for her facility only. This data transformation was made to help control for extraneous situational variance in these measures. The accuracy data for three months were used; the T-scores for a subject for the three months were summed to provide a single measure of each accuracy criterion.

The criteria means, standard deviations and tests of significance of mean differences for the white and Negro samples are presented in Table 3. There were no significant differences between the Negro and white samples on any criterion measure.

Validity

The correlations between the predictors and criteria for the total toll collector sample, the white subgroup and the Negro subgroup are shown in Table 4. If a predictor-criterion relationship fits one of the models proposed by Bartlett and O'Leary (1969), a number indicating the appropriate reference figure in Appendix A is enclosed in parentheses beneath the Negro subgroup correlation. The most striking fact evident from Table 4 is the general lack of validity of either test.

Table 4: Predictor - Criterion Correlations

Toll Collectors (1, 2)

Criterion	Group	Predictor			
		Clerical Checking Test		Arithmetic Reasoning Test	
		r	N	r	N
Attendance -	Total	-04	122	-02	137
Days Abs.	White	-03	85	09	97
	Negro	-03	37	-21	40
Attendance -	Total	-05	122	-11	137
Periods Abs.	White	00	85	09[a]	97
	Negro	-10	37	-33* (7)	40
Termination	Total	06	127	-10	142
	White	05	89	-19	101
	Negro	04	38	-03	41
Extension of	Total	-04	122	02	137
Probation	White	-06	86	09	98
	Negro	01	36	-11	39
Dollar	Total	-15	101	-05	116
Accuracy	White	-25*	71	-03	83
	Negro	04 (5)	30	-17	33
Axle	Total	-05	101	-07	116
Accuracy	White	-10	71	-04	83
	Negro	06	30	-15	33

(1) Decimals are omitted.

(2) Number in parentheses below the correlation for the Negro sample indicates the model illustrated (See Appendix A).

* $p < .05$

a Different from the Negro group correlation at the .05 level.

Models Illustrated

The relationship between the Arithmetic Reasoning Test and the attendance criterion measured in periods of absence illustrates Model 7 (Figure 7 in Appendix A) of the Bartlett and O'Leary (1969) schema. Although no significant differences on the criterion measure were found, the white sample scored significantly higher than the Negro sample on the test. The test was valid only for the Negro sample ($r = -.33$, $p < .05$); not for the white sample ($r = .09$) or the total group ($r = -.11$). Thus, this test is not appropriate for the prediction of this attentance criterion for the total group or the white sample but it would be useful with the Negro sample.

Model 5 (Figure 5 in Appendix A) is illustrated by the relationship of the Clerical Checking Test and the criterion of dollar accuracy. No significant differences on either the predictor or the criterion were found. However, validity was found only for the white sample ($r = -.25$, $p < .05$). Hence, this test is not appropriate for the prediction of this accuracy criterion for either the total group or the Negro sample. The test could appropriately be used to predict performance on this measure for the white sample.

If the more stringent criterion of a significant difference between the subgroup correlations is imposed, only the relationship between the Arithmetic Reasoning Test and the attendance criterion (periods of absence) is illustrative of a model (Model 7, in particular). This result was also found by the analyses of covariance for homogeneity of regression (Potthoff, 1966). Table 5 presents the results of this method of analysis. The significant F_2 statistic for the Arithmetic Reasoning - Attendance (Periods of Absence) relationship indicated that a common regression line cannot be used to predict both white and Negro subgroup performance. No significant F-ratios were found for any other predictor-criterion pair.

Table 5: Analysis of Covariance for Homogeneity of Regression - Toll Collector Sample

Criterion		Predictor					
		Clerical Checking			Arithmetic Reasoning		
		$F_1^{(1)}$	$F_2^{(2)}$	$F_3^{(3)}$	F_1	F_2	F_3
Attendance		.05	.00	.11	1.48	2.80	.16
Days Abs.	df	(2,118)	(1,118)	(1,119)	(2,133)	(1,133)	(1,134)
Attendance		.88	.42	1,35	3.83*	5.89*	1.72
Periods Abs.	df	(2,118)	(1,118)	(1,119)	(2,133)	(1,333)	(1,134)
Termination		.20	.00	.41	1.21	.95	1.48
	df	(2,123)	(1,123)	(1,124)	(2,318)	(1,138)	(1,139)
Extension of		.07	.10	.04	.59	1.12	.06
Probation	df	(2,118)	(1,118)	(1,119)	(2,133)	(1,133)	(1,134)
Dollar		1.71	1.46	1.94	1.15	.39	1.93
Accuracy	df	(2,97)	(1,97)	(1,98)	(2,112)	(1,112)	(1,113)
Axle		.44	.45	.44	.33	.23	.43
Accuracy	df	(2,97)	(1,97)	(1,98)	(2,112)	(1,112)	(1,113)

* $p < .05$

(1) F_1 tests hypothesis that $E(Y_{ij}|X_{ij}) = a - bX_{ij}$ for all i groups.

(2) F_2 tests hypothesis that $E(Y_{ij}|X_{ij}) = a_i - bX_{ij}$ for all i groups.

(3) F_3 tests hypothesis that $E(Y_{ij}|X_{ij}) = a + b_i X_{ij}$ for all i groups. (valid test only if F_2 is not significant.)

It should be stressed that the identification of models is for illustrative purposes only and extreme caution should be exercised in the interpretation of the relationships reported. The number of significant correlations (2 of a possible 36) was only slightly greater than expected by chance at the .05 level.

Study 2: Correctional Officers

Sample

The subjects consisted of 371 correctional officers (322 white and 49 Negro) at two state prisons. The major duties of the officers are to maintain the security of the institution and to supervise the work activities of the inmates. Biographical information for the officers is presented in Table 6.

Table 6: Biographical Data - Correctional Officers

	Group	$\bar{X}$	s	N(1)	t(2)
Age	Total	37.38	10.29	358	
	White	38.41	10.35	311	
	Negro	30.55	6.64	47	6.88**
Education (in years)	Total	10.68	1.74	358	
	White	10.52	1.72	311	
	Negro	11.72	1.51	47	4.51**
Tenure (in months)	Total	58.56	57.24	355	
	White	62.07	59.35	308	
	Negro	35.57	32.92	47	4.48**

(1) Total N is less than 371 due to incomplete data on some subjects.

(2) t ratios are between the means of the white and Negro samples.

** $p < .01$

There were significant differences between the white and Negro samples on all variables, the Negro officers being younger, having more years of formal education and having been on the job for a shorter period of time.

Predictor Comparisons

The California Test of Mental Maturity (CTMM) is the sole predictor used by the state personnel department to select correctional officers. The means, standard deviations, and the test of significance of mean differences are given in Table 7. As can be seen in Table 7, the white sample scored significantly higher on the CTMM than the Negro sample.

Table 7: CTMM - Means, Standard Deviations, N's and Test of Significance of Mean Differences

Correctional Officers

	Group	$\bar{X}$	s	N	$t^{(1)}$
CTMM	Total	78.93	6.14	248	
	White	79.33	6.14	207	
	Negro	76.91	5.83	41	2.32*

(1) t ratio is between the means of the white and Negro sample.

* $p < .05$

Criterion Comparisons

The criteria used with the correctional officer study were attendance (days absent), extension of probationary period, promotion and supervisory ratings. The attendance (days absent only) and extension of probationary period criteria were identical to those described in the toll collector study.

The promotion criterion was controlled for tenure by partial correlation techniques. This criterion measure was dichotomously scored, a "0" representing no promotion and a "1" representing a within-job-classification promotion; i.e., an increase in grade from level one to level two of the job classification.

Table 8: **Criteria** - Means, Standard Deviations, N's and Tests of Significance of Mean Differences

Correctional Officers

Criterion	Group	$\bar{X}$	s	N	$t^{(1)}$
Attendance -	Total	1.89	5.51	371	
Days	White	1.68	5.41	322	
Absent	Negro	3.27	6.03	49	1.88
Extension of	Total	0.76	0.43	355	
Probation	White	0.81	0.39	308	
	Negro	0.43	0.50	47	4.93**
Promotion	Total	1.58	0.49	368	
	White	1.60	0.49	319	
	Negro	1.45	0.50	49	1.98*
Rating by	Total	3.43	0.44	371	
Supervisor	White	3.45	0.44	322	
	Negro	3.31	0.41	49	2.09*

(1) t ratios are between the means of the white and Negro samples.

* $p < .05$

** $p < .01$

Supervisory ratings were also obtained for the correctional officer sample. The rating scale used was developed by the investigators. Recent, detailed job descriptions were available in the state personnel department. Specific job duty statements were written for the correctional officer job classification on the basis of the job descriptions. The distribution of the rating scales to the supervisors was handled by a member of the personnel department of the state correctional department.

The supervisor rated both the importance of the job duty to overall job performance (on a 4-point scale) and the performance of each of his subordinates on each job duty (on a 5-point scale). The final rating for an employee was obtained by summing the performance ratings on those duties rated as important and then dividing by the number of items rated important.

The means, standard deviations, and tests of significance of mean differences for the criteria are presented in Table 8. It can be seen that the Negro sample scored significantly lower than the white sample on three of the criterion measures, extension of probation, promotion and supervisory rating. No significant differences were found on the attendance criterion.

Validity

The correlations between the CTMM and the various criteria for the total group, white sample and Negro sample are presented in Table 9. A perusal of Table 9 again shows a general lack of validity of the test for any of the criterion measures.

The only significant correlation for the correctional officer study was between the CTMM and the attendance criterion for the Negro sample ($r = .33$, $p < .05$).

Table 9: Predictor - Criterion Correlations

Correctional Officers (1, 2)

Criterion	Group	C.T.M.M. r	C.T.M.M. N
Attendance -	Total	03	248
Days	White	-02[a]	207
Absent	Negro	33*	41
		(7)	
Extension of	Total	-03	248
Probation	White	-11	207
	Negro	01	41
Promotion	Total	-08	248
(Controlled	White	-12	207
for Tenure)	Negro	-02	41
Rating by	Total	08	248
Supervisor	White	08	207
	Negro	-01	41

(1) Decimals are omitted.

(2) Number in parentheses below the correlation for the Negro sample indicates the model illustrated (See Appendix A).

* $p < .05$

[a] Different from the Negro subgroup correlation at the .05 level.

Models Illustrated

The relationship between the CTMM and the attendance criterion for the correctional officer model fits Model 7 of the Bartlett and O'Leary (1969) schema (Figure 7 in Appendix A). There was a significant difference on the predictor between the white and Negro samples but no difference on the criterion measure. The test was a valid predictor for the Negro subgroup but lacked validity for both the total sample and the white subgroup. The correlation between the CTMM and the attendance criterion for the Negro subgroup ($r = .33$) was significantly different from that for the white subgroup ($r = -.02$) at the .05 level ($z = 2.05$). Thus, this predictor-criterion relationship is also illustrative of Model 7 when the additional criterion of a significant difference between the subgroup validity coefficients is imposed.

The results of the analyses of covariance for homogeneity of regression for the correctional officer sample are presented in Table 10. The CTMM was found to be biased for the prediction of job performance as measured by the attendance criterion if the total group regression equation were used. The significant F_2 statistic revealed that common beta weight could not be used with both subgroups. This was consistent with the results of the comparison of the validity estimates for the two subgroups.

The analysis of covariance for homogeneity of regression also revealed that the CTMM was biased for the prediction of the extension of probation criterion. Although the CTMM had no validity for the prediction of this criterion ($r = -.03$ for total group; $r = -.11$ for white subgroups; $r = .01$ for Negro subgroup), a common regression equation would underestimate the job performance of the white subgroup but overestimate the performance of the Negro subgroup because the white subgroup scores significantly higher

Table 10: Analysis of Covariance for Homogeneity of Regression - Correctional Officer Sample

CTMM

Criterion	$F_1^{(1)}$	df_1	$F_2^{(2)}$	df_2	$F_3^{(3)}$	df_3
Attendance - Days Abs.	10.47**	(2,244)	9.90**	(1,244)	10.66	(1,245)
Extension of Probation	20.03**	(2,244)	.39	(1,244)	39.76**	(1,245)
Promotion	.53	(2,244)	.70	(1,244)	.35	(1,245)
Rating by supervisor	2.06	(2,244)	.25	(1,244)	3.88	(1,245)

** $p < .01$

(1) F_1 tests hypothesis that $E(Y_{ij}|X_{ij})=a+bX_{ij}$ for all i groups.

(2) F_2 tests hypothesis that $E(Y_{ij}|X_{ij})=a_i+bX_{ij}$ for all i groups.

(3) F_3 tests hypothesis that $E(Y_{ij}|X_{ij})=a+b_iX_{ij}$ for all i groups, (valid test only if F_2 is not significant).

than the Negro subgroup on both the predictor and criterion measures. The significant F_3 statistic revealed that a common intercept value could not be used for the prediction of the extension of probation criterion measure for the two subgroups.

Study 3: Toll Facility Officers

Sample

The subjects in this investigation consisted of 74 toll facility officers employed by a state highway department. The sample included 56 white officers and 18 Negro officers. The major duties of these toll facility officers are maintaining proper traffic flow and enforcing traffic regulations within the toll facility. Table 11 presents biographical information on these employees. The only significant difference found between the white and Negro samples was that the Negro officers had attained a higher educational level than the white officers.

Table 11: Biographical Data - Toll Facility Officers

	Group	$\bar{X}$	s	N(1)	t(2)
Age	Total	34.49	7.38	72	
	White	34.31	7.21	55	
	Negro	35.06	6.33	17	0.36
Education	Total	11.08	1.62	72	
(in years)	White	10.84	1.58	55	
	Negro	11.88	1.50	17	2.37*
Tenure	Total	75.80	49.86	71	
(in months)	White	72.33	50.06	54	
	Negro	86.82	49.08	17	1.03

(1) Total N may be less than 74 because of incomplete data for some subjects.

(2) t ratios are between the means of the white and Negro groups.

* $p < .05$

Predictor Comparisons

Two tests are currently given to toll facility officer job applicants. These are the Otis Quick Scoring and a verbal reasoning test developed by the state personnel department. The verbal reasoning test (called Booklet hereafter) has been recently added; therefore, not much data is available with respect to its validity. Table 12 presents the means, standard deviations and tests of significance of mean differences for the predictors. The white officers scored significantly higher on the Otis than the Negro officers. The intercorrelations of two tests were .69 or the total group, .77 for the white sample and .09 for the Negro sample.

Table 12: Predictors - Means, Standard Deviations, N's and Tests of Significance of Mean Differences

Toll Facility Officers

Predictor	Group	$\bar{X}$	s	N	$t^{(1)}$
Otis	Total	78.21	7.58	71	
	White	79.33	7.31	54	
	Negro	74.67	7.56	17	2.24*
Booklet	Total	78.23	4.84	23	
	White	78.02	5.11	19	
	Negro	79.25	3.66	4	0.44

* $p < .05$

(1) t-ratios are between the means of the Negro and white samples.

Criterion Comparisons

The criterion measures used with the toll facility officers were attendance (days absent and periods of absence), extension of probationary period, promotion and supervisory ratings.

Table 13: Criteria - Means, Standard Deviations, N's and Tests of Significance of Mean Differences

Toll Facility Officers

Criterion	Group	$\bar{X}$	s	N	$t^{(1)}$
Extension of	Total	0.22	0.41	65	
Probation	White	0.22	0.42	50	
	Negro	0.20	0.41	15	0.16
Promotion	Total	1.69	0.46	72	
	White	1.67	0.47	55	
	Negro	1.76	0.44	17	0.69
Attendance -	Total	8.06	10.99	67	
Days Abs.	White	7.86	11.88	51	
	Negro	8.69	7.78	16	0.26
Attendance -	Total	3.04	2.61	67	
Periods Abs.	White	2.73	2.48	51	
	Negro	4.06	2.84	16	1.78
Rating by	Total	3.01	0.39	74	
Supervisor	White	3.03	0.44	56	
	Negro	2.97	0.16	18	0.85

(1) t ratios are between the means of the Negro and white samples.

The two attendance measures and the extension of probationary period criterion were defined and scored in this study in the same manner as described in Study 1 - Toll Collectors. The promotion criterion and the supervisory ratings of job performance were defined and scored in the same manner as described in Study 2 - Correctional Officers.

Table 13 presents the means, standard deviations and tests of significance of mean differences for the criterion measures. There were no significant differences between the white and Negro subgroups on any criterion measure.

Validity

The correlations between the predictors and criteria for the total toll facility officer sample, the white sample and the Negro sample are shown in Table 14. If a predictor-criterion relationship fits one of the models proposed by Bartlett and O'Leary (1969), a number indicating the appropriate reference figure in Appendix A is enclosed in parentheses beneath the Negro group correlation.

The validity of the "Booklet" test was difficult to ascertain because of the small sample to which this test had been given. The Otis Test, in general, exhibited low validity for the criterion measures. The only significant correlation for this test was that between the Otis Test and the criterion of extension of probationary period for the white subgroup only.

Models Illustrated

The relationship between the Otis Test and the extension of probationary period criterion illustrates Model 7 of the Bartlett and O'Leary schema (Figure 7 in Appendix A). There was a significant difference on the predictor between the Negro and white subgroups but no difference on the criterion.

Table 14: Predictor - Criterion Correlations

Toll Facility Officers (1, 2)

Criterion	Group	Predictor			
		Otis		Booklet	
		r	N	r	N
Extension of Probation	Total	18	65	61*	17
	White	30*	50	63*	15
	Negro	-22 (7)	15	--(3)	2
Promotion	Total	-03	71	-34	23
	White	-01	54	-34	19
	Negro	-02	17	--(3)	4
Attendance - Days Abs.	Total	-07	65	-01	17
	White	-03	50	01	15
	Negro	-26	15	-100	2
Attendance - Periods Abs.	Total	-03	65	01	17
	White	14	50	02	15
	Negro	-25	15	-100	2
Rating by Supervisor	Total	05	71	-30	23
	White	02	54	-44	19
	Negro	12	17	43	4

(1) Decimals are omitted.

(2) Number in parentheses below the correlation for the Negro sample indicates the model illustrated (See Appendix A).

(3) Nondeterminant correlation due to zero variance in one variable.

* $p < .05$

The test was a valid predictor for the white subgroup, but lacked validity for both the total group and the Negro subgroup. The correlation between the Otis and the extension of probationary period criterion for the white subgroup was not significantly different from that for the Negro subgroup. Thus, this predictor-criterion relationship is not illustrative of Model 7 when the additional restraint of a significant difference between the subgroup validity coefficients is imposed.

The results of the analyses of covariance for homogeneity of regression for the toll facility officer sample are presented in Table 15. The findings were consistent with the results of the comparison of the validity estimates for the two subgroups. No significant F-ratios were obtained for any of the predictor-criterion pairs.

Table 15: Analysis of Covariance for Homogeneity of Regression - Toll Facility Officer Sample (1)

Criterion	Otis Test					
	$F_1^{(2)}$	df_1	$F_2^{(3)}$	df_2	$F_3^{(4)}$	df_3
Extension of probation	1.64	(2,61)	3.22	(1,61)	.06	(1,62)
Promotion	.24	(2,67)	.00	(1,67)	.49	(1,68)
Attendance--Days absent	.12	(2,61)	.23	(1,61)	.02	(1,62)
Attendance--periods absent	2.56	(2,61)	1.96	(1,61)	3.12	(1,62)
Rating by supervisor	.12	(2,67)	.00	(1,67)	.24	(1,68)

(1) The analysis of covariance for homogeneity of regression was not conducted using the Booklet Test as the predictor variable due to the extremely small sample sizes.

(2) F_1 tests hypothesis that $E(Y_{ij}|X_{ij})=a+bX_{ij}$ for all i groups.

(3) F_2 tests hypothesis that $E(Y_{ij}|X_{ij})=a_i+bX_{ij}$ for all i groups.

(4) F_3 tests hypothesis that $E(Y_{ij}|X_{ij})=a+b_iX_{ij}$ for all i groups, (valid test only if F_2 is <u>not</u> significant.

Study 4: Federal Correctional Institution - Inmate Population

Sample

Study 4 consisted of 155 inmates of a Federal Correctional Institution. Education files of all inmates were searched and a sample of 119 white and 36 Negro subjects was obtained. Table 16 presents background data on the inmates.

Table 16: Biographical Data - Federal Correctional Institution

	Group	$\bar{X}$	s	N	t(1)
Age	Total	21.07	1.95	155	
	White	21.02	2.07	119	
	Negro	20.61	1.68	36	1.08
Education (years)	Total	8.48	1.80	155	
	White	8.43	1.74	119	
	Negro	8.64	2.02	36	.61

(1) t ratios are between the white and Negro samples

Inspection of the above table reveals that the average inmate age was approximately 21 years, and the average educational level (highest grade completed) was 8.5. There were no significant differences in age or educational level between the white and Negro samples.

Predictor Comparisons

Scores on the Revised Beta Examination, administered to all inmates, were recorded from inmate files. The Revised Beta is a non-verbal intelligence test commonly used in penal institutions. Table 17 presents mean scores for white and Negro subjects. Whites scored significantly higher than Negroes on this test, even though the Beta is a non-verbal test. This finding is consistent with Tenopyr's (1967) assertion that non-verbal tests do not necessarily reduce mean differences between white and Negro subjects.

Table 17: Predictor Means, Standard Deviations, N's, and Tests of Significance of Mean Differences - Federal Correctional Institution

Predictor	Group	$\bar{X}$	s	N	t(1)
Beta IQ	Total	100.63	13.00	155	
	White	103.60	11.66	119	
	Negro	90.81	12.50	36	5.63**

(1) t ratios are between the means of the white and Negro samples

** p <.01

Criterion Comparisons

Two measures of educational performance were obtained. The first was a monthly rating of the inmates' classroom performance. Inmates were rated by their instructors using a four point scale on the following: (1) Classroom Participation, (2) Utilization of Class Time, (3) Interest and Initiative, (4) Academic Aptitude, and (5) Achievement. A subject's final rating was the average of his ratings on these five traits. At least two monthly ratings were required for a case to be included in the sample.

Table 18: Criterion Means, Standard Deviations, N's, and Tests of Significance of Mean Differences - Federal Correctional Institution

Criterion	Group	$\bar{X}$	s	N	t(1)
Ratings	Total	2.97	.55	115	
	White	2.96	.60	87	
	Negro	3.01	.43	28	.40
Change Score (SAT)	Total	.00	.74	130	
	White	.10	.74	99	
	Negro	-.14	.74	31	1.33

(1) t ratios are between the means of the white and Negro samples

As shown in Table 18 white and Negro subjects were found to be approximately equal in terms of mean criterion performance based on the ratings.

The second criterion measure obtained was a residual gain score (Manning and DuBois, 1962) based on changes in Stanford Achievement Test scores before and after the inmates were exposed to educational classes. The average time between testings was approximately three months. Discussions with instructors in the educational department indicated their preference for a gain score as a

criterion measure. However, they also point out that a general increase in test scores can be expected due to general adjustment of inmates to a confined environment.

Table 18 presents the means, standard deviations, and tests of significance between means on the Stanford Achievement Test change scores. No significant differences were found between the two groups.

Validity

Correlations between the Revised Beta and the criterion measures are presented in Table 19. The Revised Beta correlated significantly with the rating criterion for the white and Negro subgroups. However, since the relationship was in the opposite direction for the two ethnic groups, the correlation for the total group was not significant. The correlation between the Beta and the change score criterion was significant for the total sample and the white subgroup but not significant for the Negro subgroup.

Table 19: Predictor - Criterion Correlations
Federal Correctional Institution(1,2)

		Predictor Beta IQ	
Criterion	Group		N
Ratings	Total	14	115
	White	31**a	87
	Negro	-49* (10)	28
Change Score (SAT)	Total	23*	130
	White	23*	99
	Negro	09 (7)	31

(1) Decimals are omitted.
(2) Number in parentheses below the correlation for the Negro sample indicates the model illustrated (See Appendix A)
* $p < .05$
** $p < .01$
a indicates those models in which a significant difference exists between the validity coefficients for the two ethnic groups.

Models Illustrated

Viewing the data in terms of the models presented by Bartlett and O'Leary (1969) reveals that Model 10 was demonstrated. (See appropriate reference Figure in Appendix A). The correlation between the predictor

and rating criterion was positive for the white inmates but negative for the Negro inmates. Moreover, combining the two groups eliminated the validity of the Revised Beta as a predictor of the ratings. Thus, unless the scores were moderated on the basis of race no linear prediction of the rating criterion would be possible. This is a situation, however, where non-linear prediction would yield validity.

The relationship between the Revised Beta and the change score criterion illustrated Model 7. Although the test is appropriate as a predictor for the white sample it is inappropriate for the Negro sample. If the test were used as a selection device the result would be the rejection of qualified Negroes.

Only the example of Model 10 met the additional criterion of a significant difference between validity coefficients, as indicated by the superscript a in Table 19.

It is important to note that motivation of inmates in the test-taking situation is indeed a problem. Discussions with instructors raised questions concerning the reliability of the measures. Thus, the above data must be interpreted with extreme caution.

Table 20: Analysis of Covariance for Homogeneity of Regression - Federal Correctional Institution

	Ratings			Change Score		
	$F_1^{(1)}$	$F_2^{(2)}$	$F_3^{(3)}$	F_1	F_2	F_3
Beta IQ	6.64** df(2,111)	10.92** (1,111)	2.17 (1,112)	.66 (2,126)	.62 (1,126)	.71 (1,127)

(1) F_1 tests hypothesis that $E(Y_{ij}|X_{ij}) = a + bX_{ij}$ for all i groups.
(2) F_2 tests hypothesis that $E(Y_{ij}|X_{ij}) = a_i + bX_{ij}$ for all i groups.
(3) F_3 tests hypothesis that $E(Y_{ij}|X_{ij}) = a_i + b_iX_{ij}$ for all i groups.
** $p < .01$

Table 20 presents the results of the regression tests for the analysis of covariance (Potthoff, 1966). The significant F_1 ratio in the relationship between the Beta IQ and the rating criterion indicates that bias is

present. The significant F_2 ratio indicates that the difference in regression slopes is the major factor contributing to this bias.

All of the F ratios in the relationship between the Beta IQ and change scores were not significant, indicating that no bias was present.

Study 5: Home Office Clerical

Sample

A representative sample of clerical employees in the home office of a large industrial organization comprised the subject population of Study 5. Selecting one out of every five employees yielded a sample of 409 subjects of whom 363 were white and 46 Negro. Table 21 presents background characteristics for the total, white and Negro samples. Inspection of Table 21 reveals that the Negro sample is older and has been with the firm for a shorter period of time than the white sample.

Table 21: Biographical Data--Home Office Clerical

	Group	$\bar{X}$	s	N(1)	t(2)
Age	Total	26.24	10.62	405	
	White	26.72	11.02	359	
	Negro	28.85	6.04	46	3.56**
Tenure	Total	3.89	3.42	405	
(years)	White	4.15	3.59	359	
	Negro	1.89	1.37	46	8.11**

(1) Total N is less than 409 because of incomplete data for some subjects

(2) t ratios are between the means of the white and Negro samples
**$p < .01$

Predictor Comparisons

The major purpose of this validation study was to determine the relative utility of a new version of the Thurstone Test of Mental Alertness (TMA), as compared to the original TMA administered at the time of employment.

In addition to the original and new TMA, a company-developed nonverbal test of reasoning ability (The Picture Selection Index) was administered to the employees. Since this test was in its early development three time limits were examined--10, 15, and 20 minutes.

Table 22 presents means, standard deviations and tests of significance between means for the white and Negro samples. No significant differences were found between racial groups on the original TMA. However, the white sample scored significantly higher than the Negro sample on the new version of the TMA. The firm's psychologists indicate that this difference may be due to the increased verbal content of the new version.

The mean performance of the two racial groups on the Picture Selection Index was approximately equal. Moreover, increasing the time limits did not produce any mean differences between the two groups.

Criterion Comparisons

Employees were rated by both their Immediate Supervisor and Office Manager on the following dimensions, using a nine point rating scale: (1) Quickness in Understanding New Material, (2) Accuracy, (3) Numerical Ability, (4) Verbal Ability, (5) Judgment--the ability to make appropriate and sound decisions, and (6) Overall Mental Alertness. In addition, employees were rated on an eight point scale on their "General Promotability"-- a rating of the employee's potential top performance level.

The correlations between the Immediate Supervisor ratings and Office Manager ratings were:

Quickness	.58	Verbal Ability	.46
Accuracy	.58	Judgment	.47
Numerical Ability	.50	Mental Alertness	.59
		Promotional Potential	.69

Because of the rather low intercorrelations between the two sets of ratings, they were not combined into an overall rating of job performance. Rather, each rating was considered separately. It should be noted that a general halo factor was present in both samples.

Criterion means for the total group, whites, and Negroes are presented in Table 23. In general, the Negro's job performance is rated as being lower than the job performance of whites. A significant difference was found between the mean job performance ratings for the two racial groups on 11 out of the 14 possible rating criteria.

Table 22: Predictor Means, Standard Deviations, N's and Tests of Significance of Mean Differences

Home Office Clerical

Predictor	Group	$\bar{X}$	s	N	t[1]
Original TMA	Total	33.08	10.74	402	
Verbal	White	33.35	10.88	363	
	Negro	31.52	9.88	46	1.08
Quantitative	Total	23.67	8.14	409	
	White	23.66	8.40	363	
	Negro	23.37	6.70	46	.22
Total Score	Total	56.87	17.29	409	
	White	58.25	27.19	363	
	Negro	54.83	15.00	46	1.29
New TMA					
Verbal	Total	46.60	15.26	409	
	White	47.32	16.02	363	
	Negro	41.22	13.35	46	2.47*
Quantitative	Total	23.10	7.61	409	
	White	23.60	8.43	363	
	Negro	20.76	6.68	46	2.24*
Total Score	Total	69.69	20.57	409	
	White	72.02	32.25	363	
	Negro	61.91	16.59	46	3.37**
Picture Selection Index					
10-min. Time Limit	Total	36.14	8.87	355	
	White	36.54	9.59	318	
	Negro	34.19	7.20	37	1.44
15-min. Time Limit	Total	48.09	9.56	355	
	White	48.43	10.18	318	
	Negro	46.38	6.72	37	1.65
20-min. Time Limit	Total	54.65	9.18	355	
	White	54.57	9.75	318	
	Negro	54.00	6.63	37	.46

(1) t ratios are between means of white and Negro samples

*p<.05

**p<.01

Table 23: Criteria-Means, Standard Deviations, Ns and Tests of Significance of Mean Differences

Home Office Clerical

Criterion	Group	$\bar{X}$	s	N	t[1]
Quickness					
Office	Total	5.84	1.42	371	
Manager	White	5.92	1.41	328	
	Negro	5.26	1.45	43	2.88**
Immediate	Total	6.08	1.43	315	
Supervisor	White	6.13	1.41	284	
	Negro	5.58	1.50	31	2.04*
Accuracy					
Office	Total	6.07	1.40	352	
Manager	White	6.15	1.39	309	
	Negro	5.49	1.33	43	2.93**
Immediate	Total	6.03	1.52	315	
Supervisor	White	6.09	1.52	284	
	Negro	5.45	1.41	31	2.24*
Numerical Ability					
Office	Total	5.74	1.50	317	
Manager	White	5.81	1.47	280	
	Negro	5.27	1.63	37	2.07*
Immediate	Total	5.81	1.39	296	
Supervisor	White	5.87	1.41	268	
	Negro	5.32	1.12	28	1.99*
Verbal Ability					
Office	Total	5.59	1.43	371	
Manager	White	5.66	1.41	328	
	Negro	5.02	1.30	43	2.82**
Immediate	Total	5.68	1.39	315	
Supervisor	White	5.73	1.36	284	
	Negro	5.23	1.59	31	1.91
Judgment					
Office	Total	5.84	1.49	350	
Manager	White	5.90	1.48	308	
	Negro	5.43	1.56	42	1.92

Table 23 (concd).

Criterion	Group	$\bar{X}$	s	N	t[1]
Immediate	Total	5.77	1.53	314	
Supervisor	White	5.82	1.53	284	
	Negro	5.37	1.45	30	1.54
Overall Mental Ability					
Office	Total	5.87	1.47	370	
Manager	White	5.95	1.45	327	
	Negro	5.33	1.54	43	2.62*
Immediate	Total	6.09	1.41	315	
Supervisor	White	6.16	1.39	284	
	Negro	5.39	1.41	31	2.92**
Promotion Potential					
Office	Total	4.71	1.42	358	
Manager	White	4.79	1.39	318	
	Negro	4.05	1.57	40	3.12**
Immediate	Total	4.54	1.45	303	
Supervisor	White	4.61	1.45	273	
	Negro	3.87	1.38	30	2.66**

(1) t ratios are between means of white and Negro samples

*p<.05

**p<.01

Since the Negro sample had been with the firm for a shorter period than the white sample, correlations between tenure and the rating criteria were computed. The results indicated that job experience was not a major factor contributing to the obtained criterion differences for the two racial groups. The only significant relationship was between tenure and ratings by Office Managers on Numerical Ability for the Negro sample (r=.36).

Validity

Correlations between the various predictors and criteria are presented in Table 24. In general, ratings by Office Managers were more predictable than ratings by Immediate Supervisors for both racial groups.

Considering both the original and new TMA, we find that ratings of Verbal Ability and Mental Alertness by Office Managers are equally predictable for both racial groups. Moreover, with the exception of the quantitative score, the new TMA predicts Office Manager ratings of Numerical Ability and Promotion Potential for both racial groups equally well.

With few exceptions, ratings by Immediate Supervisors are predicted by both the original and new TMA for the white sample but are predictable in only two cases for the Negro sample.

Increasing the time limit from ten to fifteen minutes tends to increase the validity of the Picture Selection Index for both racial groups. A further increase in the time limit from fifteen to twenty minutes tends to yield a slight increase in validity for the white sample, but in some instances results in a decrease in validity for the Negro sample.

In general, the Picture Selection Index is not as valid as the original and new TMA. This finding is consistent with studies in the literature which report that nonverbal tests are not as valid as verbal tests.

Models Illustrated

The criteria used for identifying models was whether the correlation between a test and criterion was significantly greater than zero in neither, both, or one of the subgroups. It is important to note that in a number of comparisons in Table 24, the absolute magnitude of the correlation for the Negro sample is larger than the corresponding correlation for the

Table 24: Predictor-Criterion Correlations-Home Office Clerical

Total Group, Whites, Negroes[1,2]

Criterion		Predictor										
		Original TMA			New TMA			Picture Selection Index				
		Verbal	Quant	Total	Verbal	Quant	Total	N	10 min	15 min	20 min	N
Quickness (Off. Mgr.)	Total	35**	36**	38**	35**	34**	40**	371	25**	26**	25**	319
	White	35**	37**	38**	35**	35**	40**	328	24**	25**	25**	285
	Negro	31*	31*	34*	25	17	27	43	18	29	23	34
		(3)	(3)	(3)	(8)	(8)	(8)		(6)	(6)	(6)	
Quickness (Imm. Sup.)	Total	23**	27**	27**	26**	24**	29**	315	17**	23**	25**	270
	White	24**	30**	29**	27**	26**	31**	284	17*	23**	25**	247
	Negro	08	03	06	08	02	07	31	08	19	17	23
		(6)	(6)	(6)	(8)	(8)	(8)		(6)	(6)	(6)	
Accuracy (Off. Mgr.)	Total	22**	24**	24**	21**	23**	25**	352	11*	14**	17**	302
	White	21**	24**	24**	20**	22**	24**	309	09	12	16*	268
	Negro	23	21	24	17	18	21	43	15	29	21	34
		(6)	(6)	(6)	(8)	(8)	(8)				(6)	
Accuracy (Imm. Sup.)	Total	11*	13*	13*	15**	12*	16**	315	-01	03	06	270
	White	09	12*	11	13*	10	14*	284	-02	02	05	247
	Negro	30	27	31	20	20	24	31	-02	09	12	23
			(6)		(8)	(11)	(8)					
Numerical Ability (Off. Mgr.)	Total	31**	36**	35**	30**	32**	34**	317	19**	21**	23**	271
	White	30**	36**	34**	28**	32**	33**	280	18*	20**	22**	242
	Negro	39*	40*	44**	37*	22	38*	37	27	38*	37*	29
		(3)	(3)	(3)	(1)	(8)	(1)		(6)	(3)	(3)	
Numerical Ability (Imm. Sup.)	Total	14*	24**	20**	19**	16**	21**	296	13*	18**	16*	254
	White	14*	24**	19**	18**	16**	20**	268	13*	18*	15*	233
	Negro	23	22	25	21	09	20	28	08	15	20	21
		(6)	(6)	(6)	(8)	(8)	(8)		(6)	(6)	(6)	

Table 24: Continued

Criterion		Predictor: Original TMA Verbal	Original TMA Quant	Original TMA Total	New TMA Verbal	New TMA Quant	New TMA Total	N	Picture Selection Index 10 min	Picture Selection Index 15 min	Picture Selection Index 20 min	N
Verbal Ability (Off. Mgr.)	Total	43**	35**	43**	42**	31**	44**	371	14*	16**	17**	319
	White	42**	35**	42**	41**	29**	42**	328	13*	15*	16**	285
	Negro	43**	39**	46**	41**	35*	47**	43	13	19	21	34
		(3)	(3)	(3)	(1)	(1)	(1)		(6)	(6)	(6)	
Verbal Ability (Imm. Sup.)	Total	35**	23**	32**	41**	18**	39**	315	02	07	07	270
	White	35**	22**	32**	40**	18**	38**	284	02	07	06	247
	Negro	33	31	34	42*	19	40*	31	-11	-04	06	23
		(5)	(5)	(5)	(2)	(7)	(2)					
Judgment (Off. Mgr.)	Total	25**	28**	28**	26**	27**	30**	350	12	13	14*	300
	White	24**	27**	27**	25**	26**	29**	308	11	12	14*	267
	Negro	34*	32*	37*	26	26	31	42	17	29	22	33
					(7)	(7)	(2)				(5)	
Judgment (Imm. Sup.)	Total	16**	15**	17**	18**	12*	19**	314	14*	16*	17**	269
	White	15*	15*	16*	18**	10	18**	284	13*	16*	17**	247
	Negro	25	16	22	14	23	20	30	10	11	10	22
		(5)	(5)	(5)	(7)		(7)		(5)	(5)	(5)	
Mental Alertness (Off. Mgr.)	Total	42**	39**	44**	42**	37**	46**	370	26**	29**	30**	318
	White	42**	39**	44**	42**	36**	45**	327	26**	29**	31**	284
	Negro	43**	42**	47**	43**	32*	48**	43	16	26	17	34
		(3)	(3)	(3)	(1)	(1)	(1)		(6)	(6)	(6)	
Mental Alertness (Imm. Sup.)	Total	26**	28**	29**	28**	26**	32**	315	18**	25**	26**	270
	White	27**	31**	31**	29**	27**	33**	284	20**	26**	27**	247
	Negro	10	07	08	07	13	11	31	-17	-13	02	23
		(6)	(6)	(6)	(8)	(8)	(8)		(6)	(6)	(6)	

Table 24: Continued

Criterion		Predictor: Original TMA Verbal	Original TMA Quant	Original TMA Total	New TMA Verbal	New TMA Quant	New TMA Total	N	Picture Selection Index 10 min	Picture Selection Index 15 min	Picture Selection Index 20 min	N
Promotional Potential (Off. Mgr.)	Total	42**	41**	44**	46**	39**	50**	358	23**	27**	27**	309
	White	42**	42**	45**	46**	41**	51**	318	23**	26**	27**	277
	Negro	40**	31	40**	39**	21	40**	40	16	28	24	32
		(3)	(6)	(3)	(1)	(8)	(1)		(6)	(6)	(6)	
Promotional Potential (Imm. Sup.)	Total	31**	30**	33**	36**	30**	40**	303	17**	23**	24**	262
	White	32**	32**	35**	37**	30**	40**	273	19**	24**	24**	239
	Negro	21	19	21	17	19	21	30	-10	-05	04	23
		(6)	(6)	(6)	(8)	(8)	(8)		(6)	(6)	(6)	

(1) Decimals are omitted.
(2) Number in parentheses below the correlation for the Negro sample indicates the model illustrated (see Apendix A).
*p<.05
**p<.05

white sample, but the correlation is not significant in the Negro sample due to a relatively smally sample size.

Considering the new TMA, ten examples of Model 1 emerged. The number in parentheses below the correlations for the Negro sample in Table 24 indicates the model represented (See appropriate reference figure in Appendix A). White employees obtained higher mean scores on both the predictor and criterion in this situation, but the validity coefficients were approximately equal for the two racial groups. If it can be assumed that the rating criterion is unbiased, then discrimination on the test does not constitute unfair discrimination, since the test reflects a real difference in predicted performance.

Model 2 illustrates the situation in which mean differences exist on predictor performance for the two racial groups but no difference is present in the mean criterion performance for the two groups. Also, the correlation between the predictor and criterion is significant for both groups. This model, which was illustrated in the relationship between the new TMA and ratings of Verbal Ability and Judgment, occurred three times.

Model 3 occurred 16 times. In this model, the validity coefficients are approximately equal for the two groups. In addition, there are no differences in the mean predictor scores but significant differences between racial groups on the criterion. If the tests were validated only on the total group, the result would be an underprediction of performance for the white sample and an overprediction for the Negro sample. Differential prediction would yield more accurate prediction for both groups.

Model 5 is illustrated in the relationship between the Picture Selection Index and ratings of Judgment by Immediate Supervisors. Negro and white employees perform approximately equal on both the predictor and criterion, but the test is valid only for the white sample. The frequency of this model was 10.

Forty-three cases on Model 6, as illustrated in many of the relationships between the Picture Selection Index and the various rating criteria, and in some original TMA-criterion relationships, were found. In this model the two groups differ in mean performance on the criterion as well

as validity, but there is no difference in the predictor performance for the two racial groups. If this test were used in selection, the result would be to select only those white individuals with a high probability of success on the job, but to select Negro individuals whose probability of success on the job is not known.

The relationship between the new TMA and ratings of Judgement illustrates Model 7. White employees score significantly higher than Negroes on the predictor, but mean criterion ratings were approximately equal. However, the test was valid only for the white sample. This model occurred five times.

Twenty-two examples of Model 8 were illustrated in the relationships between scores on the new TMA and the various rating criteria. White employees scored higher than Negro employees on both the predictor and criterion measures, but the test was valid only for the white sample. One can make valid predictions using a combined group validation procedure even though the test is not valid for the Negro group, since the test identifies the lower performing group of Negroes. However, it is inappropriate to use the test to select Negroes.

Model 11, the final model illustrated in this sample, represents the situation in which a test is valid for both racial groups combined but has no validity for each subgroup separately. This model is illustrated in the relationship between the quantitative section of the new TMA and ratings of Accuracy by Immediate Supervisors.

As indicated above, the criterion used for identifying the above models was whether the correlation between a test and criterion was significantly greater than zero in neither, both, or one of the subgroups. An additional criterion can be applied to Models 5 through 10--that a significant difference must exist between the validity coefficients for the two racial groups. Applying this somewhat more restrictive criterion completely eliminates the Model 5, 6, 7, and 8 examples.

The analysis of covariance for homogeneity of regression (Pottoff, 1966) yields results which are consistent with the more restrictive definition of bias.

All of the F_2 ratios were not significant, indicating that a common regression slope was appropriate for both racial groups. Table 25 presents

the results of this analysis. The original TMA demonstrated the most bias using this method of analysis as indicated by the frequently significant F_3 ratios. A significant F_3 ratio means that a common intercept cannot be used for the two racial groups.

It should be noted that comparing only mean test performance one would conclude that the original TMA was less biased than the new TMA since white employees score higher than Negro employees on the new TMA. However, considering both test and criterion performance, as well as the relationship between them, one concludes that the original TMA is more biased than the new TMA in this particular sample.

Table 25. Analysis of Covariance for Homogeneity of Regression
Home Office Clerical[1]

Predictor	Criterion														
	Quickness Off. Mgr.			Quickness Imm. Sup.			Accuracy Off. Mgr.			Accuracy Imm. Sup.			Numerical Ability Off. Mgr.		
	F_1[2]	F_2[3]	F_3[4]	F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3
Original TMA															
Verbal	3.80*	.00	7.61*	2.09	.48	3.71	3.85*	.03	7.69*	2.94	1.14	4.74	2.46	.86	4.06
Quant.	4.76*	.01	9.54**	3.34	1.	4.89	4.52*	.01	9.06**	2.93	.58	5.29*	2.77	.67	4.87
Total	4.09*	.03	8.17**	2.64	1.16	4.12	4.03*	.04	8.03**	2.99	1.05	4.92	2.77	1.30	4.24
New TMA															
Verbal	2.69	.07	5.33*	1.59	.65	2.54	3.14	.01	6.29*	2.02	.14	3.91	1.75	1.02	2.45
Quant.	3.38	.79	5.98*	2.42	1.50	3.34	3.42	.02	6.82*	2.37	.21	4.54	1.36	.11	2.61
Total	2.40	.10	4.72	1.76	1.12	2.39	2.90	.01	5.82*	2.06	.27	3.86	1.41	.73	2.08
df	(2,367)	(1,367)	(1,368)	(2,311)	(1,311)	(1,312)	(2,348)	(1,348)	(1,349)	(2,311)	(1,311)	(1,312)	(2,313)	(1,313)	(1,314)
Picture Selection Index															
10 min.	1.67	.00	3.36	.96	.06	1.88	2.62	.24	5.01	1.12	.01	2.25	.71	.50	.93
15 min.	2.15	.66	3.65	.95	.07	1.84	3.33	1.62	5.02	1.11	.13	2.10	1.52	2.00	1.03
20 min.	2.20	.15	4.25	.88	.01	1.76	2.86	.44	5.30*	1.11	.20	2.02	1.43	1.67	1.19
df	(2,315)	(1,315)	(1,316)	(2,266)	(1,266)	(1,267)	(2,298)	(1,298)	(1,299)	(2,266)	(1,266)	(1,267)	(2,267)	(1,267)	(1,268)

Table 25: Continued

Home Office Clerical

Criterion

Predictor	Numerical Ability Imm. Sup.			Verbal Ability Off. Mgr.			Verbal Ability Imm. Sup.			Judgement Off. Mgr.			Judgement Imm. Sup.		
	F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3
Original TMA															
Verbal	1.89	.11	3.69	3.58	.01	7.17*	1.70	.17	3.24	1.90	.77	3.03	1.14	.34	1.94
Quant.	2.13	.05	4.21	4.54*	.25	8.84**	2.46	.66	4.26	2.47	.71	4.24	1.20	.01	2.39
Total	1.91	.01	3.82	3.98*	.22	7.76*	2.06	.37	3.76	2.22	1.08	3.36	1.12	.17	2.08
New TMA															
Verbal	1.46	.01	2.93	2.26	.05	4.49	1.06	.56	1.56	1.05	.15	1.96	.70	.02	1.39
Quant.	1.77	.17	3.38	2.93	.19	5.68*	1.59	.08	3.12	1.28	.10	2.47	1.17	.47	1.87
Total	1.42	.01	2.84	2.22	.41	4.04	1.11	.55	1.67	1.06	.42	1.71	.68	.05	1.32
df	(2,292)	(1,292)	(1,293)	(2,367)	(1,367)	(1,368)	(2,311)	(1,311)	(1,312)	(2,346)	(1,346)	(1,347)	(2,310)	(1,310)	(1,311)
Picture Selection Index															
10 min.	.70	.03	1.37	2.72	.01	5.45*	2.14	.44	3.85	1.25	.25	2.27	.25	.01	.50
15 min.	.68	.01	1.36	2.87	.16	5.60*	1.91	.16	3.66	2.02	1.68	2.35	.26	.01	.52
20 min.	.72	.10	1.34	3.11	.22	6.01*	1.83	.03	3.65	1.67	.72	2.62	.26	.02	.50
df	(2,250)	(1,250)	(1,251)	(2,315)	(1,315)	(1,316)	(2,266)	(1,266)	(1,267)	(2,296)	(1,296)	(1,297)	(2,265)	(1,265)	(1,266)

Table 25: Continued

Predictor	Criterion: Mental Alertness Off. Mgr.			Mental Alertness Imm. Sup.			Promotional Potential Off. Mgr.			Promotional Potential Imm. Sup.		
	F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3
Original TMA												
Verbal	3.20	.31	6.11*	4.33*	.61	8.07**	4.58*	.15	9.04**	3.70	.20	7.23*
Quant.	4.30*	.67	7.95**	5.62**	1.31	9.93**	5.65**	.01	11.33**	4.57*	.43	8.73**
Total	3.73*	.76	6.70**	4.99**	1.16	8.81**	4.93*	.15	9.73**	4.21*	.46	7.93*
New TMA												
Verbal	2.08	.55	3.61	3.55	.95	6.15*	3.29	.12	6.48*	2.89	.79	4.99
Quant.	2.25	.02	4.49	4.01*	.53	7.49*	4.08*	.76	7.40*	3.32	.41	6.24*
Total	2.02	.93	3.08	3.46	.96	5.96*	2.94	.05	5.85*	2.83	.82	4.84
df	(2,366)	(1,366)	(1,367)	(2,311)	(1,311)	(1,312)	(2,354)	(1,354)	(1,355)	(2,299)	(1,299)	(1,300)
Picture Selection Index												
10 min.	1.82	.02	3.64	3.98*	2.34	5.60*	3.03	.01	6.07*	2.88	1.37	4.38
15 min.	2.12	.36	3.89	3.88*	2.12	5.61*	3.57	.78	6.36*	2.69	.97	4.42
20 min.	2.28	.01	4.58	3.05	.62	5.50*	3.79*	.35	7.24*	2.32	.30	4.35
df	(2,314)	(1,314)	(1,315)	(2,266)	(1,266)	(1,267)	(2,305)	(1,305)	(1,306)	(2,258)	(1,258)	(1,259)

(1) Degrees of freedom for each ratio are shown in parentheses below each column.

(2) F_1 tests hypothesis that $E(Y_{ij} | X_{ij}) = a + bX_{ij}$ for all i groups.

(3) F_2 tests hypothesis that $E(Y_{ij} | Y_{ij}) = a_i - bX_{ij}$ for all i groups.

(4) F_3 tests hypothesis that $E(Y_{ij} | X_{ij}) = a_i + b_i X_{ij}$ for all i groups.

*p<.05

**p<.01

Study 6: Catalog Order Plants

Sample

Study 6 consisted of 810 employees of a large retail organization of whom 472 were white, 287 Negro, and 51 Latin American. All jobs were essentially clerical in nature and most required some arithmetic skills. The sample has been broken down into specific job classifications wherever feasible.

Predictors

The same predictors were used for all job classifications. Two experimental clerical tests, developed by the firm's psychologists were administered to all employees. Clerical I consists of two columns of names and numbers and the task of the subject is to determine whether each is alike or different. Clerical II is a number cancellation task in which the subject is required to strike out all numbers in a column that are the same as the underlined number at the top of the column. Since these tests were experimental in nature two time limits were examined--5 minutes and 10 minutes. Also each test was scored in two ways: (1) Number Correct and (2) Number Correct minus Number Wrong.

In addition to the two experimental tests, scores on a company developed Arithmetic Reasoning test and a Verbal Reasoning test were obtained for all employees in the sample.

Criteria

Ratings by supervisors were obtained for all employees. The rating instrument was a seven point scale developed by the firm's psychologists covering the following dimensions:

(1) Accuracy: The ability to work without making errors.
(2) Accuracy under Pressure: The ability to turn in accurate work under differing conditions of pressure.
(3) Work Speed: The pace at which a person works.
(4) Learning Ability: The ability to understand directions and learn from the directions provided.
(5) Human Relations: The ability to maintain good relations with others.
(6) General Overall Effectiveness.

Background Data - Merchandise Handlers I

Table 26 presents the biographical data obtained for this job classification. A number of employees of Latin American extraction were employed in this job classification in addition to the Negro minority. Each minority group was compared separately with the white sample.

Table 26: Biographical Data-Merchandise Handlers I

	Group	$\bar{X}$	s	N	t
Age	Total	30.79	11.43	190	
	White	35.52	13.45	86	
	Negro	26.61	7.18	84	5.37**(1)
	Latin	28.00	8.83	20	3.01**(2)
Tenure (Years)	Total	2.35	1.15	190	
	White	2.92	1.19	86	
	Negro	1.80	.86	84	7.00**
	Latin	2.20	.83	20	2.54*
Education (Years)	Total	10.72	1.81	190	
	White	9.95	1.84	86	
	Negro	11.58	1.22	84	6.78**
	Latin	10.30	2.18	20	.74

(1) t ratios are between the means of the white and Negro samples.

(2) t ratios are between the means of the white and Latin samples.

*$p < .05$

**$p < .01$

Negro and Latin employees, as compared to their white counterparts are younger and have been with the firm for a shorter period of time. The educational level of the Negro employees is significantly higher than that of the white employees. However, the educational level of the white and Latin employees is approximately equal.

Predictor Comparisons

Mean predictor scores for the total, white, Negro, and Latin samples are presented in Table 27. White employees score significantly higher than either the Negro or the Latin sample on the Verbal Reasoning Test. There were no significant differences between the performance of the two minority groups and the white sample on any of the other predictors.

It should be noted that, although the mean differences between each minority group and the white sample were not significant, a rather consistent ranking pattern emerged across all predictors: white employees scored higher than Negro employees who, in turn, scored higher than Latin employees.

Criterion Comparisons

As indicated in Table 28, there were no differences in the job performance of the three ethnic groups as measured by supervisory ratings.

Validity

Table 29 presents validity coefficients for the total, white and Negro samples. Since a significant relationship was found to exist between tenure and the various criteria, correlations have been controlled for tenure where appropriate. The clerical tests appear equally valid across all criteria. This generalization holds regardless of the time limit imposed or the utilization of a correction-for-guessing formula.

All forms of Clerical Tests I and II were valid predictors of the six rating criteria. Moreover, with few exceptions, the validity coefficients were approximately equal for the white and Negro samples. Validities for both the Verbal Reasoning and the Arithmetic Reasoning Tests tended to be lower than those of Clerical Tests I and II.

Predictor-criterion correlations for the total, white, and Latin samples are presented in Table 30. Inspection of the table reveals that even though the absolute magnitude of the correlations for the Latin sample are relatively high, sometimes exceeding those for the white sample, only a few are statistically significant. Clerical Test I predicts more criteria for the Latin sample than any of the other predictors.

Table 27: Predictors-Means, Standard Deviations, N's, and Tests of Significance of Mean Differences

Merchandise Handlers I

Predictor	Group	$\bar{X}$	s	N	t
Verbal Reasoning	Total	18.04	8.72	190	
	White	20.29	9.66	86	
	Negro	17.15	7.44	84	2.36*(1)
	Latin	12.05	5.66	20	4.94**(2)
Arithmetic Reasoning	Total	20.71	7.37	190	
	White	21.48	7.97	86	
	Negro	20.40	6.90	84	.94
	Latin	18.65	6.40	20	1.47
Clerical I 5 minutes	Total	42.97	13.47	190	
	White	44.20	14.69	86	
	Negro	42.55	12.51	84	.78
	Latin	39.50	11.68	20	1.33
Clerical I 10 minutes	Total	89.02	26.77	190	
	White	91.58	28.18	86	
	Negro	87.55	25.71	84	.97
	Latin	84.20	24.99	20	1.07
Clerical II 5 minutes	Total	54.89	12.98	190	
	White	55.79	14.50	86	
	Negro	54.07	11.51	84	.85
	Latin	54.50	12.21	20	.37
Clerical II 10 minutes	Total	107.21	22.84	190	
	White	109.83	23 83	86	
	Negro	105.10	22.21	84	1.33
	Latin	104.85	20.89	20	.86
Clerical I (R-W) 5 minutes	Total	35.96	16.03	190	
	White	37.24	17.43	86	
	Negro	35.74	14.67	84	.60
	Latin	31.40	15.07	20	1.38
Clerical I (R-W) 10 minutes	Total	76.86	31.89	190	
	White	79.06	34.74	86	
	Negro	75.95	29.34	84	.63
	Latin	70.70	29.88	20	.99
Clerical II (R-W) 5 minutes	Total	47.47	15.05	190	
	White	47.99	16.85	86	
	Negro	47.08	13.74	84	.38
	Latin	46.85	12.44	20	.28
Clerical II (R-W) 10 minutes	Notal	92.71	26.24	190	
	White	94.74	28.05	86	
	Negro	90.93	25.65	84	.92
	Latin	91.40	20.55	20	.50

*p<.05

**p<.01

(1) t ratios are between the means of the white and Negro samples.

(2) t ratios are between the means of the white and Latin samples.

Table 28: Criteria- Means, Standard Deviations, N's,
and Tests of Significance of Mean Differences

Merchandise Handlers I

Criterion	Group	$\bar{X}$	s	N	t
Accuracy	Total	4.01	1.08	190	
	White	4.02	.96	86	
	Negro	3.92	1.16	84	.61(1)
	Latin	4.30	1.22	20	1.11(2)
Accuracy Under Pressure	Total	3.88	1.06	190	
	White	3.77	.98	86	
	Negro	3.95	1.12	84	1.11
	Latin	4.05	1.15	20	1.11
Work Speed	Total	3.92	1.01	190	
	White	3.90	1.04	86	
	Negro	3.85	1.01	84	.32
	Latin	4.30	1.15	20	1.59
Learning Ability	Total	3.98	.97	190	
	White	3.97	.93	86	
	Negro	4.00	.96	84	.21
	Latin	3.95	1.23	20	.08
Human Relations	Total	4.10	1.08	190	
	White	4.06	1.02	86	
	Negro	4.12	1.09	84	.37
	Latin	4.20	1.28	20	.52
Overall Effectiveness	Total	4.08	.91	190	
	White	4.05	.85	86	
	Negro	4.06	.95	84	.07
	Latin	4.35	.99	20	1.37

(1) t ratios are between the means of the white and Negro samples.
(2) t ratios are between the means of the white and Latin samples.

Table 29. Predictor-Criterion Correlations-Merchandise Handlers I
Total Group, Whites, and Negroes (1,2)

Criterion		Predictor										
		Verbal Reasoning	Arithmetic Reasoning	Clerical I 5 minutes	Clerical I 10 minutes	Clerical II 5 minutes	Clerical II 10 minutes	Clerical I (R-W 5 min)	Clerical I (R-W 10 min)	Clerical II (R-W 5 min)	Clerical II (R-W 10 min)	N
Accuracy	Total	21**	22**	28**	29**	35**	32**	29**	29**	37**	34**	190
	White	21	31**	22*	22*	32**	23*	24*	22*	33**	21	86
	Negro	28**	24*	33**	37**	42**	39**	35**	37**	46**	45**	84
Accuracy Under Pressure	Total	17*	24**	30**	32**	32**	27**	33**	33**	37**	(5)	
	White	25*	43**	32**	32**	35**	29**	36**	35**	41**	30**	190
	Negro	22* (2)	21 (5)	33**	38**	33**	25*	35**	39**	40**	26* 35**	86 84
Work Speed	Total	16*	29**	34**	34**	38**	34**	34**	34**	38**	34*	190
	White	19	42**	37**	34**	36**	34**	38**	35**	37**	28*	86
	Negro	17	23*	31**	35**	36**	31**	32**	36**	38**	36**	84
Learning Ability	Total	25**	29**	39**	40**	37**	38**	41**	41**	37**	36**	190
	White	34**	45**	40**	40**	49**	46**	43**	43**	48**	37**	86
	Negro	21 (7)	24*	39**	42**	30**	34**	40**	42**	32**	37**	84
Human Relations	Total	06	10	23**	24**	19**	21**	23**	24**	19**	20**	190
	White	01	16	23*	22*	22*	19	24*	24*	23*	16	86
	Negro	06	15	25*	28**	18 (5)	23* (5)	26*	28**	19 (5)	25* (5)	84
Overall Effectiveness	Total	22**	20**	39**	40**	37**	35**	39**	40**	36**	33**	190
	White	34**	38**	48**	48**	40**	40**	49**	49**	41**	39**	86
	Negro	23*	15 (5)	32**	32**	37**	37**	32**	36**	38**	39**	84

(1) Decimals are omitted.
(2) Number in parentheses below the correlation for the Negro sample indicates the model illustrated (see Appendix A).
* $p < .05$
** $p < .01$

Table 30. Predictor-Criterion Correlations-Merchandise Handlers I

Total Group, Whites and Latins[1,2]

Criterion		Predictor: Verbal Reasoning	Arithmetic Reasoning	Clerical I 5 minutes	Clerical I 10 minutes	Clerical II 5 minutes	Clerical II 10 minutes	Clerical I (R-W 5 min)	Clerical I (R-W 10 min)	Clerical II (R-W 5 min)	Clerical II (R-W 10 min)	N
Accuracy	Total	21**	22**	28**	29**	35**	32**	29**	29**	37**	34**	190
	White	21	31**	22*	22*	32**	23*	24*	22*	33**	21	86
	Latin	19	08 (5)	40 (5)	37 (5)	28 (5)	38 (5)	36 (5)	31 (5)	11 (5)	25	20
Accuracy Under Pressure	Total	17*	24**	30**	32**	32**	27**	33**	33**	37**	30**	190
	White	25*	43**	32**	32**	35**	29**	36**	35**	41**	26*	86
	Latin	02 (7)	16 (5)	33 (5)	33 (5)	28 (5)	(5)	27 (5)	26 (5)	11 (5)	25 (5)	20
Wor Speed	Total	16*	29**	34**	34**	38**	34**	34**	34**	38**	34**	199
	White	19	42**	37**	34**	36**	34**	38**	35**	37**	28*	86
	Latin	18	20 (5)	54*	55*	54*	54*	45*	44 (5)	32 (5)	39 (5)	20
Learning Ability	Total	25**	29**	39**	40**	37**	38**	41**	42**	37**	36**	190
	White	34**	45**	40**	40**	49**	46**	43**	43**	48**	37**	86
	Latin	07 (7)	0- (5)	44 (5)	48*	20 (5)	30 (5)	42 (5)	45*	09 (5)	18 (5)	20
Human Relations	Total	06	10	23**	24**	19**	21**	23**	24**	19**	20**	190
	White	04	16	23*	22*	22*	19	24*	24*	23*	16	86
	Latin	24	35	23 (5)	25 (5)	05 (5)	18	17 (5)	15 (5)	12 (5)	06	20
Overall Effectiveness	Total	22**	20**	39**	40**	37**	35**	39**	40**	36**	33**	190
	White	34**	38**	48**	48**	40**	40**	49**	49**	41**	39**	86
	Latin	14 (7)	19 (5)	50*	55*	28 (5)	32 (5)	44 (5)	48*	14 (5)	15 (5)	20

(1) Decimals are omitted.

(2) Number in parentheses below the correlation for the Latin sample indicates the model illustrated (see Appendix A).

* $p<.05$

** $p<.01$

Models Illustrated

It should be emphasized that in the majority of predictor-criterion relationships examined for the white and Negro samples, no bias was shown.

Three models were illustrated in the comparisons of the white and Negro sample The number of the model illustrated is shown below the correlations for the Negro sample in Table 29. The relationship between Verbal Reasoning and ratings of Accuracy Under Pressure demonstrates Model 2. White employees scored higher on the predictor but there was no difference between the two ethnic groups on the criteria. Moreover, the validity coefficients were approximately equal for the two groups. Using a total group validation procedure would result in the elimination of Negroes whose probability of job success is equal to that of the white employees selected.

The most frequently illustrated model was Model 5, occurring seven times. Model 5 is illustrative of the situation where a test has validity for one group, none for the other, yet mean performance on both the predictor and criterion is not significantly different for the two groups. In four of the seven cases, the test was valid only for the white sample. The use of such tests as selection instruments would result in the selection of better performing employees from the valid group, while no increase in prediction efficiency is obtained by using the test for selection of individuals from the non-valid group.

Model 7 was illustrated in the relationship between Verbal Reasoning and ratings of Learning Ability. Again, white employees score higher on the predictor than Negroes but their job performance is approximately equal. However, the test is valid only for the white subgroup. Since the Negro sample scores lower on the predictor the probability of a Negro being selected is lower than the probability of a white being selected. Thus, by using such a test as a selection device one would eliminate Negroes whose probability of of job success is equal to that of the white individuals selected.

Inspection of Table 30 reveals that forty cases of model 5 were represented in the comparisons of the validity patterns for the Latin and white samples. Because of the small sample size for the Latin sample, a rather larger correlation ($r > .44$) is required for significance at the .05

level. Thus, a number of the correlations for the Latin sample may not be significant even though the absolute magnitude of the correlation is larger than the significant correlation for the white sample.

Three cases of Model 7 were illustrated in the relationship between Verbal Reasoning and the rating criteria. Although the ratings were approximately equal for the two ethnic groups, the predictor was valid only for the white sample. Since the Latin sample obtained lower predictor scores, they would have a lower probability of being selected, even though the criterion performance of the two ethnic groups was similar.

Applying the additional criterion of a significant difference between validity coefficients eliminates all illustrations of Models 5 and 7 in both the white and Negro comparisons as well as the white and Latin comparisons.

Table 31 presents the results of the regression tests of the analysis of covariance (Potthoff, 1966). This analysis simultaneously tests the hypothesis that the regression slopes and intercepts are equal for the three ethnic groups. All the F ratios were not significant indicating that no bias was present.

Table 31. Analysis of Covariance for Homogeneity of Regression

Merchandise Handlers I [1]

	Accuracy			Accuracy Under Pressure			Work Speed			Learning Ability			Human Relations			Overall Effectiveness		
	F_1 [2]	F_2 [3]	F_3 [4]	F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3
Verbal Reasoning	1.24	.27	2.23	1.26	.15	2.39	1.50	.07	2.96	.50	.35	.65	.45	.48	.41	1.39	.05	2.77
Arithmetic Reasoning	1.19	.95	1.42	2.21	2.73	1.65	2.07	1.46	2.66	1.08	1.97	.18	1.56	2.22	.29	2.01	2.54	1.45
Clerical I 5 minutes	1.24	.95	1.54	.94	.12	1.78	1.58	.38	2.80	.39	.55	.24	.27	.08	.46	.38	.61	2.15
Clerical I 10 minutes	1.05	.73	1.37	.92	.06	1.79	1.40	.32	2.49	.46	.65	.26	.26	.09	.45	1.33	.72	1.94
Clerical II 5 minutes	1.09	1.08	1.09	.79	.09	1.50	1.06	.13	2.01	.32	.45	.19	.22	.19	.26	.78	.29	1.29
Clerical II 10 minutes	.96	.74	1.19	1.11	.54	1.69	1.32	.45	2.20	.40	.48	.33	.20	.01	.40	.78	.01	1.57
Clerical I (R-W) 5 min	1.26	.85	1.68	.95	.07	1.85	1.52	.10	2.97	.25	.26	.24	.25	.06	.45	1.23	.24	2.24
Clerical I (R-W) 10 min	1.11	.78	1.43	.89	.12	1.67	1.28	.04	2.55	.28	.38	.18	.26	.12	.39	1.10	.31	1.91
Clerical II (R-W) 5 min	1.54	1.85	1.22	1.01	.69	1.34	1.07	.09	2.06	.36	.63	.08	.59	.96	.21	.87	.51	1.23
Clerical II (R-W) 10 min	1.48	1.84	1.11	.82	.30	1.35	1.01	.10	1.93	.08	.03	.14	.24	.23	.25	.86	.54	1.19

(1) Degrees of freedom for all comparisons: F_1 - (4,184); F_2 - (2,184); F_3 - (2,186)

(2) F_1 tests hypothesis that $E(Y_{ij}|X_{ij}) = a + bX_{ij}$ for all i groups.

(3) F_2 tests hypothesis that $E(Y_{ij}|X_{ij}) = a_i + bX_{ij}$ for all i groups.

(4) F_3 tests hypothesis that $E(Y_{ij}|X_{ij}) = a_i + b_iX_{ij}$ for all i groups.

Background Data - Merchandise Handlers II

Table 32 presents the biographical data for the above job classification sample. This sample included a small number of employees of Latin-American extraction.

Negro employees in this job classification were younger than the white employees and had relatively shorter company service. The mean educational level of the Negro sample was approximately one year above the white sample. Biographical characteristics of the Latin sample tended to be similar to the white sample. Mean scores for the two groups did not differ significantly.

Table 32: Biographical Data-Merchandise Handlers II

	Group	$\bar{X}$	s	N	t
Age	Total	29.58	10.09	264	
	White	32.10	12.15	122	
	Negro	27.09	6.95	125	3.95**
	Latin	29.76	9.27	17	.76[2]
Tenure (Years)	Total	2.50	1.12	264	
	White	2.84	1.16	122	
	Negro	2.16	.97	125	4.97**
	Latin	2.59	1.18	17	.83
Education (Years)	Total	11.04	1.92	259	
	White	10.65	2.18	118	
	Negro	11.48	1.53	124	3.40**
	Latin	10.53	2.00	17	.21

[1] t ratios are between the means of the white and Negro samples.

[2] t ratios are between the means of the white and Latin samples.

**$p < .01$

Predictor Comparisons

Mean predictor scores for the total group, whites, Negroes, and Latins are presented in Table 33. White employees scored significantly higher than Negro employees on all tests except Clerical Test II. It is important to note that increasing the time limits of the tests did not reduce these racial differences.

Predictor scores for the Latin sample tended to approximate those of the white sample. Scores for these two ethnic groups differed only in one comparison; white employees obtained higher scores than Latins on the Verbal Reasoning Test.

Criterion Comparisons

Mean criterion data for the three ethnic groups is presented in Table 34. Ratings for white employees were significantly higher than those for Negro employees only on the criterion of Learning Ability. Correlations between tenure and the rating criteria were not significant, indicating that experience was not a major factor contributing to the obtained mean criterion differences for the white and Negro samples.

Comparisons of the mean criterion performance of the Latin and white samples yielded no significant differences.

Validity

Correlations between the predictors and criteria are presented in Table 35. Again the clerical tests produced higher correlations with the various criteria than either the Verbal or Arithmetic Reasoning Test. Similar validity patterns were exhibited by both of the clerical tests with Accuracy, Learning Ability, and Work Speed being the most predictable criteria.

Comparing the Negro and white sample, we find that in 18 out of a possible 60 instances, a test correlated significantly with the criterion for one racial group but not the other. It should be noted that it was not always the white group which was more predictable. In fact, in over half of these cases the test was valid for the Negro sample, but not valid for the white sample.

With few exceptions, increasing the time limit on the clerical tests from five to ten minutes resulted in an increase in the validity coefficients for all ethnic groups.

Table 33: Predictors-Means, Standard Deviations, N's and Tests of Significance of Mean Differences

Merchandise Handlers II

	Group	$\bar{X}$	s	N	t[1]
Verbal Reasoning	Total	19.61	10.44	264	
	White	22.44	11.02	122	
	Negro	17.80	9.42	125	3.54**(1)
	Latin	12.65	6.96	17	3.54**(2)
Arithmetic Reasoning	Total	21.14	8.81	264	
	White	23.75	9.53	122	
	Negro	18.42	7.49	125	4.87**
	Latin	22.41	6.69	17	.71
Clerical I 5 minutes	Total	40.38	12.21	264	
	White	42.84	12.08	122	
	Negro	37.56	11.77	125	3.46**
	Latin	43.53	12.47	17	.22
Clerical I 10 minutes	Total	85.08	24.92	264	
	White	91.02	24.10	122	
	Negro	78.34	24.20	125	4.11**
	Latin	92.06	24.50	17	.17
Clerical II 5 minutes	Total	53.48	12.06	264	
	White	54.61	12.83	122	
	Negro	52.65	11.14	125	1.27
	Latin	51.53	12.81	17	.92
Clerical II 10 minutes	Total	103.45	22.60	264	
	White	105.80	21.79	122	
	Negro	101.38	23.32	125	1.53
	Latin	101.94	22.67	17	.68
Clerical I (R-W) 5 minutes	Total	33.31	14.37	264	
	White	36.32	14.27	122	
	Negro	30.02	13.73	125	3.52**
	Latin	35.94	15.21	17	.10
Clerical I (R-W) 10 minutes	Total	73.26	28.96	264	
	White	80.69	27.50	122	
	Negro	69.10	28.09	125	4.39**
	Latin	79.94	30.96	17	.10
Clerical II (R-W) 5 minutes	Total	46.73	12.95	264	
	White	48.68	13.69	122	
	Negro	45.11	11.60	125	2.20**
	Latin	44.59	15.60	17	1.13
Clerical II (R-W) 10 minutes	Total	90.52	23.46	264	
	White	93.71	24.01	122	
	Negro	87.43	22.26	125	2.12**
	Latin	90.29	26.28	17	.54

*p< .05

**p< .01

(1) t ratios are between the means of the white and Negro samples.

(2) t ratios are between the means of the white and Latin samples.

Table 34: Criteria- Means, Standard Deviations, N's and Tests of Significance of Mean Differences

Merchandise Handlers II

	Group	$\bar{X}$	s	N	t
Accuracy	Total	3.81	1.01	264	
	White	3.86	1.06	122	
	Negro	3.77	.94	125	.70(1)
	Latin	3.77	1.20	17	.32(2)
Accuracy Under Pressure	Total	3.74	1.01	264	
	White	3.75	1.06	122	
	Negro	3.70	.97	125	.39
	Latin	3.94	1.03	17	.69
Work Speed	Total	3.79	1.03	264	
	White	3.91	1.14	122	
	Negro	3.66	.88	125	1.92
	Latin	3.88	1.22	17	.10
Learning Ability	Total	3.66	.88	264	
	White	3.79	.90	122	
	Negro	3.53	.84	125	2.34*
	Latin	3.65	1.00	17	.59
Human Relations	Total	3.85	1.05	264	
	White	3.86	1.15	122	
	Negro	3.86	.97	125	.00
	Latin	3.76	.83	17	.34
Overall Effectiveness	Total	3.80	1.08	264	
	White	3.89	1.21	122	
	Negro	3.70	.94	125	1.44
	Latin	4.00	1.06	17	.32

(1) t ratios between the means of the white and Negro samples.
(2) t ratios between the means of the white and Latin samples.
*$p<.05$

Table 35: Predictor-Criterion Correlations-Merchandise Handlers II

Total Group, Whites, and Negroes[1,2]

Criterion		Predictor: Verbal Reasoning	Arithmetic Reasoning	Clerical I 5 minutes	Clerical I 10 minutes	Clerical II 5 minutes	Clerical II 10 minutes	Clerical I (R-W 5 min)	Clerical I (R-W 10 min)	Clerical II (R-W 5 min)	Clerical II (R-W 10 min)	N
Accuracy	Total	21**	24**	23**	28**	20**	29**	21**	25**	22**	30**	264
	White	22*	34**	30**	35**	26**	38**	25**	29**	24**	34**	122
	Negro	21*	13	20*	26**	14	22**	19*	25**	22**	27**	125
		(2)	(7)	(2)	(2)	(5)		(2)	(2)	(2)	(2)	
Accuracy Under Pressure	Total	10	13*	20**	24**	17**	25**	16**	20**	18**	25**	264
	White	05	17	20*	24**	21*	30**	16	19*	19*	28**	122
	Negro	22*	06	20*	25**	13	19*	17*	24**	19*	22**	125
		(7)		(2)	(2)	(5)		(7)	(2)	(2)	(2)	
Work Speed	Total	17**	24**	23**	26**	21**	29**	20**	22**	23**	30**	264
	White	16	22*	22*	25**	22*	30**	20*	22*	22*	29**	122
	Negro	18*	23**	23**	28**	22**	29**	19*	23**	26**	31**	125
		(7)	(2)	(2)	(2)			(2)	(2)	(2)	(2)	
Learning Ability	Total	21**	21**	24**	30**	27**	35**	21**	26**	29**	35**	264
	White	13	25**	23**	29**	30**	39**	21*	26**	28**	36**	122
	Negro	27**	11	29**	36**	29**	35**	25**	32**	34**	37**	125
		(8)	(8)	(1)	(1)	(3)	(3)	(1)	(1)	(1)	(1)	
Human Relations	Total	05	11	19**	22**	22**	25**	12*	16**	20**	21**	264
	White	-07a	06	11	11a	16	21*	07	07	16	18*	122
	Negro	22**	19*	31**	38**	30**	28**	22**	30**	26**	26**	125
		(7)	(7)	(7)	(7)	(5)		(7)	(7)	(7)	(2)	
Overall Effectiveness	Total	12*	15*	17**	19**	18**	25**	12*	15*	19**	24**	264
	White	11	17	18*	21*	22*	30**	11	17	20*	27**	122
	Negro	17*	11	18*	19*	15	20*	12	15	18*	20*	125
		(7)		(2)	(2)	(5)				(2)	(2)	

(1) Decimals are omitted.

(2) Number in parentheses below the correlation for the Negro sample indicates the model illustrated (See Appendix A).

* $p<.05$

** $p<.01$

a indicates those models in which a significant difference exists between the validity coefficients for the two ethnic groups.

Table 36: Predictor-Criterion Correlations-Merchandise Handlers II
Total Group, Whites, and Latins[1,2]

Criterion		Predictor: Verbal Reasoning	Arithmetic Reasoning	Clerical I 5 minutes	Clerical I 10 minutes	Clerical II 5 minutes	Clerical II 10 minutes	Clerical I (R-W 5 min)	Clerical I (R-W 10 min)	Clerical II (R-W 5 min)	Clerical II (R-W 10 min)	N
Accuracy	Total	21**	24**	23**	28**	20**	29**	21**	25**	22**	30**	264
	White	22*	34**	30**	35**	26**	38**	25**	29**	24**	34**	122
	Latin	10	06	01	-08	04	18	-03	-12	06	19	17
		(7)	(5)	(5)	(5)	(5)	(5)	(5)	(5)	(5)	(5)	
Accuracy Under Pressure	Total	10	13*	20**	24**	17**	25**	16**	20**	18**	25**	264
	White	05	17	20*	24**	21*	30**	16	19*	19*	28**	122
	Latin	-20	19	06	07	10	30	01	01	05	22	17
				(5)	(5).	(5)	(5)		(5)	(5)	(5)	
Work Speed	Total	17**	24**	23**	26**	21**	29**	20**	22**	23**	30**	264
	White	16	22*	22*	25**	22*	30**	20*	22*	22*	29**	122
	Latin	-02	-02	09	03	03	15	01	-05	01	13	17
			(5)	(5)	(5)	(5)	(5)	(5)	(5)	(5)	(5)	
Learning Ability	Total	21**	21**	24**	30**	27**	35**	21**	26**	29**	35**	264
	White	13	25**	23**	29**a	30**	39**	21*	26**a	28**	36**	122
	Latin	09	-01	-29	-44	-16	-02	-27	-43	-05	03	17
			(5)	(5)	(5)	(5)	(5)	(5)	(5)	(5)	(5)	
Human Relations	Total	05	11	19**	22**	22**	25**	12*	16**	20**	21**	264
	White	-07	06	11	11	16	21*	07	07	16	18*	122
	Latin	-08	06	11	14	19	34	-01	-02	14	22	17
							(5)				(5)	
Overall Effectiveness	Total	12*	15*	17**	19**	18**	25**	12*	15*	19**	24**	264
	White	11	17	18*	21*	22*	30**	14	17	20*	27**	122
	Latin	-28	-12	-27	-20	-03	18	-26	-20	05	15	17
				(5)	(5)	(5)	(5)			(5)	(5)	

(1) Decimals are omitted.
(2) Number in parentheses below the correlation for the Latin sample indicates the model illustrated (See Appendix A).
* $p<.05$
** $p<.01$
a Indicates those models in which a significant difference exists between the validity coefficients for the two ethnic groups.

Although the Latin sample closely resembles the white sample with regard to both predictor and criteria performance, the test validity pattern of the white sample was not mirrored closely by the Latin sample. It should be noted that a relatively large correlation ($r>.46$) was required for significance at the .05 level for the small Latin sample.

Models Illustrated

Six cases of Model 1 were illustrated in the comparisons between the white and Negro samples. The number of the specific model illustrated is shown below the correlations for the Negro sample in Table 35. The white sample scored significantly higher than the Negro sample on the clerical tests and also on the criterion of Learning Ability. Moreover, the validity is approximately equal for the two racial groups. In this situation discrimination on the test reflects a real difference in predicted performance. Thus, selection with the test does not constitute unfair discrimination.

Model 2, occurring 24 times, represents the situation where there is a significant difference between the mean predictor scores for the two racial groups, yet no significant difference in the criterion. The correlation between the predictor and criterion is approximately equal for the two groups. If a cutting score were set on the basis of the total sample, the Negro group would not have an equal probability of being selected, even though their chances of job success were essentially equal.

Two illustrations of Model 3 were represented in the relationship between Clerical Test II and ratings of Learning Ability. Validities for the two racial groups were essentially equal. Although there was no difference in the mean predictor performance for the two racial groups, the white sample obtained higher ratings of job performance. Total group validation would result in an underprediction for white employees and an overprediction for Negro employees.

Examining the relationship between scores on Clerical Test II (5 minutes) and four criteria, we find four cases of Model 5. Predictor

and criterion performance was approximately equal for the two ethnic groups. However, the test was valid for one group but not the other. In over half of these cases the predictor was valid for the white sample, but invalid for the Negro sample. The result of using such a test would be the selection of better performing persons from the valid group than from the invalid group.

Illustrated twelve times, Model 7 is representative of the situation where there is a difference in predictor performance but no difference in criterion performance for the two groups. Also, the predictor-criterion correlations are valid for only one subgroup. It is interesting to note that the predictor is valid for the Negro sample in eleven out of the twelve cases.

The final model illustrated in the Negro-white comparison was Model 8. Performance of white employees is higher than Negroes not only on the tests of Verbal and Arithmetic Reasoning, but also on the ratings of Learning Ability. The Arithmetic Reasoning Test was valid for the white sample while the Verbal Reasoning Test was valid for the Negro sample.

Forty-two illustrations of Model 5 were found in the comparisons of the white and Latin samples. The two ethnic groups are approximately equal on the criterion measures and differ only on one predictor---Verbal Reasoning. Since none of the predictors are valid for the Latin sample, any significant correlation in the white sample (except Verbal Reasoning) produces a Model 5.

Only one additional model appeared in the white-Latin comparisons. Model 7 was illustrated in the relationship between Verbal Reasoning and Ratings of Accuracy.

The criterion of whether the correlation between a test and criterion was significantly greater than zero in neither, both, or one of the subgroups was used to identify the above mentioned models. Applying the additional criterion of a significant difference between validity coefficients for the two racial groups (this criterion applies only to Models 5 through 10) only four models emerge. Two Model 7 cases meet this additional criteria, namely, the correlations between Verbal Reasoning and Clerical I (10 minutes) and ratings of Human

Relations in the Negro-white comparisons. Two models in the Latin-white comparisons satisfy this additional criterion--the relationships between Clerical I (both ten minute forms) and ratings of Learning Ability. The superscript a in Tables 35 and 36 indicates those models which meet this additional criterion.

Table 37 presents the results of the regression tests for the analysis of covariance (Potthoff, 1966). This analysis simultaneously tests the hypothesis that the regression slopes and intercepts are equal for the three ethnic groups. Inspection of Table 37 reveals that using this method of analysis only two relationships demonstrated bias as indicated by the significant F_1 ratio. Both forms of the ten minute clerical test were biased in predicting ratings of Learning Ability. The significant F_2 ratio indicated that a common regression slope could not be used with the three ethnic groups.

Table 37: Analysis of Covariance for Homogeneity of Regression

Merchandise Handlers II (1)

Predictor	Accuracy			Accuracy Under Pressure			Work Speed			Learning Ability			Human Relations			Overall Effectiveness		
	$F_1^{(2)}$	$F_2^{(3)}$	$F_3^{(4)}$	F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3
Verbal Reasoning	.05	.01	.10	1.13	1.58	.68	.78	.15	1.42	1.09	.69	1.49	1.36	2.68	.03	1.09	1.10	1.09
Arithmetic Reasoning	.63	1.08	.17	.34	.34	.34	.45	.31	.59	.78	.58	.99	.43	.64	.21	.50	.52	.48
Clerical I 5 minutes	.56	1.00	.11	.23	.19	.27	.53	.18	.89	2.07	2.86	1.25	.75	1.13	.37	1.07	1.60	.54
Clerical I 10 minutes	1.04	1.84	.23	.31	.24	.39	.50	.39	.61	3.29*	5.75**	.81	1.28	1.93	.62	.88	1.40	.36
Clerical II 5 minutes	.37	.64	.11	.36	.23	.49	1.08	.37	1.79	2.01	1.90	2.10	.37	.69	.05	.89	.68	1.09
Clerical II 10 minutes	.78	1.52	.05	.61	.78	.45	1.05	.58	1.52	1.69	1.59	1.79	.06	.01	.11	.98	1.05	.90
Clerical I (R-W) 5 minutes	.40	.75	.06	.26	.22	.30	.70	.35	1.06	1.88	2.37	1.36	.46	.73	.19	.98	1.23	.73
Clerical I (R-W) 10 minutes	.95	1.74	.16	.40	.43	.37	.70	.65	.76	3.06*	5.22*	.88	.97	1.58	.35	.82	1.15	.49
Clerical II (R-W) 5 minutes	.19	.35	.03	.38	.27	.49	.99	.56	1.42	1.65	1.72	1.57	.30	.50	.10	.61	.32	.91
Clerical II (R-W) 10 minutes	.17	.32	.02	.30	.18	.42	.77	.32	1.22	1.31	1.20	1.43	.14	.13	.15	.60	.46	.73

(1) Degrees of freedom for all comparisons: F_1 - (4,258); F_2 - (2,258); F_3 - (2,260)

(2) F_1 tests hypothesis that $E(Y_{ij}|X_{ij}) = a + bX_{ij}$ for all i groups.

(3) F_2 tests hypothesis that $E(Y_{ij}|X_{ij}) = a_i - bX_{ij}$ for all i groups.

(4) F_3 tests hypothesis that $E(Y_{ij}|X_{ij}) = a_i + b_iX_{ij}$ for all i groups.

*p<.05

**p<.01

Background Data - Clerical I

Table 38 presents the biographical data for the Clerical I job classification sample. Eight employees of Mexican-American extraction were included in the original sample. A separate subgroup analysis was not performed on this ethnic group since it was too small to make reliable comparisons. White employees in this sample were older and had longer company service than their Negro counterparts. The educational level of the Negro sample, however, was approximately two years above that of the white sample.

Table 38: Biographical Data-Clerical I

	Group	$\bar{X}$	s	N	t(1)
Age	Total	35.19	13.57	129	
	White	37.77	13.84	99	
	Negro	28.00	7.63	22	4.50**
Tenure (Years)	Total	2.90	1.06	129	
	White	3.04	1.09	99	
	Negro	2.41	.80	22	2.52*
Education (Years)	Total	10.67	1.87	129	
	White	10.31	1.69	99	
	Negro	12.55	1.60	22	5.63**

*p<.05

**p<.01

(1)t ratios are between the means of the white and Negro samples

Predictor Comparisons

Mean predictor scores for the two racial groups are presented in Table 39. There were no significant differences between the mean performance of the two racial groups on any of the predictors.

Criterion Comparisons

Table 40 presents the mean criterion scores for the white and Negro samples. Like the predictor scores, there were no differences between the two samples on any of the mean criterion scores.

Validity

The correlations between the predictors and criteria were rather disappointing as indicated by inspection of Table 41. In fact, out of 180 possible relationships, only 41 were significant at the .05 level. Furthermore, of the 60 white-Negro comparisons, in only one case was the correlation significant for both racial groups.

Despite a considerable differential in sample size, both racial groups appear equally predictable. The rating of Work Speed was the most predictable criterion for both racial groups.

Models Illustrated

Nineteen cases of Model 5 were illustrated in this sample. The number in parentheses below the correlation for the Negro sample in Table 41 indicates the model represented. Model 5 is illustrative of the situation where no significant mean differences exist between the two racial groups on either the predictor or criterion, but the test is valid for only one racial group. In eight out of the nineteen cases, the Negro group was the most predictable racial group.

The relationship between Verbal Reasoning and ratings of Work Speed and the relationship between Clerical I (10 minutes) and ratings of Overall Effectiveness were the only illustrations of Model 5 which remained when the additional criterion of a significant difference between validity coefficients was utilized. The superscript"a" in Table 41 indicates those models which meet this additional criterion. In both of these cases the validity coefficient

Table 39: Predictors-Means, Standard Deviations, N's, and Tests of Significance of Mean Differences

Clerical I

Predictor	Group	$\bar{X}$	s	N	t[1]
Verbal Reasoning	Total	22.03	9.73	129	
	White	21.85	9.69	99	
	Negro	25.27	9.49	22	1.49
Arithmetic Reasoning	Total	23.57	8.18	129	
	White	23.44	8.62	99	
	Negro	25.23	6.50	22	.91
Clerical I 5 minutes	Total	48.26	11.46	129	
	White	48.05	10.66	99	
	Negro	49.27	15.12	22	.44
Clerical I 10 minutes	Total	100.46	20.37	129	
	White	99.59	19.40	99	
	Negro	105.91	24.38	22	1.30
Clerical II 5 minutes	Total	57.80	11.29	129	
	White	57.91	10.57	99	
	Negro	57.91	15.13	22	.00
Clerical II 10 minutes	Total	111.88	19.63	129	
	White	111.90	18.27	99	
	Negro	113.27	26.64	22	.29
Clerical I (R-W) 5 minutes	Total	41.63	13.17	129	
	White	41.29	12.09	99	
	Negro	42.68	17.71	22	.44
Clerical I (R-W) 10 minutes	Total	89.91	22.93	129	
	White	89.22	21.57	99	
	Negro	97.00	25.96	22	1.65
Clerical II (R-W) 5 minutes	Total	51.16	12.56	129	
	White	51.39	12.10	99	
	Negro	50.73	14.91	22	.22
Clerical II (R-W) 10 minutes	Total	99.12	22.45	129	
	White	99.46	21.88	99	
	Negro	98.55	26.62	22	.17

(1) t ratios are between the means of the white and Negro samples.

Table 40: Criteria-Means, Standard Deviations, N's,
and Tests of Significance of Mean Differences

Clerical I

Criterion	Group	$\bar{X}$	σ	N	t (1)
Accuracy	Total	3.95	.98	129	
	White	3.89	.99	99	
	Negro	4.00	.87	22	.48
Accuracy	Total	3.90	.95	129	
Under	White	3.86	.91	99	
Pressure	Negro	3.86	1.04	22	.00
Work	Total	3.95	1.03	129	
Speed	White	3.89	1.06	99	
	Negro	4.18	.96	22	1.17
Learning	Total	3.97	.98	129	
Ability	White	3.86	.94	99	
	Negro	4.14	1.04	22	1.23
Human	Total	3.95	.92	129	
Relations	White	3.96	.96	199	
	Negro	3.91	.81	22	.23
Overall	Total	4.06	1.00	129	
Effectiveness	White	4.02	1.03	99	
	Negro	4.05	.84	22	.13

(1) t ratios are between the means of the white and Negro samples.

Table 41: Predictor-Criterion Correlations-Clerical I

Total Group, Whites and Negroes[1,2]

Criterion		Predictor: Verbal Reasoning	Arithmetic Reasoning	Clerical I 5 minutes	Clerical I 10 minutes	Clerical II 5 minutes	Clerical II 10 minutes	Clerical I (R-W 5 min)	Clerical I (R-W 10 min)	Clerical II (R-W 5 min)	Clerical II (R-W 10 min)	N
Accuracy	Total	05	18*	12	13	08	05	15	13	08	09	129
	White	07	23*	14	12	10	05	19	18	10	10	99
	Negro	13	05 (5)	13	27	10	11	06	17	04	06	22
Accuracy Under Pressure	Total	-04	12	24**	22*	12	18*	21*	16	11	16	129
	White	-05	15	22*	17	14	17	18*	14	10	15	99
	Negro	07	06	31 (5)	43* (5)	18	23	27 (5)	38	16	20	22
Work Speed	Total	07	26**	27**	24**	21*	26**	26**	20*	18*	24**	129
	White	-02a	25*	26*	17	21*	22*	24*	16	18	21*	99
	Negro	45* (5)	39 (5)	11 (5)	54** (5)	26 (5)	-06*	37 (5)	49* (5)	22	40 (5)	22
Learning Ability	Total	05	20*	20*	21*	14	17*	19*	16	07	12	129
	White	-03	25*	21*	19	17	20	19	17	09	15	99
	Negro	37	25 (5)	39 (5)	51* (5)	25	26	31	43* (5)	21	20	22
Human Relations	Total	-02	06	19*	17	16	10	16	16	13	10	129
	White	-03	10	24*	19	18	13	17	15	14	13	99
	Negro	11	-03	12 (5)	19	15	-02	13	17	03	-09	22
Overall Effectiveness	Total	02	11	17	13	10	11	17	11	10	14	129
	White	-02	14	14	06a	06	06	13	06	06	09	99
	Negro	33	09	37	51* (5)	28	30	32	44* (5)	27	28	22

(1) Decimals are omitted.

(2) Number in parentheses below the correlation for the Negro sample indicates the model illustrated (See Appendix A).

*p<.05

**p<.01

a indicates those models in which a significant difference exists between the validity coefficients for the two ethnic groups.

was not significant for the total group. Thus, traditional validation procedures, using only total group analysis, would result in the elimination of potentially valid predictors.

Table 42 presents the results of the regression tests of the analysis of covariance (Potthoff, 1966). All of the F ratios were not significant indicating that no bias was present. It should be noted that using this method the two relationships mentioned above which met the additional criterion of a significant difference between validity coefficients fail to demonstrate bias.

Table 42: Analysis of Covariance for Homogeneity of Regression

Clerical I[1]

Predictor	Criterion: Accuracy			Criterion: Accuracy Under Pressure			Criterion: Work Speed			Criterion: Learning Ability			Criterion: Human Relations			Criterion: Overall Effectiveness		
	F_1[2]	F_2[3]	F_3[4]	F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3
Verbal Reasoning	.08	.04	.13	.15	.30	.01	2.26	3.29	1.21	1.78	2.39	1.16	.17	.30	.05	.75	1.50	.01
Arithmetic Reasoning	.21	.34	.09	.02	.03	.01	.75	.55	.95	.60	.14	1.06	.16	.22	.09	.01	.01	.01
Clerical I 5 minutes	.15	.11	.18	.02	.03	.01	.61	.01	1.22	.78	.25	1.33	.50	.90	.11	.08	.15	.01
Clerical I 10 minutes	.12	.14	.09	.64	1.21	.08	1.05	1.30	.80	1.29	1.74	.83	.17	.11	.23	.99	1.98	.01
Clerical II 5 minutes	.14	.04	.24	.01	.00	.01	.77	.07	1.48	.78	.02	1.56	.15	.25	.05	.15	.29	.01
Clerical II 10 minutes	.11	.01	.22	.01	.01	.00	.74	.16	1.33	.71	.01	1.42	.37	.67	.07	.20	.39	.01
Clerical I (R-W) 5 minutes	.47	.76	.18	.01	.02	.01	.61	.01	1.23	.69	.05	1.34	.19	.29	.09	.04	.07	.01
Clerical I (R-W) 10 minutes	.06	.08	.05	.56	1.05	.07	.93	1.08	.78	1.01	1.19	.82	.11	.01	.20	.76	1.51	.01
Clerical II (R-W) 5 minutes	.20	.15	.26	.02	.04	.01	.78	.01	1.57	.90	.23	1.59	.21	.38	.04	.18	.35	.02
Clerical II (R-W) 10 minutes	.17	.08	.25	.01	.03	.01	.91	.22	1.60	.82	.03	1.62	.51	.98	.05	.12	.22	.02

(1) Degrees of freedom for all comparisons: F_1 - (2,117); F_2 - (1,117); F_3 - (1,118)

(2) F_1 tests hypothesis that $E(Y_{ij}|X_{ij}) = a + bX_{ij}$ for all i groups.

(3) F_2 tests hypothesis that $E(Y_{ij}|X_{ij}) = a_i + bX_{ij}$ for all i groups.

(4) F_3 tests hypothesis that $E(Y_{ij}|X_{ij}) = a_i + b_iX_{ij}$ for all i groups.

Background Data - Machine Clerical I and II

Biographical data for the combined job classifications Machine Clerical I and II is presented below.

The original sample contained six employees of Mexican-American extraction. These subjects were not included in the subgroup analysis since the sample was too small to make reliable comparisons.

The trend which has occurred throughout study 6 was again demonstrated in this combined job classification. White employees were older and had more company service than their Negro counterparts. The mean educational level of the Negro employees, as reported on the application blank, was significantly higher than the educational level of the white employees.

Table 43: Biographical Data--Machine Clerical I and II

	Group	$\bar{X}$	s	N	t(1)
Age	Total	29.74	11.20	91	
	White	32.37	12.65	60	
	Negro	25.45	5.97	31	3.50**
Tenure (Years)	Total	2.75	1.25	91	
	White	3.25	1.24	60	
	Negro	1.84	.69	31	6.89**
Education (Years)	Total	11.65	1.17	91	
	White	11.37	1.22	60	
	Negro	12.10	.87	31	3.25**

** $p < .01$

(1) t ratio between the white and Negro sample.

Predictor Comparisons

Mean predictor scores for the two ethnic groups are presented in Table 44. White employees scored significantly higher than Negro employees on Clerical Test I. This difference occurred both on the 5 and 10 minute time limit as well as on both the corrected (guessing factor) and non-corrected scores.

No significant differences existed between the two racial groups on any of the other predictor measures.

Criterion Comparisons

As shown in Table 45, the mean performance ratings of the two ethnic groups were approximately equal on four out of the six criteria. White employees, however, had higher mean performance ratings on both Human Relations and Overall Effectiveness.

Because of the differential length of service for the two ethnic groups, correlations were computed between tenure and the various criteria. No significant correlations emerged from this analysis indicating that job experience was not a major factor contributing to the obtained criterion differences for the two ethnic groups.

Validity

Validity coefficients for the two racial groups are presented in Table 46. Inspection of the table reveals a distinct differential validity pattern for the two racial groups. In fact, the predictor correlated positively with the criterion for the white sample but correlated negatively for the Negro sample in a large number of the predictor-criterion relationships.

Examining specific predictors, we find that the Verbal Reasoning test did not predict any of the criteria for either racial group. Likewise, the Arithmetic Reasoning Test possessed little validity for either racial group. The clerical tests, on the other hand, predicted most criteria for both ethnic groups.

Table 44: Predictors- Means, Standard Deviations, N's and Tests of Significance of Mean Differences

Machine Clerical I and II

Predictor	Group	$\bar{X}$	s	N	t(1)
Verbal	Total	24.45	9.75	97	
Reasoning	White	25.23	10.41	60	
	Negro	24.13	8.90	31	.50
Arithmetic	Total	26.97	7.86	97	
Reasoning	White	27.82	8.82	60	
	Negro	25.26	6.12	31	1.60
Clerical I	Total	51.86	13.06	97	
5 minutes	White	54.18	14.08	60	
	Negro	46.52	10.17	31	2.66*
Clerical I	Total	109.71	24.51	97	
10 minutes	White	114.32	26.14	60	
	Negro	99.13	19.08	31	2.83*
Clerical II	Total	60.34	14.37	97	
5 minutes	White	61.67	13.67	60	
	Negro	59.87	14.61	31	.57
Clerical II	Total	119.02	25.43	97	
10 minutes	White	122.75	25.11	60	
	Negro	116.55	22.29	31	1.15
Clerical I	Total	46.07	14.19	97	
(R-W)	White	48.15	14.69	60	
5 minutes	Negro	41.26	12.97	31	2.18*
Clerical I	Total	100.38	26.59	97	
(R-W)	White	104.67	27.96	60	
10 minutes	Negro	89.94	22.38	31	2.51*
Clerical II	Total	54.96	14.63	97	
(R-W)	White	56.37	14.16	60	
5 minutes	Negro	53.84	14.83	31	.79
Clerical II	Total	108.25	25.84	97	
(R-W)	White	111.93	26.11	60	
10 minutes	Negro	104.81	22.46	31	1.28

*p<.05

( ). ratios are between the means of the white and Negro samples.

Table 48: Criteria - Means, Standard Deviations, N's, and Tests of Significance of Mean Differences

Machine Clerical I and II

	Group	$\bar{X}$	s	N	t(1)
Accuracy	Total	4.46	1.39	97	
	White	4.65	1.49	60	
	Negro	4.19	1.25	31	1.46
Accuracy	Total	4.21	1.28	97	
Under	White	4.40	1.30	60	
Pressure	Negro	3.90	1.27	31	1.73
Work	Total	3.99	1.24	97	
Speed	White	4.08	1.32	60	
	Negro	3.81	1.08	31	.97
Learning	Total	4.28	1.22	97	
Ability	White	4.43	1.27	60	
	Negro	4.10	1.14	31	1.20
Human	Total	4.59	1.28	97	
Relations	White	4.83	1.30	60	
	Negro	4.23	1.23	31	2.10*
Overall	Total	4.59	1.41	97	
Effectiveness	White	4.87	1.41	60	
	Negro	4.13	1.38	31	2.36*

*p<.05

(1) t ratios are between the means of the white and Negro samples.

Table 46: Predictor-Criterion Correlations-Machine Clerical I and II

Total Group, Whites, and Negroes[1,2]

Criterion		Predictor: Verbal Reasoning	Arithmetic Reasoning	Clerical I 5 minutes	Clerical I 10 minutes	Clerical II 5 minutes	Clerical II 10 minutes	Clerical I (R-W 5 min)	Clerical I (R-W 10 min)	Clerical II (R-W 5 min)	Clerical II (R-W 10 min)	N
Accuracy	Total	09	21*	13	16	-03	05	15	17	02	11	97
	White	07	27*	23	22*a	06a	11	27*a	29*a	10	18	60
	Negro	07	-01	-32	-32	-40*	-31	-27	-26	-30	-26	31
			(5)		(7)	(5)		(7)	(7)			
Accuracy Under Pressure	Total	04	22*	17	20*	01	10	17	19	04	13	97
	White	01	32*	27*a	33**a	16a	21a	31*a	33**a	17a	25	60
	Negro	05	-11	-27	-30	-43*	-37*	-26	-28	-36*	-33	31
			(5)	(7)	(7)	(5)	(5)	(7)	(7)	(5)		
Work Speed	Total	01	18	21*	24*	12	18	21*	24*	16	24*	97
	White	-02	19	30*a	35**a	24	26*	35**a	36**a	24	31*	60
	Negro	04	11	-21	-25	-26	-17	-29	-24	-18	-10	31
				(7)	(7)		(5)	(7)	(7)		(5)	
Learning Ability	Total	05	13	15	18	04	16	15	17	07	19	97
	White	-03	17	27*a	32*a	15a	25a	31a	32**a	14	28*a	60
	Negro	11	-03	-37*	-37*	-43*	-36*	-36*	-35	-34	-29	31
				(10)	(10)	(5)	(5)	(10)	(7)		(5)	
Human Relations	Total	03	-01	02	05	-05	05	-06	-01	-09	03	97
	White	-05	07a	12a	18a	06a	16a	06a	14a	04a	18a	60
	Negro	-06	-40*	-43*	-50**	-43*	-38*	-46**	-51**	-47**	-48**	31
			(6)	(8)	(8)	(6)	(6)	(8)	(8)	(6)	(6)	
Overall Effectiveness	Total	01	09	12	13	12	-01	09	10	-08	04	97
	White	-07	13	18	21	06a	05a	19	20	04a	11a	60
	Negro	09	-16	-29	-35	-15*	-39*	-31	-35	-39*	-36*	31
						(6)	(6)			(6)	(6)	

(1) Decimals are omitted.

(2) Number in parentheses below the correlation for the Negro sample indicates the model illustrated (See Appendix A).

* $p<.05$

** $p<.01$

a Indicates those models in which a significant difference exists between the validity coefficients for the two ethnic groups.

Models Illustrated

Five different models were represented in this sample. The specific model illustrated is again represented by the number in parentheses below the correlation for the Negro sample in Table 46.

Illustrated in eleven cases, Model 5 represents the situation where a test has validity for one group, none for the other, yet mean performance on both the criterion and predictor is not significantly different. In a number of these situations the test correlates positively with the criteria for one racial group and negatively for the other which tends to eliminate the validity of the test based on the total sample. Inspecting the specific illustrations of Model 5, we find that both racial groups appear equally predictable. That is, in approximately one-half the cases the test is a valid predictor for the Negro sample.

Model 6, as illustrated in the correlation between the Clerical II test and ratings of Human Relations and Overall Effectiveness, occurred nine times. It is interesting to note that in all cases of Model 6, the test correlated significantly with the criteria for Negro employees but not for white employees.

Model 7 was the most frequently occurring model in this sample. Twelve cases were represented in the relationships between all versions of the Clerical I test and the various criteria. In all illustrations of this model, the test possessed validity only for the white sample. Using this test as a selection instrument would result in the elimination of Negro subjects whose probability of job success is equal to those of the white subjects selected since the Negroes score lower on the predictor.

Four cases of Model 8 were illustrated. White employees scored higher on the Clerical I test and also on the criterion of Human Relations. However, the test is a valid predictor only for the Negro sample. It is somewhat ironical that even though the test is valid for the Negro sample, the probability of a Negro being selected is lower than the probability of a white individual since the Negro group scores lower on the predictor. This situation reinforces the need, not only for subgroup validation, but also for a comparison of validity coefficients as well as mean differences for the two racial groups.

The relationship between ratings of Learning Ability and performance on Clerical I test illustrates Model 10. Although there was no difference on the criterion between the two racial groups, white employees obtained higher scores on the Clerical I test. Because the test correlated in opposite directions for the two racial groups, combining them results in no validity. Either differential or non-linear prediction is required to yield valid predictions.

The criteria used for identifying the above models was whether the correlation between the test and criterion was significantly greater than zero in neither, both, or one of the subgroups. If the additional criterion of a significant difference between validity coefficients for the two racial groups is imposed (this applies only to Models 5 through 10) the frequency of the various models illustrated changes only slightly. Only seven illustrations of model 5 are represented using this somewhat more restrictive criterion, while the frequency of the other models ramains unchanged. The superscript a in Table 46 indicates those models which meet this additional criterion.

Table 47 presents the results of the regression tests for the analysis of covariance (Potthoff, 1966). The F_2 ratio of this analysis conformed with the more restrictive definition of bias (i.e., no bias was demonstrated unless the validity coefficients for the two racial groups differed significantly). It should be noted that this analysis yielded significant F_2 ratios in cases where the validity coefficients for the two racial groups were not significant but there was a significant difference between the two validity coefficients.

Table 47: Analysis of Covariance for Homogeneity of Regression

Machine Clerical I and II[1]

	Accuracy			Accuracy Under Pressure			Work Speed			Learning Ability			Human Relations			Overall Effectiveness		
	F_1[2]	F_2[3]	F_3[4]	F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3
Verbal Reasoning	.99	.01	2.00	1.47	.04	2.94	.52	.05	1.00	.93	.36	1.51	2.31	.01	4.67	3.03	.44	5.64*
Arithmetic Reasoning	1.19	1.01	1.36	2.45	2.79	2.07	.31	.04	.58	.80	.48	1.12	4.76*	4.56	4.77	3.31	1.44	5.15
Clerical I 5 minutes	3.33	5.30*	1.30	3.59	5.31*	1.79	2.34	4.42	.25	4.25*	7.62*	.81	5.50*	6.30*	4.44	4.54*	4.34	4.56
Clerical I 10 minutes	3.54	5.94*	1.09	4.56*	7.48*	1.53	3.11	6.05*	.17	4.82*	8.95**	.64	7.14**	9.77**	4.10	5.42*	6.17*	4.14
Clerical II 5 minutes	2.93	3.54	2.25	5.27*	7.26*	3.06	2.91	4.88	.91	4.00*	6.32*	1.58	4.29*	4.53	4.85	4.69*	2.90	6.35*
Clerical II 10 minutes	2.65	3.16	2.08	4.94*	6.89*	2.80	2.05	3.39	.70	4.36*	7.33*	1.30	5.29*	5.85*	4.84	5.19*	4.11	6.05*
Clerical I (R-W) 5 minutes	3.46	5.52*	1.34	4.23*	6.39*	1.96	4.13*	7.86*	.37	5.07*	9.13**	.92	5.53*	5.50*	5.28	5.03*	4.85	4.99
Clerical I (R-W) 10 minutes	3.26	5.37*	1.09	4.56*	7.29*	1.71	3.42	6.61*	.22	4.70*	8.59**	.74	6.99**	8.86**	4.69	5.61*	6.17*	4.77
Clerical II (R-W) 5 minutes	2.49	2.79	2.14	4.40*	5.65*	2.98	2.19	3.52	.84	3.00	4.41	1.54	5.19*	5.10	5.05	4.37*	2.33	6.32*
Clerical II (R-W) 10 minutes	2.63	3.35	1.85	4.77*	6.76*	2.64	1.86	3.15	.56	3.74	6.23*	1.18	6.92**	8.88**	4.56	5.23*	4.52	5.72*

(1) Degrees of freedom for all comparisons: F_1 - (2,87); F_2 - (1,87); F_3 - (1,88)

(2) F_1 tests hypothesis that $E(Y_{ij}|X_{ij}) = a + bX_{ij}$ for all i groups.

(3) F_2 tests hypothesis that $E(Y_{ij}|X_{ij}) = a_i + bX_{ij}$ for all i groups.

(4) F_3 tests hypothesis that $E(Y_{ij}|X_{ij}) = a + b_iX_{ij}$ for all i groups.

* $p<.05$

** $p<.01$

Biographical Data - Miscellaneous Clerical

Table 48 presents biographical data on the remaining clerical positions. White employees in this sample are older and have longer company service than Negro employees. The Negro employees' educational level is approximately a year higher than the educational level of the white employees.

Table 48:Biographical Data
Miscellaneous Clerical

	Group	$\bar{X}$	s	N	t(1)
Age	Total	31.87	12.44	130	
	White	33.13	13.02	106	
	Negro	26.29	7.32	24	3.44**
Tenure (Years)	Total	2.91	1.06	130	
	White	3.04	1.05	106	
	Negro	2.29	.91	24	3.26**
Education (Years)	Total	11.44	1.38	129	
	White	11.27	1.44	105	
	Negro	12.21	.72	24	4.57**

**p < .01

(1) t ratios are between the means of the white and Negro samples

Predictor Comparisons

Table 49 presents the mean predictor data for the total, white and Negro samples. Mean predictor performance for the two racial groups is approximately equal across all predictors.

Criterion Comparisons

Mean criterion scores for the two racial groups are presented in Table 50. Like the predictor comparisons no significant mean differences were found between the two racial groups on any of the criterion measures.

Table 49: Predictors-Means, Standard Deviations, N's and Tests of Significance of Mean Differences

Miscellaneous Clerical

	Group	$\bar{X}$	s	N	t(1)
Verbal Reasoning	Total	25.08	9.04	130	
	White	25.45	9.26	106	
	Negro	23.42	7.98	24	.99
Arithmetic Reasoning	Total	25.98	7.90	130	
	White	26.04	8.26	106	
	Negro	25.71	6.21	24	.18
Clerical I 5 minutes	Total	53.68	13.35	130	
	White	54.46	14.11	106	
	Negro	50.25	8.74	24	1.83
Clerical I 10 minutes	Total	111.14	22.71	130	
	White	112.48	23.89	106	
	Negro	105.21	15.49	24	1.83
Clerical II 5 minutes	Total	62.10	13.28	130	
	White	62.57	13.74	106	
	Negro	60.04	11.00	24	.84
Clerical II 10 minutes	Total	121.42	22.81	130	
	White	121.58	22.82	106	
	Negro	120.67	23.21	24	.17
Clerical I (R-W) 5 minutes	Total	48.10	13.86	130	
	White	48.74	14.74	106	
	Negro	45.29	8.68	24	1.49
Clerical I (R-W) 10 minutes	Total	102.18	24.04	130	
	White	103.27	25.51	106	
	Negro	97.33	15.56	24	1.45
Clerical II (R-W) 5 minutes	Total	55.94	13.97	130	
	White	56.26	14.56	106	
	Negro	54.54	11.15	24	.54
Clerical II (R-W) 10 minutes	Total	109.48	24.75	130	
	White	109.45	25.16	106	
	Negro	109.58	23.23	24	.02

(1) t ratios are between the means of the white and Negro samples

Table 50: Criteria- Means, Standard Deviations, N's and Tests of Significance of Mean Differences

Miscellaneous Clerical

	Group	$\bar{X}$	s	N	t[1]
Accuracy	Total	4.31	.91	130	
	White	4.25	.91	106	
	Negro	4.54	.88	24	1.41
Accuracy Under Pressure	Total	4.13	.99	130	
	White	4.06	.99	106	
	Negro	4.46	.93	24	1.21
Work Speed	Total	4.16	1.04	130	
	White	4.19	.99	106	
	Negro	4.04	1.27	24	.63
Learning Ability	Total	4.12	.82	130	
	White	4.12	.79	106	
	Negro	4.08	.97	24	.21
Human Relations	Total	4.27	.93	130	
	White	4.26	.91	106	
	Negro	4.29	1.04	24	.14
Overall Effectiveness	Total	4.28	.90	130	
	White	4.23	.88	106	
	Negro	4.54	.98	24	1.52

[1] t ratios are between the means of the white and Negro samples

Because the rating criteria were confounded with differential tenure for the two racial groups, correlations between tenure and ratings were computed. The nonsignificant correlations obtained indicated that the mean criterion ratings for the Negro sample would not have increased substantially if they had been on the job as long as the white sample.

Validity

Validity coefficients for the total, white and Negro samples are presented in Table 51. The most striking characteristic of this analysis was the general lack of validity exhibited by the predictors for either racial group. The heterogeneity of job classifications included in this sample may have been a major factor contributing to this general lack of validity.

Only two predictors show validity for the racial subgroups. The Arithmetic Reasoning Test predicted ratings of Accuracy and Learning Ability for Negro employees but not for white employees. Clerical Test II, on the other hand, predicted ratings of Accuracy and Work Speed for white employees, but not for Negro employees.

Models Illustrated

Model 5, the only Model illustrated in this sample, was represented five times. The number in parentheses below the correlation for the Negro sample in Table 51 indicates the relationship represented

Although the performances of the white and Negro samples were approximately equal on all the predictor and criterion measures, the Arithmetic Reasoning test was a valid predictor of ratings of Accuracy and Learning Ability for the Negro sample only. Thus, the test may be used with some confidence to select Negroes but is not appropriate for the selection of white employees. In contrast, Clerical TI (R-W, 10 min.) predicted ratings of Accuracy and Work Speed for the white sample but not the Negro sample. Likewise, Clerical II (10 min.) is a valid predictor of Work Speed for the white sample only.

Applying the additional criterion of a significant difference between validity coefficients for the two racial groups eliminated these five examples of Model 5.

Because only a few validity coefficients reached a statistically significant level, caution should be exercised in the interpretation of this study.

Table 51: Predictor-Criterion Correlations-Miscellaneous Clerical[1,2]

Total Group, Whites, and Negroes[1,2]

Criterion		Predictor: Verbal Reasoning	Arithmetic Reasoning	Clerical I 5 minutes	Clerical I 10 minutes	Clerical II 5 minutes	Clerical II 10 minutes	Clerical I (R-W 5 min)	Clerical I (R-W 10 min)	Clerical II (R-W 5 min)	Clerical II (R-W 10 min)	N
Accuracy	Total	12	16	08	06	10	16	09	09	11	16	130
	White	15	12	09	07	10	18	10	09	13	22*	106
	Negro	07	46* (5)	11	18	15	09	10	18	-31	-08 (5)	24
Accuracy Under Pressure	Total	02	14	01	01	08	13	-01	02	07	13	130
	White	05	14	02	01	08	14	01	03	09	16	106
	Negro	04	21	03	11	13	13	-02	05	02	-01	24
Work Speed	Total	12	10	17*	17*	12	20*	15	16	10	18*	130
	White	12	04	17	17	15	23*	16	17	14	23*	106
	Negro	07	38	19	20	06	09 (5)	08	18	06	-01 (5)	24
Learning Ability	Total	13	22*	07	08	14	22*	06	07	08	17*	130
	White	15	17	05	03	10	18	04	04	07	18	106
	Negro	06	51** (5)	20	31	33	37	13	21	12	17	24
Human Relations	Total	03	13	11	11	16	20*	08	10	14	16	130
	White	02	10	11	10	14	17	10	11	13	-15	106
	Negro	30	31	15	18	33	33	01	08	18	21	24
Overall Effectiveness	Total	-01	10	01	-02	02	09	01	01	-01	05	130
	White	15	06	04	-01	04	11	04	02	02	08	106
	Negro	-07	32	-05	-01	-01	03	-05	01	-08	-08	24

(1) Decimals are omitted.

(2) Number in parentheses below the correlation for the Negro sample indicates the model illustrated (See Appendix A).

* $p<.05$

** $p<.01$

A rather large number of correlations were examined and some statistically significant coefficients would be expected by chance.

Table 52 presents the results of the regression tests of the analysis of covariance (Potthoff, 1966). All of the F ratios were not significant indicating that no bias was present.

Table 52: Analysis of Covariance for Homogeneity of Regression

Miscellaneous Clerical[1]

	Accuracy			Accuracy Under Pressure			Work Speed			Learning Ability			Human Relations			Overall Effectiveness		
	F_1[2]	F_2[3]	F_3[4]	F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3
Verbal Reasoning	1.22	.07	2.39	1.73	.14	3.35	.13	.01	.26	.02	.04	.01	1.29	2.54	.03	1.26	.15	2.39
Arithmetic Reasoning	2.41	2.68	2.11	1.79	.19	3.41	2.11	3.85	.37	2.54	5.04	.03	.81	1.60	.02	2.23	1.94	2.49
Clerical I 5 minutes	1.18	.05	2.33	1.64	.01	3.31	.25	.36	.15	.47	.93	.01	.15	.21	.08	1.29	.11	2.48
Clerical I 10 minutes	1.30	.35	2.27	1.74	.18	3.32	.28	.41	.15	1.29	2.58	.01	.24	.39	.08	1.17	.00	2.36
Clerical II 5 minutes	1.16	.10	2.24	1.79	.07	3.53	.15	.02	.28	1.00	2.00	.01	.76	1.45	.07	1.25	.02	2.49
Clerical II 10 minutes	1.14	.19	2.10	1.69	.01	3.40	.32	.28	.36	.71	1.38	.03	.40	.77	.03	1.28	.11	2.48
Clerical I (R-W) 5 minutes	1.15	.04	2.29	1.63	.02	3.27	.01	.00	.22	.17	.32	.02	.04	.03	.05	1.29	.12	2.48
Clerical I (R-W) 10 minutes	1.28	.29	2.28	1.67	.02	3.34	.24	.28	.20	.54	1.06	.02	.04	.01	.06	1.22	.01	2.45
Clerical II (R-W) 5 minutes	1.19	.23	2.16	1.73	.06	3.42	.45	.37	.32	.10	.17	.03	.14	.24	.04	1.29	.20	2.40
Clerical II (R-W) 10 minutes	1.77	1.52	2.01	1.85	.43	3.29	.63	.86	.40	.05	.04	.05	.10	.19	.02	1.46	.52	2.41

(1) Degrees of freedom for all comparisons: F_1 = (2,126); F_2 = (1,126); F_3 = (1,127)

(2) F_1 tests hypothesis that $E(Y_{ij}|X_{ij}) = a + bX_{ij}$ for all i groups.

(3) F_2 tests hypothesis that $E(Y_{ij}|X_{ij}) = a_i + bX_{ij}$ for all i groups.

(4) F_3 tests hypothesis that $E(Y_{ij}|X_{ij}) = a_i + b_iX_{ij}$ for all i groups.

Study 7: Keypunch Operators

Sample

Study 7 consists of 135 keypunch operators of whom 107 were white and 28 were Negro. As shown in Table 53, the two ethni. groups were approximately equal in terms of age, but white employees had longer company service. Again, Negro employees have a significantly higher educational level as compared to their white counterparts.

Table 53: Biographical Data-Keypunch Operators

	Group	$\bar{X}$	s	N	t(1)
Age	Total	26.12	8.97	135	
	White	26.26	9.74	107	
	Negro	25.57	5.13	28	.50
Tenure (Months)	Total	24.03	26.19	135	
	White	25.54	28.79	107	
	Negro	18.25	10.36	28	2.12*
Education (in years)	Total	11.82	1.09	135	
	White	11.75	1.18	107	
	Negro	12.11	.57	28	2.27*

(1) t ratios are between means of the white and Negro samples.
*p < .05

Predictor Comparisons

Scores on four predictor measures were obtained. The first was a company developed test of mental alertness. Using this measure two subscores were obtained--a verbal and a quantitative score. The sum of these two scores provided a measure of general mental alertness.

Secondly, the Thurstone Temperament Schedule was administered. This personality inventory is designed to measure the following seven aspects of temperament:

Active (A)	Emotionally Stable (Es)
Vigorous (V)	Sociable (S)
Impulsive (I)	Reflective (R)
Dominant (D)	

Table 54: Predictors-Means, Standard Deviations, N's, and Tests of Significance of Mean Differences-Keypunch Operators

Predictor	Group	$\bar{X}$	s	N	t[1]
Test of Mental Alertness					
Verbal	Total	34.10	13.68	128	
	White	34.60	14.37	104	
	Negro	31.96	10.11	24	.84
Quant.	Total	17.70	6 39	128	
	White	18.04	6.79	104	
	Negro	16.21	4.02	24	1.71
Total	Total	51.77	17.40	128	
	White	52.63	18.45	104	
	Negro	48.17	11.44	24	1.49
Clerical	Total	115.77	20.32	128	
	White	116.41	20.04	104	
	Negro	113.00	21.74	24	.73
Clerical (R-W)	Total	107.19	21.55	128	
	White	108.49	20.43	104	
	Negro	101.54	25.51	24	1.42
Arithmetic	Total	26.72	7.28	128	
	White	27.15	7.39	104	
	Negro	24.83	6.60	24	1.40
Arithmetic (R-W)	Total	22.62	8.38	128	
	White	23.18	8.49	104	
	Negro	20.17	7.56	24	1.59

Table 54: Continued

Predictor	Group	$\bar{X}$	s	N	t[1]
Thurstone Temperament Schedule					
Active	Total	9.36	3.05	107	
	White	9.60	3.10	88	
	Negro	8.21	2.59	19	1.81
Vigorous	Total	7.74	3.20	107	
	White	7.67	3.39	88	
	Negro	8.11	2.16	19	.70
Impulsive	Total	11.12	3.27	107	
	White	11.25	3.33	88	
	Negro	10.53	3.06	19	.86
Dominant	Total	9.55	4.67	107	
	White	9.40	4.53	88	
	Negro	10.26	5.35	19	.72
Emotionally Stable	Total	11.53	3.31	107	
	White	11.50	3.36	88	
	Negro	11.68	3.15	19	.22
Sociable	Total	13.05	3.36	107	
	White	13.06	3.44	88	
	Negro	13.00	3.02	19	.07
Reflective	Total	6.99	3.13	107	
	White	7.07	3.37	88	
	Negro	6.63	1.67	19	.82

(1) t ratios are between the means of the White and Negro samples.

Two company developed tests were also included. The Clerical Aptitude Test is a measure of perceptual speed and accuracy while the Arithmetic Skills Test is a measure of the ability of the employee to check the accuracy of simple arithmetic problems. Two scoring procedures were used with the Arithmetic and Clerical Tests: (1) the number of correct responses, and (2) the number correct minus the number of incorrect responses.

Mean scores for the two racial groups are presented in Table 54. There were no significant differences between the performance of the two racial groups on any of the predictors.

Criterion Comparisons

Employees were rated by their immediate supervisor in committee with their department head on a company-developed rating scale. This nine-point rating scale covered the following dimensions: (1) Concentration, (2) Learning Ability, (3) Work Sharing, (4) Error Detection, (5) Social Interaction, and (6) Overall Effectiveness. Two objective criteria were also available: Keypunching Speed and Error Percentage.

The raw ratings were converted to standard scores within raters in all cases where sufficient numbers of people were rated by a pair of raters. This was an attempt to compensate for errors of leniency and central tendency. Mean criterion scores for the two racial groups are presented in Table 55. Considering the standardized criteria we find that white employees obtained higher ratings than Negro employees on Concentration. No significant differences existed between the two racial groups on any of the other standardized criteria.

White employees obtained higher ratings than Negro employees on two raw score ratings: Error Detection and Social Interaction. No mean score differences existed between the two racial groups on the standardized objective criteria.

Validity

Validity coefficients for the two racial groups are presented in Table 56. All correlations have been controlled for tenure when appropriate. In general, the coefficients were rather low. The most promising tests were the Arithmetic and Clerical tests developed by the firm's psychologists.

The most predictable criteria were ratings of Learning Ability and the

Table 55: Criteria-Means, Standard Deviations, N's, and Tests of Significance of Mean Differences-Keypunch Operators

Criterion	Group	$\bar{X}$	s	N	t[1]
Ratings - Standardized					
Concentration	Total	50.29	9.95	131	
	White	51.26	10.21	103	
	Negro	46.72	7.98	28	2.16*
Learning Ability	Total	49.95	9.70	131	
	White	50.17	9.63	103	
	Negro	49.18	10.08	28	.47
Work Sharing	Total	49.92	9.93	131	
	White	50.47	9.90	103	
	Negro	47.91	9.95	28	1.20
Error Detection	Total	50.03	9.95	131	
	White	50.66	10.09	103	
	Negro	47.71	9.23	28	1.39
Social Interaction	Total	50.38	10.12	131	
	White	51.25	10.28	103	
	Negro	47.19	8.97	28	1.89
General Overall Effect	Total	49.98	9.96	131	
	White	50.39	10.33	103	
	Negro	48.50	8.49	28	.88
Ratings - Raw Score					
Concentration	Total	6.01	1.47	131	
	White	6.03	1.58	103	
	Negro	5.93	.98	28	.41
Learning Ability	Total	5.75	1.71	131	
	White	5.73	1.68	103	
	Negro	5.82	1.87	28	.25

Table 55: Continued

Criterion	Group	$\bar{X}$	s	N	t[1]
Work Sharing	Total	5.65	1.47	131	
	White	5.76	1.45	103	
	Negro	5.21	1.57	28	1.73
Error Detection	Total	5.62	1.79	131	
	White	5.84	1.69	103	
	Negro	4.82	1.96	28	2.70**
Social Interaction	Total	5.67	1.75	131	
	White	5.95	1.67	103	
	Negro	4.68	1.70	28	3.55**
General Overall Effectiveness	Total	5.53	1.57	131	
	White	5.64	1.62	103	
	Negro	5.14	1.35	28	1.48
Standardized Objective Criteria					
Speed	Total	49.99	9.93	100	
	White	49.79	10.13	79	
	Negro	50.77	9.33	21	.40
Error Percentage	Total	49.84	10.09	94	
	White	50.65	10.73	76	
	Negro	47.28	6.31	18	1.61

(1) t ratios are between the means of the White and Negro samples.

Table 56: Predictor-Criterion Correlations-Keypunch Operators
Total Group, Whites and Negroes[1,2]

Criterion		Predictor							
		Test of Mental Alertness							
Standardized Ratings		Verbal	Quant	Total	Clerical	Clerical (R-W)	Arith	Arith (R-W)	N
Concentration	Total	02	-04	01	-08	-06	03	08	125
	White	-07a	-10	-09	-11	-12	-05	01	101
	Negro	44* (6)	07	32	-20	-21	18	19	24
Learning Ability	Total	11	15	14	26**	25**	28**	26**	125
	White	06	20*	12	38**	38**a	30**	30**	101
	Negro	39	05 (5)	34	07 (5)	-25 (5)	30 (5)	21 (5)	24
Work Sharing	Total	01	11	05	03	02	04	05	125
	White	-02	11	03	11	12	02	07	101
	Negro	06	11	09	-01	07	11	-04	24
Error Detection	Total	08	02	07	16	16	24*	25**	125
	White	05	-01	04	14	14	21*	24*	101
	Negro	35	-01	29	-01	-09	28 (5)	19 (5)	24
Social Interaction	Total	03	-04	01	-08	-11	-04	02	125
	White	-03	-07	-05	-13	-17	-10	-03	101
	Negro	37	14	30	17	-01	25	27	24
General Overall Effectiveness	Total	05	02	05	12	09	14	14	25
	White	01	03	02	22*	19	12	13	101
	Negro	25	03	24	-16 (5)	10	28	20	24

Table 56: Continued

Criterion: Standardized Ratings		Predictor: Thurstone Temperament Schedule							
		Active	Vigorous	Impulsive	Dominant	Emotionally Stable	Sociable	Reflective	N
Concentration	Total	08	-05	-13	-11	10	-07	-06	105
	White	06	-02	-10	-07	07	-06	-04	86
	Negro	09	-29	-35	-28	29	-17	-26	19
Learning Ability	Total	20*	-03	08	-07	12	-04	-23*	105
	White	18	03	11	-13	05	-04	-24*	86
	Negro	37	-31	-19	-06	17	-23	-22	19
								(5)	
Work Sharing	Total	08	-03	05	07	26**	06	-02	105
	White	01	02	04	09	16a	09	-03	86
	Negro	33	-14	01	-11	67**	-09	21	19
						(5)			
Error Detection	Total	22*	04	-03	-14	11	-05	-20*	105
	White	23*	08	01	-13	11	-04	-19	86
	Negro	06	-35	-37	-25	-04	-29	-35	19
		(5)							
Social Interaction	Total	-04	-10	-21	-18	04	-2..*	10	105
	White	-07	-12	-18	-12	01	-23*	10	86
	Negro	05	04	-42	-43	26	-25	13	19
							(5)		
General Overall Effectiveness	Total	19	-04	02	-02	17	-06	-17	105
	White	18	02	04	-04	11	-04	-17	86
	Negro	26	-45	-36	-20	+8*	-37	-24	19
						(5)			

Table 56: Continued

Criterion		Predictor							
Raw Score Ratings		Test of Mental Alertness Verbal	Quant	Total	Clerical	Clerical (R-W)	Arith	Arith (R-W)	N
Concentration	Total	-05	-06	-0o	-04	-05	-04	04	125
	White	-08	-07	-08	-06	-07	-06	02	101
	Negro	36	02	32	-24	-25	09	10	24
Learning Ability	Total	09	19*	14	37**	36**	30**	30**	125
	White	06	22*	13	40**	40**	31**	32**	101
	Negro	37	0-	34	08	07	29	21	24
			(5)		(5)	(5)	(5)	(5)	
Work Sharing	Total	02	15	07	14	13	07	10	125
	White	01	15	06	14	15	04	11	101
	Negro	06	11	09	01	-09	12	-01	24
Error Detection	Total	09	08	10	28**	28**	26**	29**	125
	White	05	06	06	22*	22*	21*	27**	101
	Negro	33	-02	29	05	-01	32	25	24
					(6)	(6)	(6)	(6)	
Social Interaction	Total	04	-01	03	-04	-06	-02	05	125
	White	-02	-04	-03	-09	-12	-09	-01	101
	Negro	31	08	30	17	10	20	22	24
General Overall Effectiveness	Total	04	07	05	25**	23**	18*	19*	125
	White	01	05	03	24*	22*	14	15	101
	Negro	25	06	24	-14	-20	31	23	24
					(5)	(5)			

Table 56: Continued

Criterion		Predictor							
		Thurstone Temperament Schedule							
Raw Score Ratings		Active	Vigorous	Impulsive	Dominant	Emotionally Stable	Sociable	Reflective	N
Concentration	Total	05	-05	-12	-09	08	-06	-06	105
	White	07	-04	-08	-06	05	-06	-05	86
	Negro	10	-33	-36	-29	27	-18	-29	19
Learning Ability	Total	19	-03	09	-10	09	-03	-23*	105
	White	19	01	15	-13	05	-02	-24*	86
	Negro	37	-31	-18	-07	16	-24	-20 (5)	19
Work Sharing	Total	12	-01	01	-04	15	01	-01	105
	White	07	03	01	01	04a	03	-03	86
	Negro	34	-21	-06	-17	63** (5)	-14	12	19
Error Detection	Total	23*	02	01	-20*	03	-05	-19	105
	White	25*	05	07	-16	02	-02	-20	86
	Negro	05 (6)	-26	-29	-29	-04	-34	-22	19
Social Interaction	Total	02	-11	-21*	-27**	-05	-28**	10	105
	White	-03	-11	-20	-22*	-11	-29**	09	86
	Negro	05	04	-40	-43 (6)	24	-28 (6)	15	19
General Overall Effectiveness	Total	20	01	-03	-12	09	-06	-13	105
	White	20	03	01	-10	0-	-05	-1+	86
	Negro	26	-44	-32	-20	4-	-36	-21	19

Table 56: Continued

Criterion		Predictor							
		Test of Mental Alertness				Clerical		Arith	
		Verbal	Quant	Total	Clerical	(R-W)	Arith	(R-W)	N
Speed	Total	10	19	15	-13	-17	02	-03	97
	White	10	22	16	04	01	03	-04	77
	Negro	14	34	26	-42	-43	11	11	20
Error	Total	10	-02	07	-01	-01	07	07	91
Percent	White	06	01	05	08	10	08	09	74
	Negro	-12	-49* (5)	-31	-08	-02	-08	-11	17

Thurstone Temperament Schedule

		Active	Vigorous	Impulsive	Dominant	Emotionally Stable	Sociable	Reflective	N
Speed	Total	05	14	11	04	16	15	07	87
	White	09	20	13	07	07	17	13	62
	Negro	-47	24	-18	-29	49	17	-12	15
Error	Total	02	-08	-02	-06	-06	-12	-10	73
Percent	White	-04	-02	-08	-09	-12	-09	-10	60
	Negro	04	01	-10	-36	-11	-62 (5)	37	13

(1) Decimals are omitted.

(2) Number in parentheses below the correlation for the Negro sample indicates the model illustrated (See Appendix A).

*p<.05

**p<.01

a indicates those models in which a significant difference exists between the validity coefficients for the two ethnic groups.

least predictable criteria were the objective measures of Keypunching Speed and Error Percentage.

Models Illustrated

Only two different models were illustrated in this sample. The number in parentheses below the correlation for the Negro sample in Table 56 indicates the specific model illustrated. Model 5, the most frequently illustrated model, was represented in 24 predictor-criterion relationships. Model 5 illustrates the situation where there is no difference on either the predictor or criterion for the two racial subgroups and the test is valid only for one subgroup. The relationship between ratings of Work Sharing and the Emotional Stability scale of the Thurstone Temperament Schedule clearly illustrated this model.

The final model illustrated in this sample was model 6. Eight illustrations of this model occurred but it was most clearly illustrated in the relationship between the Clerical test and the raw score ratings of Error Detection. The mean test performance was approximately equal for the two groups on both forms of the Clerical test but the white employees were rated higher on Error Detection. The validity coefficient was significant only for the white sample. Total group validation procedures would recommend the use of the test for selection even though the test is clearly not appropriate for the Negro sample.

The frequency of the various models was greatly reduced when the additional criterion of a significant difference between validity coefficients was applied. Only four illustrations of Model 5 met this criterion. The superscript a in Table 56 indicates those models which met this criterion.

Table 57 presents the results of the regression tests of the analysis of covariance (Potthoff, 1966). A significant F_3 statistic was obtained in a large number of the comparisons of the predictors with the raw score ratings of Social Interaction. A significant F_3 statistic indicates that a common intercept value could not be used for the two ethnic groups. Only four significant F_2 statistics were illustrated. A significant F_2 statistic indicates that a common beta weight could not be used for the two ethnic groups.

Table 57. Analysis of Covariance for Homogeneity of Regression

Keypunch Operators Raw Score Ratings[1]

Predictor	Criterion: Concentration			Learning Ability			Work Sharing			Error Detection			Social Interaction			General Overall Effectiveness		
	F_1[2]	F_2[3]	F_3[4]	F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3
Test of Mental Alertness																		
Verbal	.89	1.77	.01	1.31	2.07	.55	1.13	.13	2.15	2.76	2.04	3.44	5.85**	2.36	9.23**	.95	.83	1.08
Quantative	.09	.15	.02	.38	.02	.74	.93	.05	1.82	1.78	.01	3.58	4.98*	.34	9.68**	.57	.09	1.06
Total	.84	1.66	.01	1.14	1.61	.67	1.07	.15	2.00	2.59	1.79	3.37	5.94**	2.48	9.29**	.95	.89	1.04
Clerical	.08	.15	.01	.44	.01	.88	1.01	.02	2.02	2.00	.63	3.3[illegible]	5.77**	1.80	9.66**	.48	.03	.94
Clerical (R-W)	.01	.01	.01	.57	.01	1.16	1.18	.56	1.81	1.59	.20	3.00	5.13*	.74	9.55**	.57	.44	.70
Arithmetic	.29	.58	.01	.85	.47	1.24	1.30	.59	2.02	2.26	1.74	2.76	5.64**	1.69	9.53**	.88	1.03	.73
Arithmetic (R-W)	.11	.21	.01	.71	.08	1.35	.94	.01	1.89	1.56	.63	2.05	5.09*	1.20	8.96**	.58	.50	.67

Table 57: Continued

Keypunch Operators — Raw Score Ratings

Predictor: Thurstone Temperament Schedule	Criterion: Concentration F_1 (2)	Concentration I_2 (3)	Concentration F_3 (4)	Learning Ability F_1	Learning Ability F_2	Learning Ability F_3	Work Sharing F_1	Work Sharing F_2	Work Sharing F_3	Error Detection F_1	Error Detection F_2	Error Detection F_3	Social Interaction F_1	Social Interaction F_2	Social Interaction F_3	General Overall Effectiveness F_1	General Overall Effectiveness F_2	General Overall Effectiveness F_3
Active	.04	.01	.07	.40	.29	.70	.89	1.32	.47	1.94	.47	3.42	2.55	.06	5.09	.32	.02	.62
Vigorous	.25	.47	.02	.40	.62	.19	.99	1.06	.92	2.84	.57	5.14	2.63	.18	5.12	1.12	.86	1.38
Impulsive	.28	.56	.01	.82	1.32	.32	.46	.03	.91	3.16	1.56	4.73	3.73	1.13	6.32*	.98	.67	1.30
Dominant	.09	.16	.03	.24	.24	.24	.65	.44	.86	2.29	.01	4.63	2.66	.64	4.71	.61	.02	1.22
Emotionally Stable	.40	.79	.01	.84	1.50	.17	4.69*	8.35**	.96	2.76	.49	5.05	3.65	2.03	5.23*	1.74	2.06	1.41
Sociable	.01	.01	.01	.10	.01	.19	.55	.24	.87	2.59	.22	5.01	2.83	.04	5.68*	.70	.05	1.36
Reflective	.16	.31	.01	09	.04	.13	.60	.37	.83	2.62	.01	5.30*	2.56	.14	5.02	.74	.01	1.43

Table 57: Continued

Keypunch Operators — Standardized Ratings

Predictor: Thurstone Temperament Schedule	Criterion: Concentration			Learning Ability			Work Sharing			Error Detection			Social Interaction			General Overall Effectiveness		
	F_1 (2)	F_2 (3)	F_3 (4)	F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3
Active	1.86	.16	3.58	.15	.27	.04	1.02	1.06	.98	.46	.47	.46	.50	.11	.39	.04	.01	.08
Vigorous	2.16	.33	4.01	.41	.77	.04	72	.34	1.12	1.05	.86	1.23	.43	.13	.73	.65	.89	.42
Impulsive	3.21	2.06	4.31	.56	1.18	.01	.76	.58	.95	1.09	1.08	1.11	1.07	1.01	1.14	.71	1.09	.35
Dominant	2.43	1.08	3.77	.14	.27	.01	.90	.64	1.16	.49	.01	.99	.78	.94	.62	.20	.03	.38
Emotionally Stable	4.16*	3.90	4.30	.62	1.20	.04	4.16*	6.95*	1.29	.62	.06	1.20	.74	.58	.81	.93	1.41	.45
Sociable	2.08	.07	-.13	.02	.01	.03	.70	.28	1.12	.60	.06	1.15	.46	.05	.86	.26	.12	.40
Reflective	3.11	1.91	4.07	.05	.02	.07	.77	.39	1.16	.67	.07	1.28	.34	.01	.68	.23	.00	.47

Table 57: Continued

Keypunch Operators Standardized Ratings

Predictor	Criterion																	
	Concentration			Learning Ability			Work Sharing			Error Detection			Social Interaction			General Overall Effectiveness		
Test of Mental Alertness	F_1 (2)	F_2 (3)	F_3 (4)	F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3
Verbal	3.95*	4.37	3.43	1.05	2.10	.01	.70	.04	1.36	.99	1.43	.55	2.52	2.67	2.34	.54	.84	.24
Quantative	2.98	2.10	3.82	.01	.01	.01	.64	.04	1.24	.37	.05	.69	1.70	.74	2.67	.17	.07	.27
Total	4.65*	5.63*	3.53	.84	1.67	.01	.68	.07	1.29	.98	1.40	.56	2.78	3.09	2.41	.55	.86	.24
Clerical	2.45	1.17	3.72	.04	.04	.03	.66	.01	1.34	.48	.41	.55	2.16	1.64	2.66	.12	.06	.19
Clerical (R-W)	2.00	.42	3.59	.08	.06	.10	.80	.40	1.22	.23	.05	.40	1.67	.68	2.66	.26	.42	.10
Arithmetic	3.25	3.20	3.23	.25	.36	.13	.74	.27	1.22	.42	.56	.28	2.17	1.77	2.55	.55	.99	.11
Arithmetic (R-W)	2.62	2.22	2.98	.11	.07	.15	.61	.07	1.16	.16	.10	.22	1.82	1.33	2.31	.26	.43	.09

(*) Degrees of freedom for comparisons using :
(a) Raw Score and Standardized Ratings : F_1-(2,121); F_2-(1,121); F_3-(1,122)
(b) Objective Criteria and Test of Mental Alertness, Clerical and Arithmetic Tests; F_1-(2,93); F_2-(1,93); F_3-(1,94)
(c) Speed and Thurstone Temperament Schedule: F_1-(2,73); F_2-(1,73); F_3-(1,74)
(d) Error Detection and Thurstone Temperament Schedule: F_1-(2,69); F_2-(1,69); F_3-(1,70)

Table 57: Continued

Keypunch Operators — Objective Criteria

Predictor: Test of Mental Alertness	Criterion: Speed F_1	Speed F_2	Speed F_3	Error Percent F_1	Error Percent F_2	Error Percent F_3
Verbal	.42	.08	.76	.76	.21	1.32
Quant.	.82	.69	.94	1.56	1.51	1.59
Total	.69	.44	.94	1.11	.79	1.43
Clerical	2.58	5.12	.06	2.46	.80	4.12
Clerical (R-W)	3.75	7.43*	.06	2.78	1.30	4.25
Arithmetic	.35	.70	.02	1.89	.02	3.81
Arithmetic (R-W)	1.42	2.81	.02	2.02	.46	3.59

Predictor: Thurstone Temperament Schedule	Criterion: Speed F_1	Speed F_2	Speed F_3	Error Percent F_1	Error Percent F_2	Error Percent F_3
Active	2.67	4.56	.74	2.07	.12	4.08
Vigorous	1.03	.92	1.14	2.39	.89	3.90
Impulsive	.96	1.29	.62	2.04	.02	4.12
Dominant	1.10	1.44	.76	1.93	.19	3.71
Emotionally Stable	1.27	1.60	.93	1.81	.04	3.62
Sociable	.51	.21	.80	1.96	.11	3.87
Reflective	.72	.68	.77	2.16	.45	3.91

(2) F_1 tests hypothesis that $E(Y_{ij} | X_{ij}) = a + bX_{ij}$ for all i groups.

(3) F_2 tests hypothesis that $E(Y_{ij} | X_{ij}) = a_i + bX_{ij}$ for all i groups.

(4) F_3 tests hypothesis that $E(Y_{ij} | X_{ij}) = a_i + b_i X_{ij}$ for all i groups.

Section IV: Summary and Discussion

Any attempt to summarize the data presented in the preceding seven studies is necessarily open to question and limited by the very nature of the data. Since a basic purpose of this research project was to obtain an estimate of the parameters of subcultural differences in the prediction of job performance, predictor-criterion relationships across studies were examined with respect to type of valid predictor, type of predictable criterion, and type of subgroup for which the predictor was valid. Several assumptions about the data were made before these comparisons were attempted. First, within each study each predictor-criterion relationship was treated as if it were independent of all other predictor-criterion relationships. Thus, the intercorrelations of the predictor set and the criterion set were ignored. Secondly, no attempt was made to weight the results of a study by the sample size of the study. This served to place the emphasis on the statistical significance of a result rather than its absolute magnitude. This is consistent with the decision that primary attention should be paid to the significance of validity coefficients when comparing different ethnic subgroups because of its practical implications.

All samples in the seven studies consisted of current employees. Thus, data were not available for the applicant populations. A further assumption that had to be made, therefore, was that the current employees in all ethnic subgroups were representative of their respective subgroup applicant population with respect to the predictor-criterion relationships. Finally, the assumption was made that there was no bias in the criterion measures. Unfortunately, no estimates of such bias were available; therefore, all subgroup differences and lack of differences on criteria were assumed to be a function of actual subgroup job performance.

These assumptions, in addition to the fact that small sample sizes permitted only a single estimate of each predictor-criterion relationship to be made, lead to somewhat equivocal statements in summarizing the data. All of these assumptions and restrictions must be considered when attempting to generalize from these data.

Table 58 presents a summary of predictor mean subgroup differences and validity with respect to type of predictor. It can be seen from Table 58 that the white subgroup (W) scored significantly higher than the non-white subgroup (N, either Negro or Latin American) on approximately one-fourth of the predictors. It should be remembered that a subgroup mean difference on a predictor does not necessarily indicate that the predictor is biased against one of the subgroups. If the difference on the predictor is associated with a corresponding difference on the criterion measure, the predictor may not be biased, but rather may be reflecting a difference in criterion performance. Table 59 presents the instances of unfairness with respect to type of predictor. Unfairness may exist when a difference on either the predictor or criterion measure is not associated with a corresponding subgroup difference on the other measure. From Table 59 it can be seen that the type of test most frequently (in terms of percentage of total comparisons) associated with instances of unfairness was the non-verbal intelligence test. This type of test failed to predict a criterion difference 75% of the time. The type of test which fared best with regard to unfairness was the perceptual test. When a perceptual test was the predictor, there was no unfairness in 84% of the predictor-criterion comparisons.

The concept of unfairness does not involve the validity of the predictor. Of course, both fairness and validity are desirable attributes of a predictor. In the right half of Table 58, the validity patterns of the

Table 58 : Mean Differences and Validity with Respect to Type of Predictor

Predictor	Mean Differences				Valid For				
	W>N	W<N	No Diff.	Total	W(only)	N(only)	Both	Neither	Total
Gen. IQ	3	0	2	5	16	1	12	22	51
N.-V. IQ	1	0	3	4	31	0	3	10	44
Verbal	5	0	5	10	22	7	15	40	84
Arithmetic	3	0	10	13	47	6	11	54	118
Perceptual	10	0	49	59	114	34	95	127	370
Personality	0	0	7	7	7	4	0	87	98
Total	22	0	76	98	237	52	136	340	765

Table 59 : Instances of Unfairness with Respect to Type of Predictor

Type of Predictor	Instances of Unfairness			
	Differences on Only		No Unfairness	Total
	Predictor	Criterion		
Gen. IQ	9 (18%)	14 (27%)	28 (55%)	51
N.-V. IQ	2 (5%)	33 (75%)	9 (20%)	44
Verbal	26 (31%)	16 (19%)	42 (50%)	84
Arithmetic	14 (12%)	22 (19%)	82 (69%)	118
Perceptual	46 (12%)	16 (4%)	308 (84%)	370
Personality	0 (0%)	21 (21%)	77 (79%)	98
Total	97 (13%)	122 (16%)	546 (71%)	765

predictors with respect to type of predictor are presented. A most striking fact evident from Table 58 is the large proportion of instances where the predictor was valid for only one of the subgroups. In particular, predictors were valid for only the white subgroup 237 times (of a total of 765 predictor-criterion comparisons) and valid for only the non-white subgroup 52 times. This contrast of frequency of subgroup validity lends support to the commonly held hypothesis that tests tend to be valid for white persons but not for minority group members. It must be remembered, however, that the sample sizes of white and Negro subgroups were quite dissimilar and a smaller correlation in terms of magnitude was required for significance with the white subgroups. The perceptual tests again were superior when validity was considered, being valid for at least one subgroup in about two-thirds of the total comparisons and being valid for both subgroups in approximately one-fourth of the instances. The superiority of the perceptual type of test with respect to validity was not surprising, since most of the samples consisted of clerical workers.

Table 60 presents criteria mean subgroup differences and criterion predictability summarized over the seven studies. The white subgroup scored significantly higher on about one-fourth of the criterion measures, and there were no differences on the rest. Table 61 presents instances of unfairness with respect to type of criterion. The predictability of each type of criterion measure is given in the right half of Table 60.

Since, in all instances where either predictor or criterion subgroup mean differences were found, the white subgroup scored higher on the measure than the non-white subgroup, certain consistent results concerning unfairness were found. When the difference in mean subgroup performance was on the predictor variable only, the non-white subgroup would be discriminated

Table 60 : Mean Differences and Predictability with Respect to Type of Criterion

Criterion	Mean Differences				Predictable For				
	W>N	W<N	No Diff	Total	W(only)	N(only)	Both	Neither	Total
Attendance	0	0	5	5	0	2	0	5	7
Termination	0	0	1	1	0	0	0	2	2
Extension of Probation	1	0	2	3	1	0	0	3	4
Promotion	1	0	1	2	0	0	0	2	2
Objective	0	0	4	4	2	2	0	28	32
Rating	18	0	53	71	233	48	136	300	717
Test Score Change	0	0	1	1	1	0	0	0	1
Total	20	0	67	87	237	52	136	340	765

Table 61 : Instances of Unfairness with Respect to Type of Criterion

Type of Criterion	Instances of Unfairness			
	Differences on Only		No	
	Predictor	Criterion	Unfairness	Total
Attendance	5 (71%)	0 (0%)	2 (29%)	7
Termination	1 (50%)	0 (0%)	1 (50%)	2
Extension of Probation	2 (50%)	0 (0%)	2 (50%)	4
Promotion	1 (50%)	0 (0%)	1 (50%)	2
Objective	2 (6%)	0 (0%)	30 (94%)	32
Rating	85 (12%)	122 (16%)	510 (72%)	717
Test Score Change	1 (100%)	0 (0%)	0 (0%)	1
Total	97 (13%)	122 (16%)	546 (71%)	765

against if selection were made using a common regression equation. In those cases where the difference was on the criterion only, the white subgroup would be discriminated against if the common regression line were used. Thus, the non-white subgroup was discriminated against in 13% of the instances reported in this investigation and the white subgroup in 16% of the instances, if the criterion of unfairness as defined previously is used to determine discrimination. An examination of Table 61 reveals that a rating criterion is involved in all cases of unfairness against the white subgroup. Any conclusion reached with only a rating criterion is equivocal.

All predictor-criterion relationships were also analyzed to determine the frequency of occurrence of the eleven different relationships presented in the Bartlett and O'Leary (1969) model. Table 62 presents, by sample, the frequency of each model.

Clearly, the model most often illustrated was Model 5 (no differences on criterion or predictor, but differential validity). This is not surprising since in a large number of the predictor-criterion relationships both racial groups performed equally well on both the predictor and criterion, and thus a significant correlation in either sample would produce a Model 5. It is important to note that in a majority of the illustrations of this model, the test was valid for the white sample and not valid for the minority sample.

It is unlikely that these cases would produce any differential selection rates for the ethnic groups since there was no difference in mean test performance for the two groups. Thus, viewed in terms of equal opportunity, these models do not appear to illustrate bias. However, subsequent mean job performance for the two groups would be discrepant, and one might erroneously conclude that the minority sample's ability to perform on the job was inferior

Table 60 : Mean Differences and Predictability with Respect to Type of Criterion

Criterion	Mean Differences				Predictable For				
	W>N	W<N	No Diff	Total	W(only)	N(only)	Both	Neither	Total
Attendance	0	0	5	5	0	2	0	5	7
Termination	0	0	1	1	0	0	0	2	2
Extension of Probation	1	0	2	3	1	0	0	3	4
Promotion	1	0	1	2	0	0	0	2	2
Objective	0	0	4	4	2	2	0	28	32
Rating	18	0	53	71	233	48	136	300	717
Test Score Change	0	0	1	1	1	0	0	0	1
Total	20	0	67	87	237	52	136	340	765

Table 61 : Instances of Unfairness with Respect to Type of Criterion

Type of Criterion	Instances of Unfairness			
	Differences on Only		No	
	Predictor	Criterion	Unfairness	Total
Attendance	5 (71%)	0 (0%)	2 (29%)	7
Termination	1 (50%)	0 (0%)	1 (50%)	2
Extension of Probation	2 (50%)	0 (0%)	2 (50%)	4
Promotion	1 (50%)	0 (0%)	1 (50%)	2
Objective	2 (6%)	0 (0%)	30 (94%)	32
Rating	85 (12%)	122 (16%)	510 (72%)	717
Test Score Change	1 (100%)	0 (0%)	0 (0%)	1
Total	97 (13%)	122 (16%)	546 (71%)	765

against if selection were made using a common regression equation. In those cases where the difference was on the criterion only, the white subgroup would be discriminated against if the common regression line were used. Thus, the non-white subgroup was discriminated against in 13% of the instances reported in this investigation and the white subgroup in 16% of the instances, if the criterion of unfairness as defined previously is used to determine discrimination. An examination of Table 61 reveals that a rating criterion is involved in all cases of unfairness against the white subgroup. Any conclusion reached with only a rating criterion is equivocal.

All predictor-criterion relationships were also analyzed to determine the frequency of occurrence of the eleven different relationships presented in the Bartlett and O'Leary (1969) model. Table 62 presents, by sample, the frequency of each model.

Clearly, the model most often illustrated was Model 5 (no differences on criterion or predictor, but differential validity). This is not surprising since in a large number of the predictor-criterion relationships both racial groups performed equally well on both the predictor and criterion, and thus a significant correlation in either sample would produce a Model 5. It is important to note that in a majority of the illustrations of this model, the test was valid for the white sample and not valid for the minority sample.

It is unlikely that these cases would produce any differential selection rates for the ethnic groups since there was no difference in mean test performance for the two groups. Thus, viewed in terms of equal opportunity, these models do not appear to illustrate bias. However, subsequent mean job performance for the two groups would be discrepant, and one might erroneously conclude that the minority sample's ability to perform on the job was inferior

Table 62

Frequency of Models for each Study

Study		No. of comparisons between subgroups	1	2	3	4	5	6	7	8	9	10	11	Total
1		12	0	0	0	0	1	0	1	0	0	0	0	2
2		4	0	0	0	0	0	0	1	0	0	0	0	1
3		5	0	0	0	0	0	0	1	0	0	0	0	1
4		2	0	0	0	0	0	0	1	0	0	1	0	2
5		126	10	3	16	0	10	43	5	22	0	0	1	110
6a	W-N	60	0	1	0	0	7	0	2	0	0	0	0	10
	W-L	60	0	0	0	0	40	0	3	0	0	0	0	43
6b	W-N	60	6	24	2	0	[illegible]	0	12	2	0	0	0	50
	W-L	60	0	0	0	0	42	0	1	0	0	0	0	43
6c		60	0	0	0	0	19	0	0	0	0	0	0	19
6d		60	0	0	0	0	11	9	12	4	0	3	0	39
6e		60	0	0	0	0	5	0	0	0	0	0	0	5
7		192	0	0	0	0	24	8	0	0	0	0	0	32
Total		765	16	28	18	0	103	60	39	28	0	4	1	357

to that of the white sample. This mean difference in criterion performance would be a direct result of an inappropriate selection procedure. The only solution to the selection problems of Model 5 appears to be to use the test for which it is valid and to search for other valid predictors for the non-valid group.

In view of the relatively high frequency of this model, it would seem that more research should be directed toward the development of valid predictors for minority populations. An examination of Table 58 (page [illegible]) reinforces this belief since in a large number of the total predictor-criterion relationships, the test was valid only for the white sample.

The second most frequently occurring model was Model 6 (mean difference on criterion only and differential validity). In all illustrations of this model, the white sample obtained higher ratings of job performance while there was no difference in test performance for the two groups. The use of a common regression line would always result in an over-prediction of job performance for the minority group. Thus, this model does not deny opportunity to minority group members. In fact, it systematically provides opportunity to minority groups. It is unlikely, however, that such over-prediction would benefit the minority group members in the long run. It is likely to lead only to temporary employment since the minority group member would have a low probability of success on the job.

It is also important to note that if a common regression line were employed, one would under-predict job performance for the white sample and thereby systematically reject qualified white applicants. This model illustrates the fact that not all bias is against minority groups.

Model 7 was the third most frequently occurring model (mean difference on predictor only and differential validity). As was the case with Model 5, in most illustrations of this model the test was valid only for the white

sample. However, since the minority group scored lower on the predictor, utilization of this test in selection is more detrimental to the minority group member than is Model 5. Because there was a difference in mean test performance, the minority group member has less of an opportunity to be selected. But, perhaps more important is the fact that by using such a test, one is systematically denying opportunities to minority group members on the basis of a non-valid test.

Another clear illustration of unfair discrimination is represented in Model 2 (mean difference on predictor only but equal subgroup validities). In all illustrations of this model, opportunity would be denied to minority group members since they score lower on the test, but perform as well as the white sample on the job. Since the test is valid for both groups, differential prediction is a solution to the problem. Separate regression lines and separate expectancy tables for minority and white samples would eliminate the unfair discrimination in this model.

Occurring as frequently as Model 2 was Model 4 (difference on both predictor and criterion and differential validity). Since there is a differential in both the predictor and criterion performance for the two ethnic groups, one would expect a difference in selection rates. Valid predictions can be made because the test identifies the lower performing minority group members. Nonetheless, the test is certainly not appropriate for prediction within the non-white sample.

The development of a valid predictor of job performance for minority group members will not eliminate the differential in selection rates since the minority group members do not perform as well as the white individuals on the job. However, a valid predictor for the non-white sample will insure that the most qualified minority group members will be selected.

Model 3 (difference on criterion only, equal subgroup validities), occurring 18 times, illustrates a situation where job performance is over-predicted for the non-white sample. Again, job opportunity is not denied minority group members. In this instance, the bias is against the white sample. Separate regression lines and expectancy tables will eliminate this inequality.

Perhaps the most important finding of this phase of the research project is the fact that Model 1 (no difference on predictor or criterion, equal subgroup validities) occurred so infrequently. Traditional personnel selection procedures assume that Model 1 is operative (i.e., a single regression line can be used for all subgroups in a population). The results of this study indicate that the traditional model is inappropriate in most cases. Homogeneous populations are the exception rather than the rule. Thus, it is imperative that tests be validated separately for subgroups in a population if inadvertant discrimination is to be avoided.

Models 10 (difference on predictor only, opposite subgroup validities) and 11 (differences on both predictor and criterion, no subgroup validity) occurred relatively infrequently (4 and 1 times, respectively), while Models 4 (difference on both predictor and criterion, equal subgroup validity) and 9 (no differences on predictor or criterion but opposite validity) did not appear. This would tend to indicate that these models are probably rare and are not contributing a significant amount to inadvertant discrimination in testing.

Two separate methods of model identification were utilized in those situations where differential validity was demonstrated for the two racial groups (Models 5-10). The above summary of the relative frequency of the various models utilized the first method of model identification. All

predictor-criterion relationships in which a validity coefficient was significant for one racial group but not significant for the other, were identified as models using this method.

The second method of model identification imposed an additional criterion of a statistically significant difference between the validity coefficients for the two racial groups. Table 63 presents a comparison of the relative frequency of each model, using the two methods of model identification.

Table 63

Frequency of Models Illustrated

	Method of Model Identification	
Model	Total Occurrences	Significant Differences(1)
1	16	16
2	28	28
3	18	18
4	0	0
5	163	15
6	60	9
7	39	16
8	28	4
9	0	0
10	4	4
11	1	1

(1) Using the second method of model identification, Models 5 through 10 require a significant difference between validity coefficients for the two ethnic groups to be included as an illustration of that model.

As can be seen in Table 63, the relative frequency of the various models was greatly reduced using this additional criterion of model identification. However, it is important to note that even with this more stringent criterion, inadvertant test bias was demonstrated in over 25% of the relationships.

Throughout the report we have identified those models which met the first criterion and those which met both criteria. Greater emphasis, however, has been placed on the first method of model identification because of its practical implications. That is, it is difficult to justify using a test for a given subgroup where it does not correlate significantly with the criterion, despite the fact that the correlation may not differ significantly from a valid correlation for another subgroup of the population.

Each predictor-criterion relationship was also analyzed using the regression tests of the analysis of covariance (Potthoff, 1966) to test the equality of slopes and intercepts for the ethnic groups. In general, the results of this analysis were similar to the second method of model identification. However, the analysis of covariance method identified regression intercept differences even in those cases where the test possessed no validity for either subgroup.

Table 64 presents the frequency of the various models for each of the six general classifications of predictor variables. As can be seen in the the table, it is not possible to predict which type of test is likely to produce a certain model. That is, no model was clearly associated with a particular type of test. Although the perceptual tests illustrate the most models, they were also the most frequently utilized test, since most jobs were clerical in nature. The non-verbal I.Q. tests do not reduce bias, as is sometimes assumed. The non-verbal I.Q. tests illustrated biased relationships in 33 out of a possible 44 predictor-criterion relationships.

Table 64

Frequency of Models by Type of Predictor

Type	No. of comparisons between subjects	1	2	3	4	5	6	7	8	9	10	11	Total
Gen. I.Q.	51	4	2	5	0	2	5	3	7	0	0	0	28
N.-V. I.Q.	44	0	0	2	0	4	26	1	0	0	1	0	34
Verbal	84	4	3	5	0	3	6	12	8	0	0	0	41
Arithmetic	116	2	1	4	0	28	10	5	9	0	0	1	60
Perceptual	370	6	22	2	0	118	10	18	4	0	3	0	183
Personality	96	0	0	0	0	8	3	0	0	0	0	0	11
	765	16	28	18	0	163	60	39	28	0	4	1	357

Kirkpatrick, Ewen, Barrett and Katzell (1968) have developed a useful means of summarizing data concerning the relationship between subgroup membership and predictor validity. The data from the seven studies of the present investigation have been organized according to the procedure of Kirkpatrick, et. al., and are presented in Table 65. For each sample, a number of tests were compared with a number of criteria; the product of these numbers is the number of instances where comparisons of test fairness and validity could be made, and it is listed in column 1 of Table 65. In column 2 appears the number of these predictor-criterion comparisons in which a significant mean difference between subgroups in either a test or a criterion was not associated with a significant mean difference in the other, i.e., the number of instances in which unfairness, as defined in this report, occurred. Column 3 shows the number of predictor-criterion comparisons where the test was valid for at least one of the subgroups. It might be noted that the smaller the number in column 3 is in comparison to the number in column 1, the less appropriate are the tests as a whole for predicting the job success of any of the subgroups (Kirkpatrick, et. al., 1968). Column 4 presents the number of instances in which the test was valid in one subgroup but not in the other. The larger the number in column 4 in comparison to the number in column 3, the greater the evidence that differential validity in population subgroups may exist. Column 5 indicates differential validity in the sense of the number of instances in which the validity coefficient between a given predictor and criterion significantly differs in magnitude for the two subgroups. It is useful to compare columns 4 and 5 with column 3, as well as column 1, when attempting to draw a conclusion about the relative frequency of differential validity, since column 1 contains many instances where the tests lacked validity in

Table 6?

Summary Table

Study	No. of comparisons between ethnic groups	No. of instances of unfairness	No. of instances where test is valid in at least one ethnic group	No. of instances where test is valid in one group but not other	No. of instances of significant differences in degree of validity
	1	2	3	4	5
1	12	6	2	2	1
2	4	1	1	1	1
3	5	5	1	1	0
4	2	2	2	1	1
5	120	75	112	80	0
6a (W & N)	60	5	57	7	0
6a (W & L)	60	6	55	44	0
6b (W & N)	60	42	56	18	2
6b (W & L)	60	6	43	43	2
6c	60	0	20	19	2
6d	60	28	39	36	35
6e	60	0	5	5	0
7	192	42	32	32	4
Total	765	219	425	289	48

either subgroup. Such instances may be regarded as irrelevant to the issue of differential validity, as the tests were apparently inappropriate for these situations (Kirkpatrick, et. al., 1968).

In summary, within the limitations of the data gathered and the assumptions required, the results of the present study indicate that test bias is clearly present in a large number of cases where heterogeneous groups are combined in making predictions. However, it is erroneous to conclude that all inadvertent test bias denies opportunities to minority group members. The present study has demonstrated the need to validate tests separately for minority and majority group members. The traditional validation model which assumes homogeneous populations is clearly inappropriate.

REFERENCES

Anastasi, Anne. Some implications of cultural factors for test construction. In Anastasi (Ed.), Testing Problems in Perspective. Washington, D.C.: American Council on Education, 1966.

Amrine, M. The 1965 congressional inquiry into testing: a commentary. American Psychologist, 1965, 20, 859-870.

APA Task Force on Employment Testing of Minority Groups. Job Testing and the Disadvantaged. American Psychologist, 1969, 24, 637-650.

Arvey, R. D. Unfair discrimination and tests: some issues. Unpublished paper, University of Minnesota, 1967.

Banas, P. A. Moderator variables: a review of the literature. U. S. Army Personnel Research Office, Research Memorandum 65-4, June 1965.

Bartlett, C. J. and O'Leary, B. S. A differential prediction model to moderate the effects of heterogeneous groups in personnel selection and classification. Personnel Psychology, 1969, 22, 1-17.

Boneau, C. A. The effects of violations of assumptions underlying the t test. Psychological Bulletin, 1960, 57, 49-64.

Cleary, T. Anne. Test Bias: Validity of the Scholastic Aptitude Test for Negro and White Students in Integrated Colleges. College Entrance Examination Board, Research and Development Reports, RDR-65-6, No. 18, Educational Testing Service, June 1966.

Equal Employment Opportunity Commission. Guidelines on Employment Testing Procedures. Washington, D. C.: E.E.O.C. 1966.

Guion, R. M. Personnel Testing. New York: McGraw-Hill, 1965.

Guion, R. M. Employment tests and discriminatory hiring. Industrial Relations, 1966, 5, 20-37.

Gulliksen, H. Theory of mental tests. New York: Wiley, 1950.

Hays, W. L. Statistics for psychologists. New York: Holt, Rinehart, and Winston, 1963.

Kirkpatrick, J. J., Ewen, R. B., Barrett, R. S., and Katzell, R. A. Differential selection among applicants from different socioeconomic or ethnic backgrounds. Final report to the Ford Foundation, New York University, May, 1967.

Kirkpatrick, J. J., Ewen, R. B., Barrett, R. S., and Katzell, R. A. Testing and fair employment. New York: New York University Press, 1968.

Krug, R. E. Some suggested approaches for test development and measurement. Personnel Psychology, 1966, 19, 24-34.

Lockwood, H. C. Critical problems in achieving equal employment opportunity. Personnel Psychology, 1966, 19, 3-9.

Lopez, F. M. Current problems in test performance of job applicants. Personnel Psychology, 1966, 19, 10-17.

Manning, W. H. and DuBois, P. H. Correlational methods in research on human learning. Perceptual and Motor Skills, 1962, 15, 287-321.

Mitchell, M. D., Albright, L. E., and McMurray, F. D. Biracial validation of selection procedures in a large southern plant. Proceedings of the 76th Annual Convention of the American Psychological Association, 1968, 575-576.

Potthoff, R. F. Statistical aspects of the problem of biases in psychological tests. Chapel Hill, N.C.: University of North Carolina, Institute of Statistics Mimeo Series No. 479, May 1966.

Ronan, W. W. and Prien, E. P. Toward a criterion theory: a review and analysis of research and opinion. Greensboro, N.C.: The Richardson Foundation, June 1966.

Ruda, E. and Albright, L. E. Racial differences on selection instruments related to subsequent job performance. Personnel Psychology, 1968, 21, 31-41.

Saunders, D. R. Moderator variables in prediction. Educational and Psychological Measurement, 1956, 16, 209-222.

Tenopyr, Mary L. Race and socioeconomic status as moderators in predicting machine-shop training success. Paper presented at the annual meeting of the American Psychological Association, Washington, D.C., 1967.

Wallace, Phyliss, Kissinger, Beverly, and Reynolds, Betty. Testing of minority group applicants for employment, Research Report 1966-7, Equal Employment Opportunity Commission, March 1966.

Welch, B. L. The generalization of student's problems when several different population variances are involved. Biometrika, 1947, 34, 28-35.

APPENDIX A

FIGURES ILLUSTRATING POSSIBLE EFFECTS OF A HETEROGENEOUS APPLICANT POPULATION IN PERSONNEL SELECTION PROCEDURES

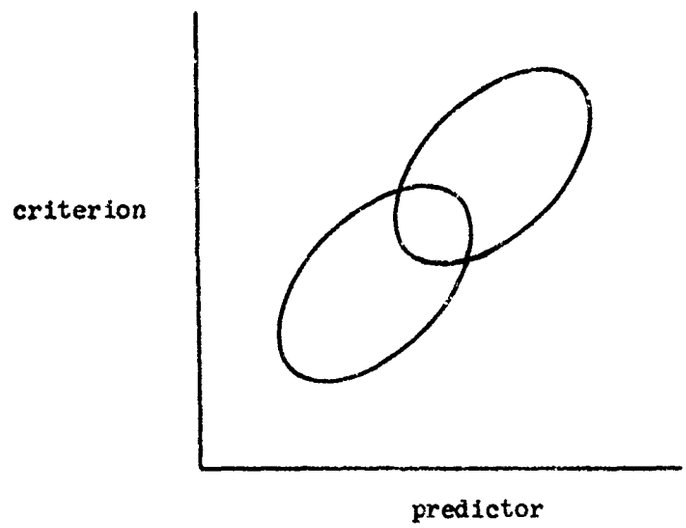


Fig. 1: Difference on Both Predictor and Criterion, Equal Validity for Both Groups

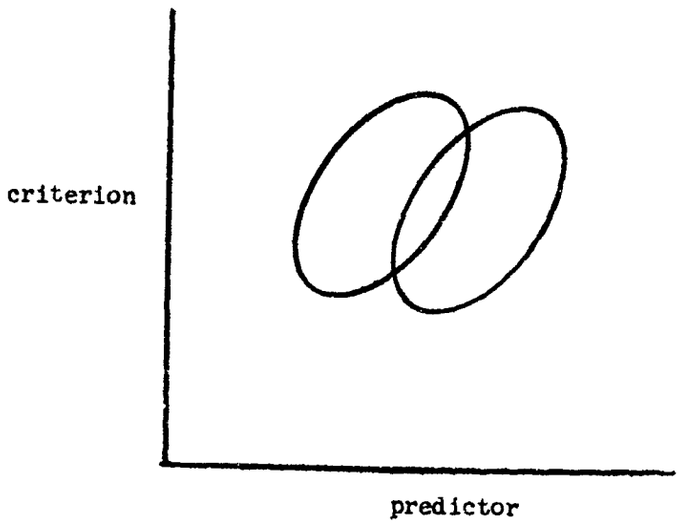


Fig. 2: Differences on Predictor Only, Equal Validity for Both Groups

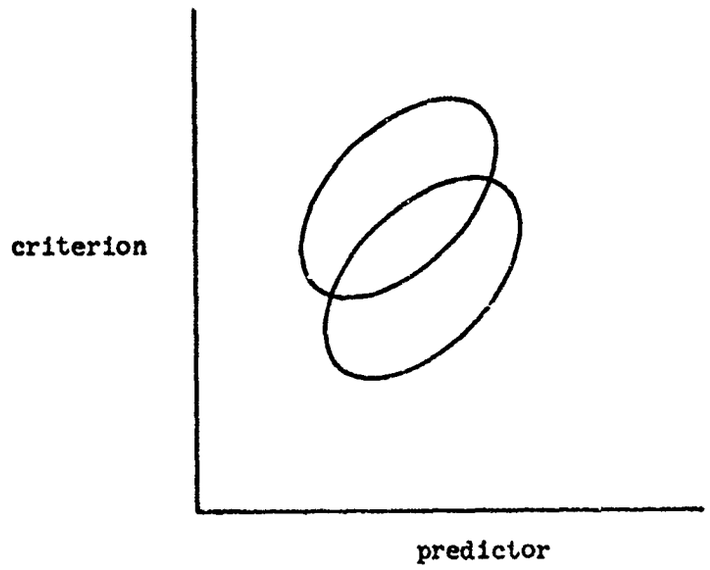


Fig. 3: Differences on Criterion Only, Equal Validity for Both Groups

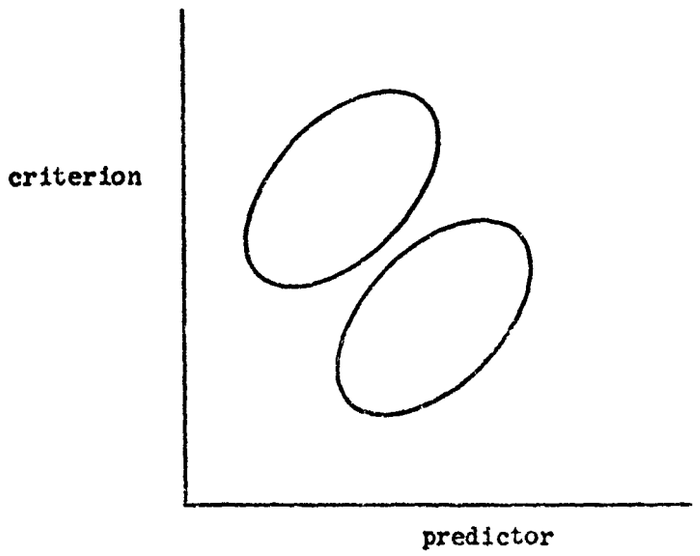


Fig. 4: Differences on Predictor and Criterion But in Opposite Direction, Equal Validity for Both Groups

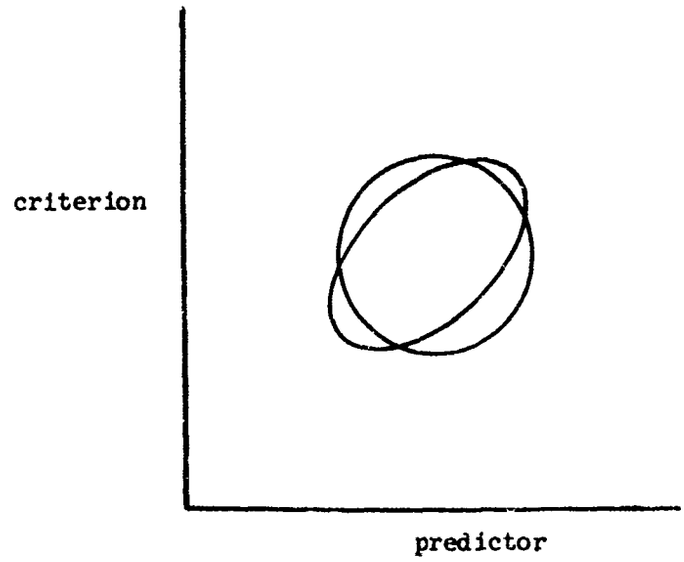


Fig. 5: Valid for Only One Subgroup, No Differences on Predictor or Criterion

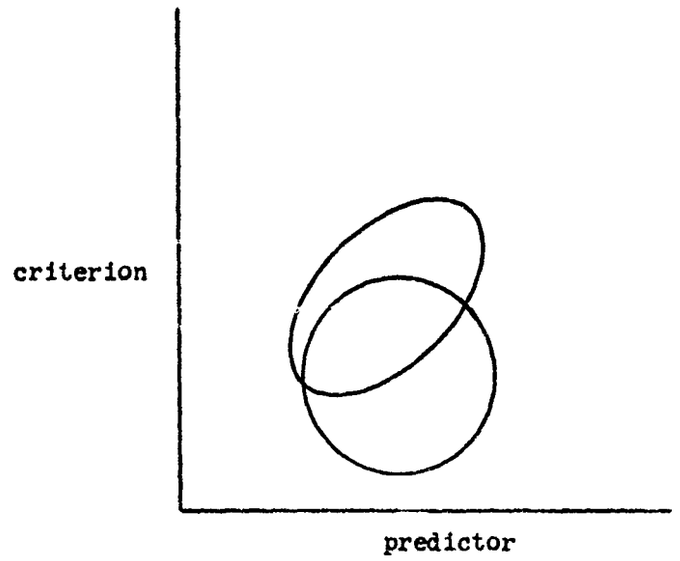


Fig. 6: Valid for Only One Subgroup, Difference on Criterion

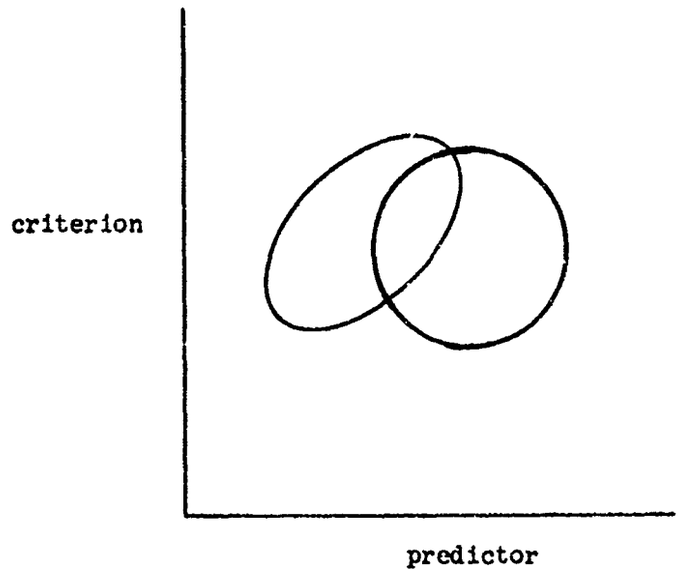


Fig. 7: Valid for Only One Subgroup, Difference on Predictor

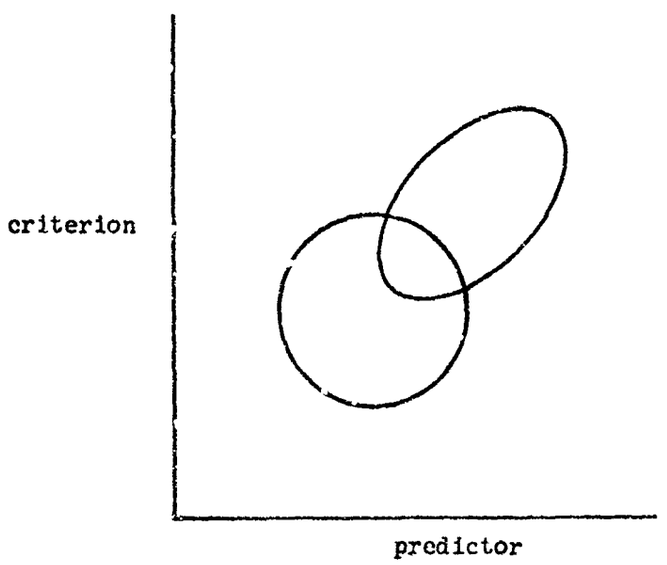


Fig. 8: Valid for Only One Subgroup, Differences on Both Predictor and Criterion

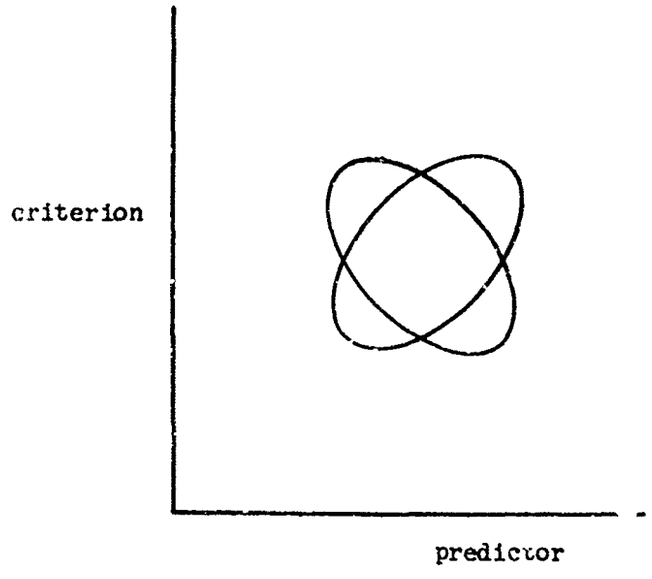


Fig. 9: Equal but opposite Validity for Subgroups, No Differences on Predictor or Criterion

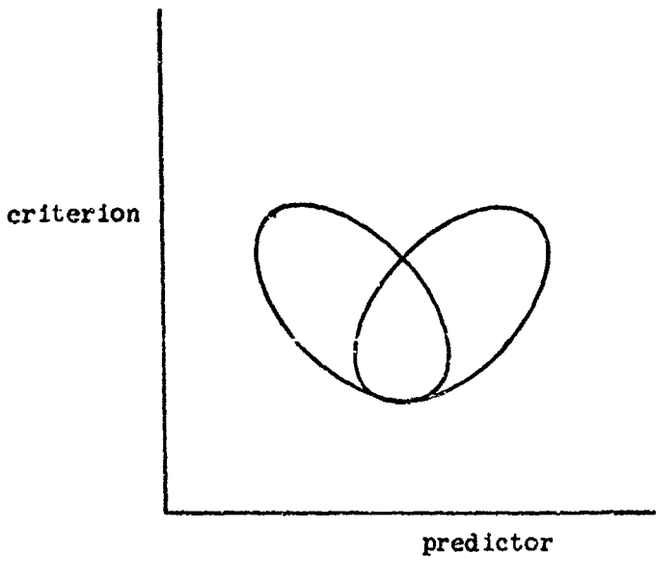


Fig. 10: Opposite Validity, Difference on Predictor Only

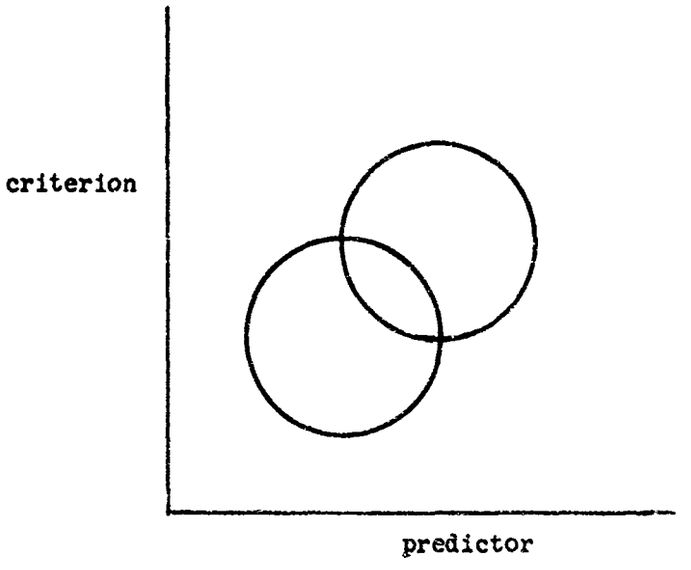


Fig. 11: No Validity in Subgroups

Security Classification

DOCUMENT CONTROL DATA - R & D

Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified

1. ORIGINATING ACTIVITY (Corporate author) American Institutes for Research 135 North Bellefield Avenue Pittsburgh, Pennsylvania 15213	2a. REPORT SECURITY CLASSIFICATION Unclassified 2b. GROUP

3. REPORT TITLE

Ethnic Group Membership as a Moderator of Job Performance

4. DESCRIPTIVE NOTES (*Type of report and inclusive dates*)
Technical Report Number 1

5. AUTHOR(S) (*First name, middle initial, last name*)

O'Leary, Brian S., Farr, James L., and Bartlett, Claude J.

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
April, 1970	ix + 152	27

8a. CONTRACT OR GRANT NO	9a. ORIGINATOR'S REPORT NUMBER(S)
N00014-68-C-0341 b. PROJECT NO NR 151-277 c d	AIR-753-4/70-TR-1 9b. OTHER REPORT NO(S) (*Any other numbers that may be assigned this report*)

10. DISTRIBUTION STATEMENT

This document has been approved for public release and sale; its distribution is unlimited. Reproduction in whole or in part is permitted for any purpose of the United States Government.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Office of Naval Research Department of the Navy Washington, D. C.

13. ABSTRACT

This report presents the findings of the first phase of a research project to investigate the problems which exist regarding subcultural differences in the prediction of job performance. Phase I of the project was an attempt to obtain an adequate picture of the effects of cultural factors on existing selection procedures. Seven independent studies were conducted in which the validity of commercial and industrially developed selection tests was examined separately for white and Negro subgroups of the population using the eleven different relationships presented in the Bartlett and O'Leary (1969) model. Occupational groups which were studied included toll collectors, correctional officers, toll facility officers, various clerical workers, and keypunch operators. A sample of inmates in a federal correctional institution was also studied.

The results of Phase I indicated that test bias is clearly present in a large number of cases where heterogeneous groups are combined in making predictions of job performance. However, it is erroneous to conclude that all inadvertent test bias denies opportunities to minority group members. The present study has demonstrated the need to validate tests separately for minority and majority group members. The traditional validation model which assumes homogeneous populations is clearly inappropriate.

The second phase of the project will involve the evaluation of procedures to control or eliminate bias. Differential prediction models, culture-equivalent tests, learning measures, as well as some non-cognitive measures will be examined.

DD FORM 1 NOV 65 1473 (PAGE 1)

S/N 0101-807-6801

Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Test Validation						
Moderator Variable						
Ethnic Group Membership						