ACT Research Report Series 87-18

A Procedure to Adjust Individual Scores for Construct Invalidity

LeAnn M. Gamache

September 1987

For additional copies write:
ACT Research Report Series
P.O. Box 168
Iowa City, Iowa 52243

A PROCEDURE TO ADJUST INDIVIDUAL SCORES FOR CONSTRUCT INVALIDITY

LeAnn M. Gamache

ABSTRACT

The study outlined a modification to Lord's procedure (1980, p. 220) as a more parsimonious method of identifying and adjusting for constructs intended to be extraneous to the measurement, by making individual adjustments only as needed. Reading level and gender were considered constructs extraneous but potentially potent relative to the particular testing situation and intended score use. Item response data by item for Sample 1 were examined for influence of the gender and reading level using item response theory and the log transformation and coincidence of regression procedure similar to that outlined by Hulin, Drasgow, and Komocar (1982). Based on the results of these examinations, a total of 10 items identified as influenced by one or more of the extraneous variables were targeted for removal. The effectiveness of the process and the improvement over that suggested by Lord were examined with ability estimates for examinees in Sample 2 using the various subsets of items. The procedure was found to be of both statistical and practical significance in reducing the influence of the extraneous characteristics in the final test score, thus improving construct validity. The procedure also was found to provide an improved alternative to that suggested by Lord due to a significant reduction, both statistically and practically, in the error of estimate for the overall testing. In summary, for these data, the suggested modification to the procedure reduced the level of influence of the extraneous variables and improved the accuracy of the ability estimate over that obtained using Lord's procedure.

Note: This study was completed in 1985.

A PROCEDURE TO ADJUST INDIVIDUAL SCORES FOR CONSTRUCT INVALIDITY

Attempts to reduce the error in measuring examinee attributes traditionally focus on testing procedures and instrument development techniques (e.g., employment of writing guidelines for item development, consideration of item difficulty and discrimination indices in item selection, examination of items for bias, etc.). Such procedures benefit the overall measurement situation and any improvement benefits the total group and/or subgroups of examinees.

Some minor focus is aimed at the other input to the testing situation -- the examinee -- who is cautioned to be well rested, to be prepared, to control anxiety, etc. It can not be assumed, however, that individual examinees are comparable in these characteristics or, perhaps more importantly, in other characteristics not intended to affect measurement of the construct to be tested. For example, it might be decided that mathematics achievement test scores should not be affected by reading ability, gender, or ethnicity, when these are considered extraneous to the measurement of mathematics achievement.

When decisions are made at the group level, the relevant information concerning the effect of these extraneous variables is reflected in the reliability and validity information typically reported and indicative of testing quality. Increasingly, however, test scores are interpreted and major, irreversible decisions made at the level of the individual as in selection, certification, and licensure situations. The group level information typically reported provides relatively little information concerning the validity or reliability of a given individual's score.

Recent advances in psychometric theory and estimation procedures may now make it possible to identify, quantify, and account for such sources of error in individual test scores. An extension of the procedure outlined by Lord (1980) to purify a test for item dysfunction at the group level after test administration may provide a procedural correction for individual score contamination by any source specified *a priori*. The purpose of this study is to examine the effectiveness of a modification/extension of Lord's procedure in improving the construct validity of individual scores by adjusting for prespecified sources of contamination.

Lord's suggestion is to analyze item data using item response theory estimation procedures, identify items functioning differentially for specified groups, remove these items from scoring, and re-estimate examinee trait levels on the "purified" subset of items. One problem with this procedure is that not all examinees have responses contaminated by the extraneous construct(s), so valid information is needlessly discarded and confidence intervals grow unnecessarily large. An improvement would be to use the "purified" estimate of examinee trait level only if it was not equal to the original estimate, i.e., only if the individual's estimate had been atypically contaminated by those prespecified extraneous constructs that were not intended to affect the measurement. Scores of both those disadvantaged and those advantaged by the contamination would be adjusted. The resulting scores would be more valid in that they are freed from the prespecified contaminating influences. Table 1 provides a summary of the entire procedure. Many choices must be made to operationalize this procedure and are noted below the table.

Review of the Literature

Efforts to Correct Individual Scores

A variety of efforts have been extended to identify and correct for error in measurements of individuals in the areas of social and psychological measurement (Cronbach, 1970; McKinley, Hathaway, & Meehl, 1948; Wiener, 1948; Seeman, 1952; Navran & Stauffacher, 1954; Dahlstrom & Welsh, 1960; Ruch & Ruch, 1967) and of human physical/perceptual measurement (Morell, 1974; Witkin & Berry, 1975; McGarvey, Maruyama, & Miller, 1977; Haller & Edgington, 1982).

In the measurement of educational aptitude and achievement, relatively less attention has been paid to identifying and correcting for sources of error resulting from the influence of extraneous factors in individual scores. The major exception is the attempt to correct for the varying tendency of individuals to guess in answering multiple-choice items. Much research has focused on scoring formulae to correct individual's scores for this confounding characteristic (Cronbach, 1970; Collet, 1971; Reilly, 1975; Frary, Cross, & Lowry, 1977; Cross & Frary, 1977; Wilcox, 1980; Abu-Sayf, 1977). Another exception is recent research concerned with development of appropriateness indices and the potential of these measures to identify individuals who did not respond in the typical manner to the testing and for whom the usual score interpretation would be inappropriate. (See Hulin, Drasgow, & Parsons, 1983; Levine & Rubin, 1979; Wright, 1977; Donlon & Fischer, 1968; Drasgow, 1982; Levine & Drasgow, 1982; Harnisch & Linn, 1981; Tatsuoka & Linn, 1983; Harnisch, 1983).

The majority of investigation of extraneous factors in ability testing, however, has attempted to identify systematic error at the group level through investigations of reliability and validity issues, including test and item bias, with any corrections to be made at the group level by, for example, omitting items or variables for all examinees if they are found to be biased for or against some examinees (Lord, 1980; Ironson, 1983; Gamache & Novick, 1985).

Lord's Procedure

Lord's suggestion is to analyze item data using item response theory estimation procedures, identify items functioning differentially for specified groups, omit those items found to function differentially, and re-estimate examinee trait levels on the "purified" subset of items. Since it is not reasonable to assume, however, that examinees are comparable in all characteristics given that they share one, identification of the contamination and correction at the level of the individual examinee is more appropriate. Smith (1981) attempted to address this issue by comparing Rasch ability estimates from items favoring males, those favoring females, and neutral items. He reports that corrections based on group membership would have been inappropriate for the majority of examinees. Fifty-six percent of the individuals had their highest estimate from a set of items other than that favoring their gender group, and 30% had their highest estimate on items biased against their gender group. Although highest estimate is not synonymous with most appropriate estimate, it appears a group-based correction may be inappropriate.

The modification of Lord's procedure proposed here to individualize the correction for contamination is diagramed in Table 1. Several decisions necessary to operationalize the procedure are noted below the table.

Operationalizing Lord's Procedure

Contaminators to Control.

Solution of the issue of which contaminators to control is situation-specific and related to test content as well as testing purpose. The host of extraneous constructs investigated in achievement testing is generally limited by available procedures and resources. The expectation often is that demographic characteristics such as gender and race should not directly affect response to a given item intended to measure achievement, nor should achievement in other content areas influence response to items (e.g., reading ability influencing response to math items). The determination of which of the many potential confounding constructs to be identified must be made with respect to the specific testing purpose, test used, and examinee group.

Choice of IRT Model.

The issue of which model to use seems to be dictated by the dimensionality of the data, the appropriateness of assumptions concerning the equality of item discrimination parameters and existence of guessing, the level of response (dichotomous, polychotomous, or continuous), as well as practical considerations of sample size, availability of local expertise, computer program resources, and cost (Gamache, 1985). It is the critical assumption of unidimensionality, prerequisite to appropriate use of the one-, two-, and three-parameter logistic models most often used with achievement data, that most requires empirical investigation with each application.

Unidimensionality is commonly tested operationally in terms of factor analysis (Ansley, 1984). Lord (1980) suggests that there is evidence of unidimensionality if the ratio of the first eigenvalue to the second is large and the remaining eigenvalues are similar in magnitude to the second. Reckase

(1979) indicates that the first factor should account for at least 20% of the variance for acceptable calibration, although good ability estimates can be obtained even when the first factor accounts for less than 10% of the total variance. Other procedures for assessing the dimensionality of a data set have been specified (Bejar, 1980; McDonald, 1980; Hambleton & Murray, 1983).

Choice of Item Bias Index. Several procedures have been developed and examined to identify differential item performance (item bias) across examinee subgroups. These may be categorized as those based on differences in difficulty (Cardall & Coffman, 1964; Cleary & Hilton, 1968; Angoff & Ford, 1973), differences in discrimination (Green & Draper, 1972; Hunter, 1975; Merz & Grossen, 1979), multivariate factor analysis (see Hulin, Drasgow, & Parsons, 1983), the chi-square statistic (Scheuneman, 1979; Shepard, Camilli, & Averill, 1980; Scheuneman, 1980), and those based on item response theory. Ironson (1983) indicates that measures of bias using item response theory fall into three general categories: 1) the difference between the item characteristic curves as a whole; 2) lack of fit measures; and 3) the differences between item parameters. In addition to Ironson's categories, classification may be based on the item response model used, as shown in Table 2, which categorizes the several procedures to a matrix combining both schema (Gamache, 1985).

According to Hulin, Drasgow, and Parsons, the many types of bias indices should be regarded as different, rather than as substitutable, solutions to the problem of identifying item bias. They are not competitive approaches; different hypotheses are being tested. The general consensus of those researchers summarizing comparison of the procedures (Shepard, Camilli, & Averill, 1981; Ironson, 1983; Hulin, Drasgow, & Parsons, 1983) is to use an item response theory based index for identification of differential item performance, or item bias. Furthermore, if practical considerations of sample size and other resources permit, an index based on the three-parameter model should be used that reflects area between the ICCs. Among such indices, however, there is little evidence of the superiority of a specific procedure; the choice must be made subjectively relative to practical considerations. If the distribution of the statistic is unknown, a within-subsample comparison should be made to determine a threshold estimate for critical difference.

Methodology

Data

To examine the effectiveness of the identification and correction procedure, a data set is needed that 1) is unidimensional; 2) provides a test score having important implications for individual examinees; 3) has potential contaminators that are specifiable, measurable, and can be used to distinguish subgroups of examinees; 4) provides information regarding status on these potential contaminators external to the set of items upon which the test score is to be based; and 5) contains enough items so, following identification of contaminator-sensitive items and temporary removal from the test, a sufficient number of items remains to allow IRT estimation of theta. Two cautions are appropriate to note *a priori*: 1) each subgroup defined by status on the potential contaminator must be sufficiently large and affected to permit evidence of differential item performance and 2) the null hypothesis that a contaminator did not affect an individual's performance can not be proven. Analysis can show only that there was lack of significant evidence that the contaminator affected the measurement of that individual.

Data from the October 1983 administration of the ACT Assessment Mathematics Usage test met these considerations. The Mathematics Usage Test consists of 40 multiple-choice items, each with five alternatives, measuring mathematical reasoning ability and solution of mathematical problems using techniques typically learned in high school mathematics courses. Two samples of three thousand examinees each were randomly selected from those participating in the regular October 1983 administration who had complete gender and test score information for all four tests. Sample 1 was used to identify items performing differentially for various groups in the first stage of the study. Sample 2 was used in the evaluation of the procedure in the second stage of the study. For purposes of this study, the sum of the Social Studies Reading and the Natural Sciences Reading scores for each examinee was used as a measure of reading level in the content areas. These tests are designed to measure the ability to comprehend, analyze, and evaluate material provided through the reading passages.

The assumption of unidimensionality of the data, prerequisite to appropriate application of the modification of Lord's procedure, was assessed using Sample 1 and principal factor analysis of a matrix of tetrachoric interitem correlations having estimated communalities on the diagonal. The data were to be considered sufficiently unidimensional if 1) the ratio of first to second factor eigenvalues was large with the remaining eigenvalues similar in magnitude to the second, and 2) the first factor accounted for more than 20% of the variation. The assumption was made by definition that any subset of these items is also unidimensional; thus a check of this assumption for each application of IRT within the study was not made.

Operationalization of Lord's Procedure

The steps for the procedure, as outlined in Table 1, include identification of potential contaminators, estimation of examinee trait levels and item parameters, examination of items for contaminability, exclusion of identified items and re-estimation of examinee trait levels and item parameters, comparison of original and purified trait estimates, and selection of the most appropriate trait estimate. Sample 1 was used for identification of items influenced by the extraneous characteristics. The results of the analysis of Sample 1 were implemented in Sample 2 for evaluation of the procedure.

Identification of Contaminators. Potential contaminators may be identified from a logical perspective. For this study, it was assumed that the test score was to provide information for placement of examinees into levels of college mathematics courses and selection of those for whom certain course requirements would be waived. The intended use of the test information, irreversibility of the decision, and potential benefit/detriment for the examinee determines the type and number of potential contaminators identified as well as determines other subjective decisions, such as specification of acceptable Type I error rate, made at later stages in the procedure. For purposes of this use, a subjective judgment was made that the influence of gender or reading level was extraneous to the measurement to be made and thus should be examined, identified, and accounted for. A logical case could be made for identification of many other extraneous factors that might influence examinee performance. A different set of potentially contaminating variables could appropriately be identified for each specific score use and testing situation. For purposes of examining the effectiveness of the procedure, reading level and gender were chosen for identification and correction due to preliminary evidence that some effect may exist and the relative size of the subgroups potentially affected.

Estimation of Examinee Trait Level. After logical identification was made of the potential contaminators to be controlled, examinee trait levels and item parameters were estimated. Data are multiple-choice items and thus susceptible to guessing. Three item parameters are thus reflected in examinee response to items; the three-parameter logistic model is considered, *a priori*, most appropriate to the data. To minimize problems associated with c-parameter estimation and unnecessary costs, the three-parameter model was modified to hold the c-parameter constant across items at a subjectively determined constant equal to the chance of a randomly correct response to the item using LOGIST V (Wingersky, Barton, & Lord, 1982).

Examination of Items. Examination of each item for atypical susceptibility to effects of gender and/or reading level was then made using Sample 1. A three-parameter model based index of the area between the curves is preferred for use in identifying differential item performance, but none is associated with a statistical distribution. (Lord's simultaneous chi-square test actually examines the differential item performance with respect to only two parameters.) Since the modified three-parameter model used to estimate the item characteristic curves results in identical c-parameter estimates for all subgroups, differences in lower asymptotes of the ICCs would not be identifiable in this study by any of the item bias detection techniques. It was thus possible to use a two-parameter based index, based on the linearizing transformation and regression procedure outlined by Hulin, Drasgow, and Komocar

(1982), to investigate the area between the ICCs estimated for subgroups defined according to status on the contaminator variables. The advantage of this procedure is that, in addition to providing an indirect index of the area between the curves in the form of probability of coincident regression lines, a statistical test for coincidence of regression lines for the subgroups is readily available after the equating of the estimated theta metrics, using a linearizing transformation of the proportion of correct response by estimated theta interval, and regression onto theta estimate for each subsample. A base 10 log transformation of the proportion correct was used as the linearizing transformation. Such transformation is strongly related to the logit transformation, with a correlation greater than .85 for more than 50 data points. This procedure was repeated for gender and for reading level. The procedure requires that subgroups be formed on the basis of status on the potential contaminator. Subgrouping by gender was straightforward; the entire Sample 1 was used to examine the influence of this variable on item performance. For investigation of item contaminability by reading level, the subsamples were identified as nearly as possible as the upper and lower thirds of the Sample 1 defined by the sum of Social Studies Reading and Natural Sciences Reading scores, in order to provide clearer examination of the possible effect of reading level.

Re-estimation of Examinee Trait Level. Trait estimates and item parameter estimates based on the total set of items were obtained for Sample 2. Items identified as reflecting a gender effect in the Sample 1 analyses were removed temporarily from the Sample 2 data and examinee trait levels and item parameters were estimated from the reduced item set to obtain theta purified for gender ($\hat{\theta}_{pg}$). Similarly, items identified as affected by level of reading were temporarily removed to allow re-estimation of item parameters and trait level to provide theta purified for reading ($\hat{\theta}_{pr}$). Finally, estimation of item parameters and of trait level for each examinee in Sample 2 was made with both sets of contaminated items removed, to provide trait level purified for gender and reading level ($\hat{\theta}_{pgr}$). Thus, each examinee in Sample 2 had a trait level estimate for the original total set of items, one purified for each of the contaminators individually, and one purified for the combination of the contaminators.

Choice of Estimate. The decision to use the original estimate or one of the adjusted estimates is predicated on definition of a meaningful difference in estimates. If a tolerable level of Type II error rate is specified *a priori*, the critical difference can be determined with respect to nonoverlapping confidence bands around the original and the purified estimate. The importance and irreversibility of the decision intended to be made from the measurement must be subjectively considered when determining significance level used in establishing confidence intervals. Type II error rate was more important than Type I in this context. For purposes of this study, significance level was subjectively set as necessary to limit the failure to appropriately adjust the score to 10% of the cases. The criterion for use of the adjusted estimate was therefore:

$\mathrm{Max}[(\hat{\theta}_{oh}-\hat{\theta}_{pl}),(\hat{\theta}_{ph}-\hat{\theta}_{ol})] \geq (\hat{\theta}_{oh}-\hat{\theta}_{ol}) + (\hat{\theta}_{ph}-\hat{\theta}_{pl})$, where the upper bound of the confidence interval for the original estimate ($\hat{\theta}_{oh}$) equals $\hat{\theta}_o + t_o/\mathrm{SQRT}[I(\hat{\theta}_o)]$, the lower bound for the original estimate ($\hat{\theta}_{ol}$) equals $\hat{\theta}_o - t_o/\mathrm{SQRT}[I(\hat{\theta}_o)]$, the upper bound for the purified estimate ($\hat{\theta}_{ph}$) equals $\hat{\theta}_p + t_p/\mathrm{SQRT}[I(\hat{\theta}_p)]$, the lower bound for the purified estimate ($\hat{\theta}_{pl}$) equals $\hat{\theta}_p - t_p/\mathrm{SQRT}[I(\hat{\theta}_p)]$, t_o equals the t-value at p=.95 with degrees of freedom based on the number of items for the original estimate, and tp equals the t-value at p=.95 with degrees of freedom based on the number of items for the purified estimate. Comparison of $\hat{\theta}_o$ with $\hat{\theta}_{pg}$ and with $\hat{\theta}_{pr}$ was made for each examinee in Sample 2 against this critical difference criterion. The adjusted estimates were used if those comparisons met or exceeded the criterion: 1) $\hat{\theta}_{pr}$ was used if only the $\hat{\theta}_{pr}$ comparison met the criterion; 2) $\hat{\theta}_{pg}$ was used if only the $\hat{\theta}_{pg}$ comparison met the criterion; and 3) $\hat{\theta}_{pgr}$ was used if both $\hat{\theta}_{pr}$ and $\hat{\theta}_{pg}$ comparisons met the criterion.

Evaluation of the Procedure

Examination of the effectiveness of the procedure was made from only those examinees in Sample 2 whose original estimates ($\hat{\theta}_o$) were adjusted ($\hat{\theta}_p$). If the procedure effectively corrects for undue influence of the specified extraneous variables, then statistical significance should be evidenced in that the multiple correlation resulting from the regression of trait level estimate on contaminator variables should be significantly less for the adjusted estimates than for the original estimates. A statistical test developed by Sympson (1979) was used to examine this hypothesis. Practical benefit of the procedure was examined by comparing the relative efficiency (Lord, 1980) of using the original and the purified estimates, i.e., $I(\hat{\theta}_p)/I(\hat{\theta}_o) > 1$ along the ability distribution.

Although individualized adjustment is to be preferred from a theoretical perspective since the larger number of items for estimation results in smaller error of estimate, practical considerations of resources may necessitate a less individualized adjustment. The improvement in error of estimate of the individualized correction over the group correction procedure suggested by

Lord is situation specific and reflects the proportion of items found to be influenced by the contaminators, the extent of the contamination, and proportion of examinees affected within the particular situation. The improvement of the individualized adjustment for this particular data set was examined using the total group of 3,000 examinees in Sample 2. Statistical significance was determined through use of the dependent t-statistic to test the quality of mean standard errors for $\hat{\theta}_{pgr}$ and $\hat{\theta}_p$ against the alternative that the mean standard error for $\hat{\theta}_{pgr}$ was greater than that for $\hat{\theta}_p$, where standard error was equal to $1/\text{SQRT}[I(\hat{\theta})]$. Given statistical significance, the practical significance of the improvement in error of estimate was investigated by examining the relative efficiency of use of the individually purified estimate ($\hat{\theta}_p$) and the group purified estimate ($\hat{\theta}_{pgr}$) for $I(\hat{\theta}_p)/I(\hat{\theta}_{pgr}) > 1$ along the trait level distribution. A subjective evaluation of relative efficiency with respect to cost ratio for the two scorings and importance of score use must be made to come to conclusion regarding importance of use of the procedure for each application situation.

Results

Two random samples from the October 1983 administration of Form 24B of the ACT Assessment were taken such that gender identification and all raw scores were available for the Mathematics Usage, English Usage, Natural Science Reading, and Social Studies Reading Tests. Each sample consisted of 3,000 examinees meeting these criteria. Descriptive information concerning Samples 1 and 2 is provided in Table 3. Means and standard deviations for the four test scores and reading level, defined as the sum of the Natural Sciences

Reading and the Social Studies Reading scores for each examinee, are given in the raw score metric.

Principal factor analysis of the matrix of tetrachoric correlations with estimated communalities on the diagonal was performed on Sample 1 to examine the unidimensionality of the data. Only the data for those examinees completing the test were examined; noncompletion has been thought to comprise a spurious second factor. The analysis was thus performed on 2,710 of the 3,000 examinees. Of the 290 examinees omitted from this phase of the analysis, females and those with low reading levels were somewhat overrepresented: 59% were female, 41% were male, 42.1% had reading level scores at or below 46, and 27.6% had reading level scores at or above 61. For purposes of examining the unidimensionality of the data set, omitted responses were coded as incorrect. The ratio of the first to second factor eigenvalues was 5.97, with the first factor accounting for 22.7 percent of the variance before rotation and having an eigenvalue of 9.062. Examination of the scree plot indicated that the data can be considered sufficiently unidimensional for purposes of using item response theory, given the criteria of a large first to second factor eigenvalue ratio and that the first factor account for more than 20% of the variance. The scree plot is given in Figure 1.

The first phase of the investigation used Sample 1 to identify items affected by gender or by reading level. Identification of items differentially affected by the extraneous variables required several steps using the modification of the Hulin, Drasgow, and Komocar procedure. The first step was to estimate the trait level parameters for each examinee, by subgroup. Estimates were then equated to a common scale. The proportion of correct response within each of the theta intervals was determined. The log transformation of this proportion was then regressed on the midpoint of each theta interval by group and the resulting regression lines were compared across groups for coincidence.

The first step of the procedure in identifying gender influenced items, therefore, was to run LOGIST V on each of the subgroups defined by gender. All program defaults were taken, except that the c-parameter was set to a constant equal to the reciprocal of the number of alternatives within each item. Thus, the c-parameter for all items was held to a constant of .20. Estimation of the item and trait level parameters for the males converged after fifteen stages. Item and trait level parameter estimates were similarly determined for the female sample. It was necessary to eliminate the record of one female examinee to allow the program to converge. Convergence occurred at the fifteenth stage following the omission of this record. The program was not able to estimate the trait level for 18 of the examinees in the total sample due to their perfect scores.

The trait level estimates for the female group were equated to the scale underlying those of the males, using the procedure outlined by Linn, Levine, Hastings, and Wardrop (1980). This procedure standardizes the item difficulty parameter estimates using weights that are a function of the sampling variance of the estimates. The equating constants are then a function of the weighted standard deviation of the difficulty estimates computed for the two groups. These constants are then applied as a linear transformation of the estimates of one group to the scale for the other. The constants needed to equate the trait level estimates for females to the scale for males were a slope of 1.046 and an intercept of -.431. Thus, the equation needed to equate these scales was -.431 + 1.046 (theta).where theta is the group-specific estimate for a given female examinee. These same constants would be used to standardize the b-parameters from the female calibration to a scale to be held in common with the male subgroup. The frequency distributions of theta estimates for males and equated theta estimates for females are given in Table 4. As with the raw score distributions, the mean theta estimate was higher for the male subgroup.

Once the thetas were equated to a common scale, the empirical item characteristic curves (ICCs) could be determined and compared. The Hulin, Drasgow, and Komocar procedure required that the theta scale be divided into intervals. Intervals were defined to reflect equal intervals along the scale, ranging from -3.0 to +3.0. The 141 examinees with trait level estimates exceeding this range were eliminated from the analyses used to identify contaminable items to avoid giving these outliers undue weighting in the estimation of the regression lines. The scale was divided into 61 intervals, each of length 0.1. The proportion of examinees in each interval answering the item correctly was determined by item to provide the empirical ICCs. To compare the ICCs by group, the Hulin, Drasgow, and Komocar procedure was modified so that the base 10 logarithm of the proportion of correct response within each theta interval was regressed on the midpoint of the theta interval. The hypothesis of coincident regression for the two groups was tested.

The results of testing this hypothesis for each item are given in Table 5 in terms of the probability of such a result by chance alone. The probability value can be interpreted as an indirect, relative index of area between the linearized curves for the groups. The smaller the probability value, the more different the observed linearized curves and the greater the area between them across the theta scale. It should be noted that, for some items and some theta intervals, there were few if any examinees. If there were no examinees at that level, the data were treated as missing and omitted from the analysis. It should also be noted that the accuracy of the proportion in each interval was a function of the number of examinees at that interval; the small samples at some intervals may have increased the probability of Type II error at the level of examination of the individual item. The repeated testing of the hypothesis over the forty items, however, most likely increased the probability of Type I error over the series of item examinations.

Three items were identified with probability less than or equal to .001 of having coincident regressions for the two groups. Items 3, 5, 19, 24, and 36 were identified with probability less than or equal to .05 of having coincident regressions for the groups. Eleven items were identified as having probabilities of coincident regression lines for the groups of less than or equal to .10. It was believed that a probability level set at less than or equal to .05 would control the probability of Type II error sufficiently, given the increase in probability of Type I error due to the repeated testings. Characteristics of these items are given in Table 6. Group-specific difficulty and discrimination parameter estimates from separate calibrations using the Sample 1 subgroups are listed with their respective standard errors of estimate by item. A brief content description of each item is also given. For each gender group, this group of items appears indistinguishable from other items in the test with respect to these characteristics.

The Hulin, Drasgow, and Komocar procedure modified to use the log transformation was repeated to identify items influenced by reading level. Subgroups were formed based on the sum of the Natural Sciences Reading and Social Studies Reading scores to allocate, as closely as possible given the distribution of raw scores, one-third of Sample 1 to the low reading level group and one-third to the high reading level group. Examinees with reading level scores at or below 46 (N=951) were categorized as low readers; those with reading level scores at or above 61 (N=1154) were categorized as high readers. As in the gender analyses, item parameter and theta estimates were obtained for each subgroup separately using LOGIST V, taking all program defaults with the exception that the c-parameter was held to a constant of .20. Estimation converged after fifteen stages for the low group and fifteen stages for the high group. Using the Linn, Levine, Hastings, and Wardrop (1980) procedure, the trait level estimates of the low reading group were equated to the scale for the high reading group using the equation -1.999 + 1.343(theta). Frequency distributions of theta estimates along the common theta scale are given for those with high and those with low reading level scores in Table 7. As with the distribution of raw scores on the mathematics variable, those with high reading level scores as a group had higher theta estimates.

Empirical ICCs were obtained, using the same designation of intervals along the theta scale as was used in the gender analyses. The results of the modified Hulin, Drasgow, and Komocar procedure testing the coincidence of regression lines for the low reading and high reading groups are given, by item, in Table 5.

The hypothesis of coincidence is rejected at a probability level less than or equal to .05 for 21 of the 40 items. For this data set, a correlation of .71 was found between mathematics total raw score and the reading level variable. Those in the low reading group typically also had very low mathematics raw scores. Many scored at or below the chance level. It may be that examinee response to an item for those at this level reflects some characteristic other than the mathematics achievement reflected in the responses of examinees at a level of achievement similar to that of the high reading level group. Perhaps it was this influence, rather than reading level per se, that affected the rejection of the hypothesis of coincidence for some items. Other explanations may also be possible.

Since the general academic achievement trait, intelligence, or whatever, affects most items, items were ranked by probability level, and the five most affected were identified for removal. The limit of five was chosen for the practical needs of retaining a sufficient number for use of item response theory procedures. Items 16, 13, 36, 15, and 20 therefore were targeted for removal, although many other items were not meaningfully different in the degree to which they were influenced by this "extraneous" construct. Characteristics of these items are given in Table 8 with the group-specific difficulty and discrimination parameter estimates and the respective standard errors of estimate by item. As a group, these items do not appear unique for each of the subgroups in these characteristics given the other items in the test.

The second phase of the investigation examined 1) the effectiveness of the modification of Lord's procedure and 2) the degree of improvement of the modified procedure over that suggested by Lord. Using LOGIST V and the procedures outlined for the earlier analyses, trait estimates were obtained for examinees in Sample 2 from several subsets of items: the total set of forty items, providing $\hat{\theta}_o$; the subset remaining after elimination of the five gender affected items, providing $\hat{\theta}_{pg}$; the subset of items remaining after elimination of the five most reading influenced items, providing $\hat{\theta}_{pr}$; and the subset of items resulting from elimination of the five gender influenced and the five most reading influenced items not already omitted for gender affect, to provide $\hat{\theta}_{pgr}$. Trait estimates could not be made for a total of 39 examinees due to their perfect scores or scores of 0 on one or more of these item subsets. The correlations between the theta estimates based on these item subsets for Sample 2 ranged from .94 to .98 as shown in Table 9.

To determine the most appropriate trait level estimate for each individual ($\hat{\theta}_p$), the estimate purified for gender ($\hat{\theta}_{pg}$) and that purified for reading level ($\hat{\theta}_{pr}$) were each compared against the original estimate ($\hat{\theta}_o$) based on the total set of items. If the comparison met the adjustment criterion of non-overlap of confidence intervals around the estimates, an estimate other than the original was indicated as appropriate for that individual. If the adjustment criterion was met by the $\hat{\theta}_o$ to $\hat{\theta}_{pg}$ comparison only, $\hat{\theta}_{pg}$ was used as $\hat{\theta}_p$. If the adjustment criterion was met by the $\hat{\theta}_o$ to $\hat{\theta}_{pr}$ comparison only, the $\hat{\theta}_{pr}$ was used as $\hat{\theta}_p$. If both comparisons met the adjustment criterion, $\hat{\theta}_{pgr}$ was used as the most appropriate estimate $\hat{\theta}_p$ for the individual.

Of the 2,961 examinees having estimable trait levels on all four sets of items, 821 met the criterion for use of an adjusted theta. Examination revealed that 99 of the 101 examinees at the lowest point of the theta scale were included in this adjusted group and that all but 5 received the same adjusted theta estimate. The trait level estimates for the remaining 2 examinees could not be obtained on one or more of the subsets of items. These outliers were considered inappropriate to this portion of the evaluation of the procedure in that 1) the use of an adjusted estimate for these outliers was likely artifactually related to the narrowness of confidence intervals at this extreme and 2) approximately 70% of the group had raw scores at or below 8, the raw score expected by chance. This is inconsistent with the assumption of having obtained a meaningful measurement of achievement for these examinees. Since such examinees are part of the reality of the applied measurement situation and are not generally evaluated and omitted from scoring and reporting, they were included in all stages of application of the procedure. These 99 examinees, however, were deleted from the sample for this stage of the evaluation to guard against confounding the measurement problems for these examinees with the effectiveness of the procedure.

A sample of 722 examinees thus remained to evaluate the effectiveness of the procedure. For these examinees, values of the difference between $\hat{\theta}_p$ and $\hat{\theta}_o$ ranged from -6.81 to 3.07 with a mean difference of -.277, a median difference of -1.00, and standard deviation of 1.51. The appropriate estimate, $\hat{\theta}_p$, was less than the original estimate, $\hat{\theta}_o$, for 59.1% of this group. Additional descriptive information concerning this group is provided in Table 10. It appears that the extraneous variables most affected performance of the lower scoring individuals. Within each of the subgroups defined by an extraneous variable (males, females, high readers, and low readers), those needing an adjusted estimate had lower scores and reported lower high school grade point average than those for whom the original estimate was most appropriate. Also, from the sample sizes in Table 11, it can be seen that if performance was influenced by one of the extraneous variables, it was highly likely to be affected by the other as well. Most of the examinees for whom an adjusted estimate was appropriate required use of the theta estimate purified for both gender and reading level.

This group of examinees was used to examine the effectiveness of the procedure in reducing the influence of the extraneous variables. The effectiveness of the procedure in reducing the influence of gender and reading level was examined using a test for equality of multiple correlations for large, dependent samples (Sympson, 1979). Results of testing this hypothesis and information concerning the regressions of $\hat{\theta}_o$ and $\hat{\theta}_p$ separately on reading level, gender, and the interaction are given in Table 12. A significant reduction in correlation of trait estimate with the contaminating characteristics was made with use of the purified estimates $\hat{\theta}_p$. The combination of gender, reading level, and the interaction accounted for 13.51% of the variance in $\hat{\theta}_o$, whereas these variables accounted for only 9.15% of the variance in $\hat{\theta}_p$. Practical significance of the improvement is seen from the ratio of test information functions for the two estimates. Descriptive information for the ratio of the test information function for $\hat{\theta}_p$ to that for $\hat{\theta}_o$ for this group of examinees is given in Table 13. The 33 examinees with ratio values exceeding 50000 were omitted from this descriptive information due to the effect of such extreme values on such information. For each of these examinees, $\hat{\theta}_p$ was very low and less than $\hat{\theta}_o$. Males and those with low reading level scores were overrepresented in this group of those with extremely large improvement in test information function with use of $\hat{\theta}_p$. Although the large majority had increased information, approximately 37.4% of the 722 examinees experienced a decrease in the test information function with use of the most appropriate estimate $\hat{\theta}_p$.

The entire group of examinees in Sample 2 for whom an estimate could be made was used to examine the improvement of the procedure over the elimination of all potentially influenced items for all examinees as suggested by Lord. The reduction in error of estimate for the final, purified estimate $\hat{\theta}_p$, over that for the estimate omitting all potentially influenced items, $\hat{\theta}_{pgr}$, was tested for statistical significance using the dependent t-test to test the hypothesis of the equality of mean standard errors for $\hat{\theta}_p$ and $\hat{\theta}_{pgr}$ against the alternative that the mean standard error for $\hat{\theta}_{pgr}$ was greater than that for $\hat{\theta}_p$. As indicated in Table 14, the error of estimate for the suggested procedure was significantly less than that using Lord's procedure with a t-value of -18.63 and p less than or equal to .00. The practical significance of this improvement was examined through the ratio of the test information function of $\hat{\theta}_p$ to that of $\hat{\theta}_{pgr}$. Descriptive information concerning this ratio is provided in Table 15. Although 518 examinees had relative efficiency values equal to 1 because $\hat{\theta}_{pgr}$ was their most appropriate estimate, $\hat{\theta}_p$, an increase in relative efficiency was found for 59.9% of the total group with use of $\hat{\theta}_p$. The median was 23.4% improvement in relative efficiency, although the distribution was highly skewed and a much greater increase was found for many examinees.

Discussion and Conclusions

The study outlined a modification to Lord's procedure as a more parsimonious method of identifying and adjusting for constructs intended to be extraneous to the measurement, by making individual adjustments only as needed. Reading level and gender were considered constructs extraneous but potentially potent relative to the particular testing situation and intended score use. Item response data by item for Sample 1 were examined for influence of the gender and reading level using item response theory and the log transformation and coincidence of regression procedure similar to that outlined by Hulin, Drasgow, and Komocar. Based on the results of these examinations, a total of 10 items identified as influenced by one or more of the extraneous variables were targeted for removal. The effectiveness of the process and the improvement over that suggested by Lord were examined with estimates for examinees in Sample 2 using the various subsets of items. The procedure was found to be of both statistical and practical significance in reducing the influence of the extraneous characteristics in the final test score, thus improving construct validity. The procedure also was found to provide an improved alternative to that suggested by Lord due to a significant reduction, both statistically and practically, in the error of estimate for the overall testing. In summary, for these data, the suggested modification to the procedure reduced the level of influence of the extraneous variables and improved the accuracy of the estimate over that obtained using Lord's procedure.

Some Interesting Details

It should be noted that analysis of items for reading level effect resulted in rejection of the hypothesis of coincidence for 21 of the 40 items at a probability level of less than or equal to .05. To the degree that the Social Studies Reading and Natural Sciences Reading Tests tap the same analytical, general academic, or ability traits as the Mathematics Usage Test, this may be expected. Subsequent studies should examine this finding using a more refined reading level variable (e.g., Noble, 1985) that is less associated with the mathematics variable. Since the Hulin, Drasgow, and Komocar procedure examines item performance against a statistical distribution, the conclusion of contaminability is not relative to other items in the test. The conclusion of contaminability, however, is specific to the characteristics of the particular sample of examinees and thus situation specific, rather than an intrinsic characteristic of the item.

It is interesting to note, secondly, that the majority of examinees for whom a purified estimate was appropriate, both overall and by subgroup, had consistently lower reported mean high school grade point averages as well as lower test scores than those for whom the original estimate was appropriate. It appears that performance of the lower ability examinees may be more likely to be affected by the extraneous constructs. In placement situations, in certification situations, and in others where decisions are more likely to be made at other than the high end of the score scale, it may be particularly important to examine individual estimates for contamination using such a procedure to reduce sources of construct invalidity.

Third, if a purified estimate was appropriate for a given individual, it was likely that the estimate needed to be purified for both gender and reading level influences. This may mean that the extraneous variables are highly related or that certain persons' performances may be susceptible to sources of construct invalidity in general, rather than to specific extraneous constructs. Future research with these examinees using measures of other constructs deemed extraneous to the measurement may provide additional insight.

Finally, although the majority of affected examinees experienced an increase in test information function, approximately 37.4% of the group experienced a decrease in the test information function with use of the most appropriate estimate. This may be due to 1) the false contribution of the influenced items to the information for those examinees benefited by the influence of the extraneous constructs and/or 2) the effect of fewer items underlying the estimate. The adjusted estimate is more valid by definition since some of the effect of extraneous variables has been accounted for. However, since the level of validity is limited by the level of reliability, this increment in validity may be undermined by the loss of reliability due to fewer items. If further investigation indicates this reduced level of information is due only to the effect of fewer items, the criterion for use of the adjusted estimate may need to be expanded to require that there be no loss in test information function for the given individual.

Broader Issues

The impetus for this investigation was to outline and examine a procedure to evaluate and improve construct validity of a measurement at the level of the individual examinee, to be consistent with the usual individual use of scores and individual decisions made in situations such as certification, selection, and placement. The procedure, however, must rely on some group based, sample dependent analyses in detection of contaminable items. If too small a portion of the group is affected by the extraneous construct, items can not be targeted for removal. To the extent the bias analyses do not identify items contaminated for a particular individual, the individual's estimate can not be purified. Further study should indicate the minimum proportion of the group needing to be affected to provide sufficient power for identifying potentially contaminable items. The current procedure provides improvement in the purity of an individual's score but can not be assumed to purify it completely for these extraneous constructs.

A related issue concerns the identification of the extraneous variables. It may be that a variable considered extraneous actually functions as a surrogate variable for a characteristic intended to validly be reflected in the measurement. In some settings gender, for example, may be a surrogate for differential coursework which would be appropriately reflected in measurement of mathematics achievement. It might also be argued that the observed differences reflect gender differences in interest, confidence, or cognitive style that validly influence mathematics achievement. In such case, gender would not be a contaminator at all. A similar argument might be made that reading level has an appropriate influence on mathematics achievement. To the extent that test specifications include domains such as application of mathematics concepts in realistic situations or translation or synthesis of mathematical symbols with textual information, for example, reading level would be unlikely to be considered a contaminator. If such conditions exist in a particular setting, use of this procedure with these variables might actually reduce, rather than increase, the validity of the assessment. For each assessment context, extremely careful conceptual and quantitative consideration is required before a variable is labeled as a contaminator. In the assessment situation hypothesized for purposes of this particular study, it was assumed that for the hypothetical score use gender and reading level could be considered extraneous.

Another issue concerns the choice of item bias index used in the procedure. The choice of item bias index is critical to the effectiveness of the procedure. To the extent that an index reflects the degree of contaminability relative to other items in the test, the effectiveness of the procedure is restricted, since fewer items can be identified for examination at the individual level.

One policy issue concerns the adjustment of an examinee's score. The most valid, appropriate estimate for an individual may be higher or lower than the original estimate for the particular examinee. The issue of using different items for different examinees may initially appear inequitable from a political viewpoint. The situation, however, is similar in some respects to that of adaptive testing where individuals are administered items targeted to obtain a prespecified level of accuracy for the measurement. Examinees may take both differing items and differing numbers of items, yet scores are comparable. In the current study, items are targeted to be scored for an individual to minimize the influence of extraneous constructs that either inappropriately advantage or disadvantage that individual. Although unpopular from the perspective of an affected examinee, this procedure would result in a lower, but more valid and appropriate, estimate for examinees advantaged by the influence of the extraneous constructs. Just as use of a high estimate resulting from examinee cheating behavior would be invalid, an estimate reflecting advantage due to the influence of extraneous constructs would be less valid for the intended score use.

A final issue, highlighted by the group of 101 examinees estimated at the lowest point of the continuum, is the need in testing endeavors to routinely evaluate examinee data for appropriateness from several perspectives to verify the assumption of having obtained a meaningful measurement of each individual. Although some researchers have invested considerable effort in analysis of aberrant response patterns and similar endeavors, in practice scores generally are processed and reported without evaluation from this perspective or verification that a raw score at or below chance level is indeed a meaningful measurement for a given individual. When test scores are used at the level of the individual, this issue becomes particularly central to the validity concern.

Future research is needed to incorporate ongoing improvements in equating, IRT estimation procedures, and item bias detection techniques as well as to address some of the detail-related issues outlined above, including the cause of the reduction in test information that was found for some examinees with use of the adjusted estimate and relationship to the need for the additional criterion of no loss of test information in order to assign an adjusted estimate. In the larger research arena, such a procedure may be found to facilitate investigation of characteristics that distinguish examinees whose performance is particularly susceptible to influence by extraneous constructs. Such a procedure may also be found useful in identifying the source of test bias in differential prediction situations, in improving the predictor or criterion as needed, or in reducing the number of false negative and false positive classifications in certification situations.

Several operational and policy challenges need resolution before implementation of such a procedure in a given situation. Issues include how to process examinees with perfect scores or scores equal to 0, since trait estimates can not be determined for such examinees given the estimation programs in common use. Another challenge requires development of a viable explanation for examinees whose appropriate estimate may be lower than that obtained from the total set of items as well as explanation as to why the same raw score may be associated with different trait estimates even when based on the same set of items. An additional challenge involves explication of the rationale, statistical and psychometric issues, procedures, limitations, and possible potential of such a procedure to testing practitioners and policy makers.

In conclusion, given further research and resolution of the operational challenges, the procedure may have the potential to make an operational contribution to improving construct validity in some testing situations by reducing the influence of extraneous constructs in measurements that meet the requirements for use of such a procedure and of item response theory. Such requirements include unidimensionality of the data, large sample sizes, and tests of sufficient length to allow omission of items. The need for additional refinement and use of such a procedure becomes increasingly critical as testing is used to make major, irreversible decisions at the level of the individual examinee in the wide variety of certification, selection, and placement situations. In such situations the additional costs in time and resources needed to apply such a procedure may be insignificant given the potential impact for an examinee.

TABLE 1

Outline of the Procedure

Stage 1: Examine items for influence of contaminators

1. Identification of potential contaminating variables
2. Estimation of item and examinee parameters using IRT
3. Examination of each item for contamination
4. Elimination of contaminated items to form a "purified" set of items

Stage 2: Adjust examinee scores if contaminated

1. Estimation of examinee trait level from purified item set
2. Decision to use original or adjusted estimate to be made for each examinee

Issues to Resolve for Each Application Situation:

1. Which extraneous variables to control
2. Which IRT model to use
3. Which method to use in identifying biased items
4. What criterion to establish as indicative of a real difference in trait level estimates

TABLE 2

IRT Based Item Bias Indices

	Difference Between Curves	Difference Between Parameters	Lack of fit Measures
One-Parameter		t-statistic (Wright, Mead, & Draba, 1976)	Lack of fit (Wright & Stone, 1979)
Two-Parameter	Linearized ICC (Hulin, Drasgow, & Komocar, 1982)		
Three-Parameter	Signed or unsigned area (Rudner, 1977)	Simultaneous Chi-Square (Lord, 1980)	Differences in probabilities
	Squared differences (Linn et al., 1980, 1981)		Standardized differences in probabilities (Linn & Harnisch, 1981)
	Weighted area measure Weighted squared differences		
	Sum of squares		
	Visual inspection of plots against standard error		
	Base high area Base low area Base high and low Root mean squared difference (Linn, Levine, Hastings, & Wardrop, 1981)		

TABLE 3

Characteristics of Samples 1 and 2

S1 Variables	Total	Male	Female	High Readers	Low Readers
ESCORE Mean	46.07	44.94	46.94	55.17	35.86
SD	12.63	12.57	12.61	9.66	9.86
MSCORE Mean	20.78	23.06	19.03	27.42	13.60
SD	8.93	8.98	8.49	7.18	6.11
SSCORE Mean	30.23	31.69	29.11	39.13	20.15
SD	9.19	9.07	9.13	5.10	4.52
NSCORE Mean	25.63	28.56	23.37	33.54	17.64
SD	8.17	8.47	7.16	5.65	4.00
RDLVL Mean	55.86	60.26	52.48	72.67	37.79
SD	16.13	16.33	15.13	8.61	6.26
Sample N	3,000	1,303	1,697	1,154	951

S2 Variables	Total	Male	Female	High Readers	Low Readers
ESCORE Mean	45.96	44.49	47.00	55.46	35.79
SD	12.82	12.89	12.67	9.64	10.08
MSCORE Mean	20.44	22.55	18.94	27.45	13.46
SD	8.92	8.94	8.59	7.37	5.92
SSCORE Mean	29.80	31.01	28.94	39.05	19.79
SD	9.29	9.43	9.09	5.07	4.78
NSCORE Mean	25.39	28.36	23.28	33.52	17.35
SD	8.27	8.65	7.30	5.80	3.85
RDLVL Mean	55.19	59.37	52.22	72.64	37.14
SD	16.36	16.96	15.25	8.74	6.61
Sample N	3,000	1,246	1,754	1,114	975

TABLE 4

Distributions of Equated Trait Estimates for Gender Groups

Equated Trait Estimate	Percent Cumulative Frequency	
	Male	Female
Less than -7.00	2.4	4.0
-6.99 to -2.33	4.9	9.1
-2.32 to -1.59	10.0	17.7
-1.58 to -1.16	14.9	26.4
-1.15 to -0.94	20.0	33.0
-0.93 to -0.71	24.7	41.0
-0.70 to -0.52	29.9	47.0
-0.51 to -0.38	34.6	52.8
-0.37 to -0.23	40.0	58.1
-0.22 to -0.12	44.7	61.4
-0.11 to 0.01	49.8	65.2
0.02 to 0.13	54.9	70.2
0.14 to 0.27	59.6	75.1
0.28 to 0.39	64.8	78.4
0.40 to 0.52	69.9	82.9
0.53 to 0.67	74.9	86.6
0.68 to 0.79	79.7	89.4
0.80 to 0.95	84.8	92.4
0.96 to 1.18	89.9	95.3
1.19 to 1.54	94.8	98.4
1.55 to 2.37	100.0	99.8
Greater than 2.38	100.0	100.0
Mean	-0.205	-0.768
Median	0.020	-0.452
SD	1.524	1.860
N	1,303	1,697

TABLE 5

Probability of Coincident Regression Lines

Item	P($\hat{\theta}$pg)	P($\hat{\theta}$pr)	Item	P($\hat{\theta}$pg)	P($\hat{\theta}$pr)
1	.2379	.0092	21	.9371	.0232
2	.1665	.0375	22	.8799	.0029
3	.0249	.4259	23	.9343	.0182
4	.6303	.0730	24	.0301	.0004
5	.0002	.0009	25	.0533	.1694
6	.5582	.6675	26	.0984	.0024
7	.1410	.1176	27	.9347	.0057
8	.2857	.1133	28	.5243	.1996
9	.4664	.1512	29	.9051	.1597
10	.5314	.5358	30	.2456	.0143
11	.0719	.0123	31	.6717	.0067
12	.9091	4673	32	.0788	.0113
13	.9553	.0001	33	.5426	.2667
14	.0863	.1912	34	.2245	.0193
15	.1790	.0002	35	.1096	.2710
16	.8200	.0001	36	.0047	.0001
17	.6666	.3085	37	.1565	.0005
18	.6699	.0766	38	.7580	.1785
19	.0003	.2001	39	.8410	.0260
20	.3725	.0003	40	.0645	.1131

TABLE 6

Characteristics of Gender Influenced Items

Item	Characteristics		
3	Algebraic manipulation with decimals		
		Males	Females
	a(s.e.)	0.83436(0.083)	1.00363(0.084)
	b(s.e.)	-0.83262(0.139)	-0.32357(0.082)
5	Story problem wiih percentages		
		Males	Females
	a(s.e.)	0.57531(0.070)	0.53800(0.063)
	b(s.e.)	-0.49407(0.199)	0.14480(0.152)
19	Multi-step problem with numeric multiples		
		Males	Females
	a(s.e.)	0.83681(0.100)	0.60025(0.078)
	b(s.e.)	0.29608(0.105)	0.64049(0.126)
24	Circle geometry problem with change of metric		
		Males	Females
	a(s.e.)	1.01534(0.107)	1.27702(0.137)
	b(s.e.)	0.10230(0.086)	0.80288(0.055)
36	Manipulation of algebraic expression presented in text		
		Males	Females
	a(s.e.)	1.16081(0.137)	1.69128(0.218)
	b(s.e.)	0.60942(0.070)	1.20214(0.051)

TABLE 7

Distributions of Equated Trait Estimates by Reading Group

Equated Trait Estimate	Percent Cumulative Frequency High	Low
Less than -7.00	0.1	8.9
-6.99 to -1.76	4.9	56.6
-1.75 to -1.28	10.0	72.2
-1.27 to -1.04	14.9	78.5
-1.05 to -0.86	19.9	82.6
-0.85 to -0.66	24.8	87.1
-0.65 to -0.52	29.9	90.3
-0.51 to -0.37	34.8	92.7
-0.36 to -0.23	39.9	94.6
-0.22 to -0.11	45.0	95.6
-0.10 to 0.00	49.6	96.6
0.01 to 0.13	54.7	97.3
0.14 to 0.24	59.7	98.2
0.25 to 0.37	65.0	98.6
0.38 to 0.50	69.9	98.7
0.51 to 0.62	74.8	99.2
0.63 to 0.80	79.9	99.4
0.81 to 0.98	85.0	99.6
0.99 to 1.24	89.9	99.7
1.25 to 1.71	95.0	100.0
1.72 to 2.50	100.0	100.0
Mean	-0.016	-2.892
Median	0.010	-1.945
SD	1.040	3.061
N	1,154	951

TABLE 8

Characteristics of Reading Influenced Items

Item	Characteristics		
13	Manipulation of algebraic expression presented in text		
		Low Readers	High Readers
	a(s.e.)	1.79166(0.275)	0.99108(0.115)
	b(s.e.)	0.98138(0.058)	-0.35170(0.131)
15	Figuring angle of polygon using symbols in test		
		Low Readers	High Readers
	a(s.e.)	1.28477(0.172)	0.58255(0.095)
	b(s.e.)	0.63990(0.068)	-1.50129(0.536)
16	Calculation using absolute value with textual problem		
		Low Readers	High Readers
	a(s.e.)	1.84257(0.249)	0.97473(0.114)
	b(s.e.)	0.78725(0.053)	-0.91589(0.176)
20	Story problem with percentages		
		Low Readers	High Readers
	a(s.e.)	0.66938(0.224)	0.89770(0.109)
	b(s.e.)	2.13370(0.355)	-0.61283(0.173)
36	Manipulation of algebraic expression presented in text		
		Low Readers	High Readers
	a(s.e.)	1.53417(0.379)	0.81366(0.116)
	b(s.e.)	1.73388(0.138)	0.26998(0.142)

TABLE 9

Intercorrelations of Trait Estimates from Item Sets

	$\hat{\theta}$o	$\hat{\theta}$pg	$\hat{\theta}$pr
$\hat{\theta}$pg	.9437		
$\hat{\theta}$pr	.9742	.9434	
$\hat{\theta}$pgr	.9334	.9753	.9443

TABLE 10

Characteristics of Examinees Requiring Use of Purified Estimates

	Examinees With Purified Estimates			Examinees With Original Estimates		
	N	Mean	SD	N	Mean	SD
HSGPA	634	262.15	60.65	2053	308.02	63.84
Males	181	243.24	59.82	943	298.42	67.27
Females	453	269.71	59.38	1110	316.18	59.61
High R	42	299.38	63.92	1010	324.22	61.73
Low R	415	256.12	59.19	376	280.91	61.98
MSCORE	722	10.92	3.93	2140	24.04	6.99
Males	210	11.27	4.21	985	25.10	7.06
Females	521	10.78	3.81	1155	23.13	6.80
High R	46	15.98	9.49	1034	27.67	6.58
Low R	484	10.20	2.82	400	18.83	4.90
RDLVL	722	42.31	11.54	2140	60.06	14.65
Males	201	43.71	12.16	985	62.90	15.26
Females	521	41.76	11.25	1155	57.64	13.66
High R	46	67.74	6.77	1034	72.55	8.58
Low R	484	35.91	6.79	400	39.56	5.26

TABLE 11

Characteristics of Examinees by Purified Subgroup

	$\hat{\theta}pg$			$\hat{\theta}pr$			$\hat{\theta}pgr$		
	N	Mean	SD	N	Mean	SD	N	Mean	SD
HSGPA	93	274.94	61.25	82	271.45	54.74	459	257.90	61.11
Males	29	239.97	55.75	27	251.56	52.95	125	242.21	62.34
Females	64	290.78	57.26	55	281.22	53.38	334	263.78	59.67
High R	9	325.89	62.21	12	291.67	60.62	21	292.43	66.36
Low R	52	262.39	59.41	39	266.69	48.92	324	253.84	60.22
MSCORE	104	13.57	3.91	100	13.86	4.89	518	9.82	3.10
Males	33	14.09	4.38	33	13.46	4.99	135	10.04	3.37
Females	71	13.32	3.68	67	14.06	4.86	383	9.74	3.01
High R	9	19.44	9.90	16	17.63	10.14	21	13.24	8.43
Low R	61	12.72	2.16	46	12.94	2.54	377	9.46	2.50
RDLVL	104	44.98	11.88	100	47.75	13.05	518	40.72	10.73
Males	33	46.49	12.62	33	50.76	13.38	135	41.31	10.95
Females	71	44.28	11.55	67	46.27	12.72	383	40.51	10.66
High R	9	66.67	6.21	16	69.00	6.98	21	67.24	7.01
Low R	61	36.93	7.13	46	36.85	7.04	377	35.62	6.70

TABLE 12

Impact on the Relationship with the Extraneous Variables

	Theta Estimate	
	$\hat{\theta}o$	$\hat{\theta}p$
N	722	722
Mean	-1.475	-1.752
SD	.990	1.496
Correlation with gender	-.065	.014
Correlation with reading level	.366	.300
Regression results		
Coefficient for gender	-.07541	.27406
Significance	.7942	.5404
Coefficient for reading level	.03143	.04512
Significance	.0057	.0103
Coefficient for interaction	.00017	.00346
Significance	.9795	.7293
Constant	-2.66226	-3.88185
Significance	.0000	.0000
Multiple R	.3675	.3025
R Squared	.1351	.0915

Ho: $R(\hat{\theta}o) = R(\hat{\theta}p)$
Sympson's test z = -12.24
p > .001

TABLE 13

Ratio of Test Information Functions for the Purified to Original Estimates

Ratio Value	Percent Cumulative Frequency
Less than 1.00	39.2
1.00 to 1.99	49.5
2.00 to 2.99	58.8
3.00 to 3.99	67.3
4.00 to 4.99	72.4
5.00 to 5.99	76.1
6.00 to 6.99	79.5
7.00 to 7.99	81.3
8.00 to 8.99	82.0
9.00 to 9.99	82.9
Greater than 9.99	100.0

Mean = 271.23
Median = 2.021
SD = 2464.53

TABLE 14

Equality of the Errors of Estimate

Variable	Mean	SD
s.e. $(\hat{\theta}p)$	.0120	.008
s.e. $(\hat{\theta}pgr)$	.0133	.010

Ho: s.e. $(\hat{\theta}p)$ = s.e. $(\hat{\theta}pgr)$
$t = -18.63$, df = 2,960
$p < .001$

TABLE 15

Practical Improvement in Error of Estimate

Ratio Value	Percent Cumulative Frequency
Less than 0.50	6.0
0.50 to 0.74	11.8
0.75 to 0.99	19.3
1.00 to 1.24	51.5
1.25 to 1.49	90.0
1.50 to 1.74	94.6
1.75 to 1.99	95.9
Greater than 1.99	100.0

Mean = 2.279
Median = 1.234
SD = 57.854

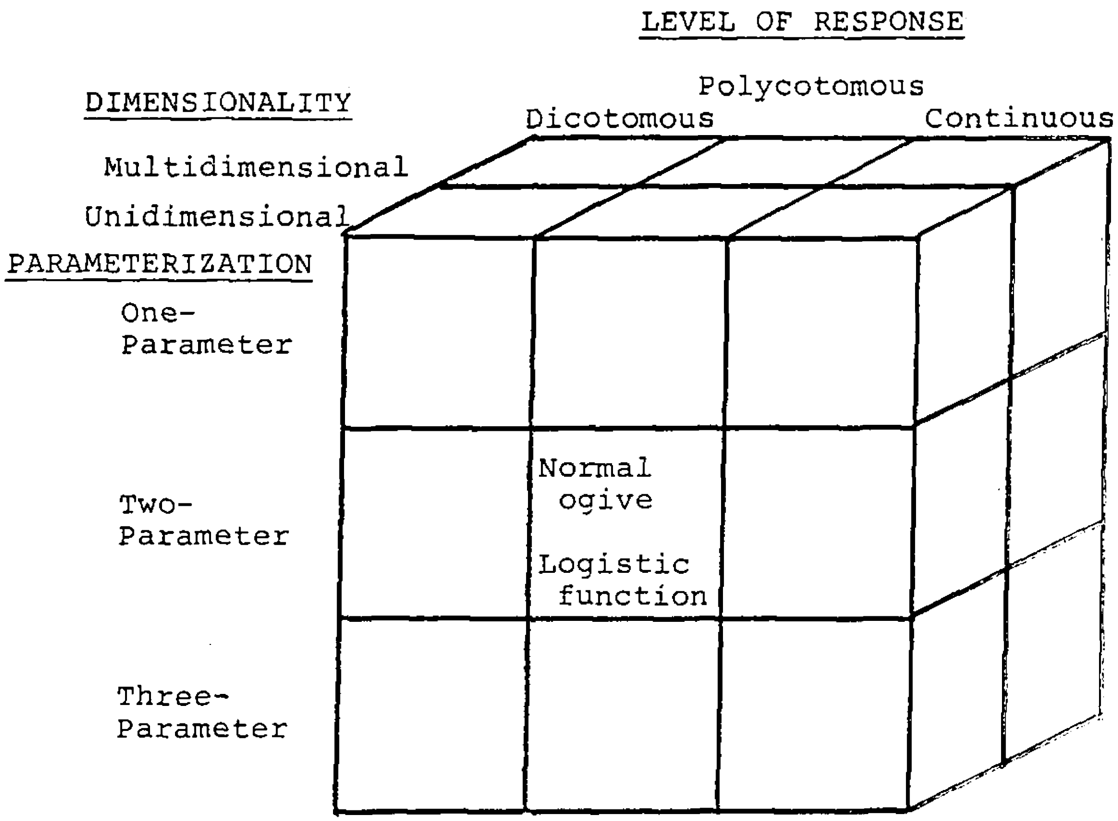

Figure 1. Framework of Static Class of Item Response Models

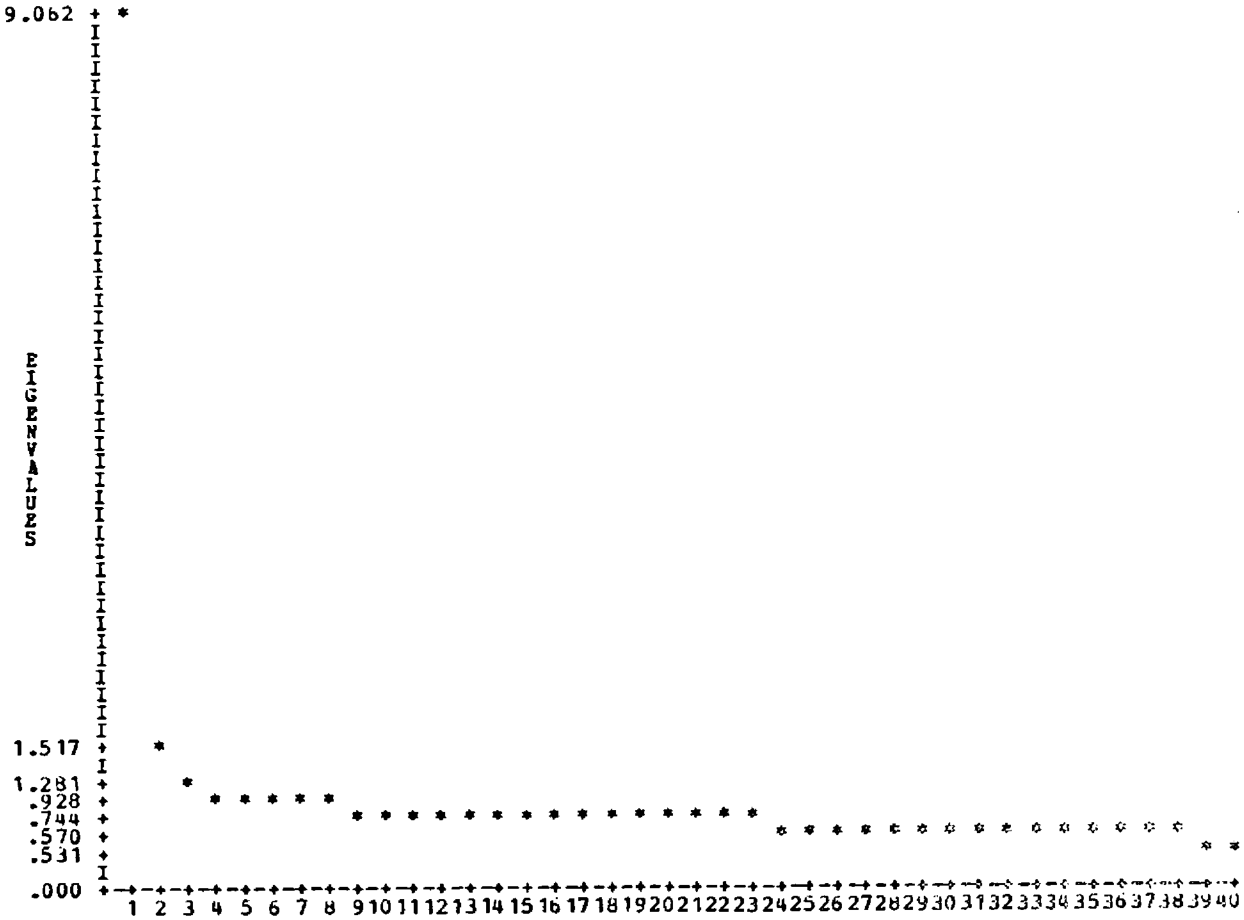

Figure 2. Scree Plot

REFERENCES

Abu-Sayf, F.K. (1977). A new formula score. *Educational and Psychological Measurement*, *37*, 853-862.

Angoff, W.H., & Ford, S.F. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement*, *10*, 95-106.

Ansley, T.N. (1984). *An empirical investigation of the effects of applying a unidimensional latent trait model to two-dimensional data.* Unpublished doctoral dissertation, The University of Iowa, Iowa City.

Bejar, I.I. (1980). A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. *Journal of Educational Measurement*, *17*, 283-296.

Cardall, C., & Coffman, W.E. (1964). *A method for comparing the performance of different groups on the items in a test.* (Research Bulletin 64-61). Princeton, N.J.: Educational Testing Service.

Cleary, T.A., & Hilton, T.L. (1968). An investigation of item bias. *Educational and Psychological Measurement*, *28*, 61-75.

Collet, L.S. (1971). Elimination scoring: An empirical evaluation. *Journal of Educational Measurement*, *8*, 209-214.

Craig, R., & Ironson, G.H. (1981). *The validity and power of selected item bias techniques using an a priori classification of items.* Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles.

Cronbach, L.J. (1970). *Essentials of Psychological Testing.* New York: Harper & Row Publishers.

Cross, L.H., & Frary, R.B. (1977). An empirical test of Lord's theoretical results regarding formula scoring of multiple-choice tests. *Journal of Educational Measurement*, *14*, 313-321.

Dahlstrom, W.G., & Welsh, G.S. (1960). *An MMPI handbook: A guide to use in clinical practice and research.* Minneapolis: University of Minnesota Press.

Donlon, T.F., & Fischer, F.E. (1968). An index of an individual's agreement with group-determined item difficulties. *Educational and Psychological Measurement*, *28*, 105-113.

Drasgow, F. (1982). Choice of test model for appropriateness measurement. *Applied Psychological Measurement*, *6*, 297-308.

Drasgow, F., & Parsons, C.K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement*, *7*, 189-199.

Frary, R.B., Cross, L.H., & Lowry, S.R. (1977). Random guessing, correction for guessing, and reliability of multiple-choice test scores. *Journal of Experimental Education*, *46*, 11-15.

Gamache, L.M. (1985). *Improving construct validity by identifying, quantifying, and accounting for sources of error in individual scores*. Unpublished doctoral dissertation, The University of Iowa, Iowa City.

Gamache, L.M., & Novick, M.R. (1985). Choice of variables and gender differentiated prediction within selected academic programs. *Journal of Educational Measurement*, *22*, 53-70.

Green, D.R., & Draper, J.F. (1972). *Exploratory studies of bias in achievement tests*. Paper presented at the Annual Meeting of the American Psychological Association, Honolulu. (ED070794)

Haller, O., & Edgington, E.S. (1982). Scoring rod-and-frame tests: Quantitative and qualitative considerations. *Perceptual and Motor Skills*, *55*, 587-593.

Hambleton, R.K., & Murray, L.N. (1983). Some goodness of fit investigations for item response models. In R.K. Hambleton, (Ed.), *Applications of Item Response Theory*. Vancouver, British Columbia: Educational Research Institute of British Columbia.

Harnisch, D.L. (1983). Item response patterns: Applications for educational practice. *Journal of Educational Measurement*, *20*, 191-206.

Harnisch, D.L., & Linn, R.L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement*, *18*, 133-146.

Hulin, C.L., Drasgow, F., & Komocar, J. (1982). Applications of item response theory to analysis of attitude scale translations. *Journal of Applied Psychology*, *67*, 818-825.

Hulin, C.L., Drasgow, F., & Parsons, C.K. (1983). *Item Response Theory: Application to Psychological Measurement*. Homewood, Ill; Dow Jones-Irwin.

Hunter, J.E. (1975). *A critical analysis of the use of item means and item-test correlations to determine the presence or absence of content bias in achievement test items*. Paper presented at the National Institute of Education Conference on Test Bias, Maryland, December, 1975.

Ironson, G.H. (1983). Using item response theory to measure bias. In R.K. Hambleton, (Ed.), *Applications of Item Response Theory*. Vancouver, British Columbia: Educational Research Institute of British Columbia.

Levine, M.V., & Drasgow, F. (1982). Appropriateness measurement: Review, critique and validating studies. *British Journal of Mathematical and Statistical Psychology*, *35*, 42-56.

Levine, M.V., & Rubin, D.B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics*, *4*, 269-290.

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

Linn, R.L., Levine, M.V., Hastings, C.N., & Wardrop, J.L. (1981). Item bias in a test of reading comprehension. *Applied Psychological Measurement*, *5*, 159-173.

Linn, R.L., Levine, M.V., Hastings, C.N., & Wardrop, J.L. (1980). An investigation of item bias in a test of reading comprehension. *Technical Report No. 163*. Urbana-Champaign, IL: Center for the Study of Reading, University of Illinois. (ED184091)

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

McDonald, R.P. (1980). A simple comprehensive model for the analysis of covariance structures: Some remarks on applications. *British Journal of Mathematical and Statistical Psychology*, *33*, 161-183.

McGarvey, B., Maruyama, G., & Miller, N. (1977). Scoring field dependence: A methodological analysis of five rod-and-frame scoring systems. *Applied Psychological Measurement*, *1*, 433-446.

McKinley, J.C., Hathaway, S.R., & Meehl, P.E. (1948). The Minnesota Multiphasic Personality Inventory: VI. The K Scale. *Journal of Consulting Psychology*, *12*, 20-31.

Merz, W.R., & Grossen, N.E. (1979). *An empirical investigation of six methods for examining test item bias*. Report submitted to the National Institute of Education, Grant NIE-6-78-0067, California State University, Sacramento. (ED178566)

Morell, J. (1974). The training of field independence. Unpublished doctoral dissertation, Northwestern University.

Navran, L., & Stauffacher, J.C. (1954). Social desirability as a factor in Edwards' Personality Preference Schedule performance. *Journal of Consulting Psychology*, *18*, 442.

Noble, J. (1985). Estimating reading skill from ACT Assessment scores. *ACT Research Report No. 88*. Iowa City, IA: The American College Testing Program.

Reckase, M.D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, *4*, 207-230.

Reilly, R.R. (1975). Empirical option weighting with a correction for guessing. *Educational and Psychological Measurement*, *35*, 613-619.

Ruch, F.L., & Ruch, W.W. (1967). The K factor as a (validity) suppressor variable in predicting success in selling. *Journal of Applied Psychology*, *51*, 201-204.

Rudner, L.M. (1977). *An approach to biased item identification using latent trait measurement theory*. Paper presented at the Annual Meeting of the American Educational Research Association, New York. (ED137337)

Scheuneman, J. (1980). A posteriori analyses of biased items. In *Test item bias methodology: The state of the art*. The Johns Hopkins University National Symposium on Educational Research, Washington, D.C.

Scheuneman, J. (1979). A method of assessing bias in test items. *Journal of Educational Measurement*, *16*, 143-152.

Seeman, W. (1952). "Subtlety" in structured personality tests. *Journal of Consulting and Clinical Psychology*, *16*, 278-283.

Shepard, L.A., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics*, *6*, 317-375.

Shepard, L., Camilli, G., & Averill, M. (1980). *Comparisons of six approaches for detecting test item bias using both internal and external ability criteria*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Boston.

Smith, R.M. (1981). *An analysis of the individual effects of sex bias*. A paper presented at the Annual Meeting of the National Council on Measurement in Education, Los Angeles.

Sympson, J.B. (1979). *Testing differences between multiple correlations*. (Research Report 79-20). Princeton, N.J.: Educational Testing Service.

Tatsuoka, K.K., & Linn, R.L. (1983). Indices for detecting unusual patterns: Links between two general approaches and potential applications. *Applied Psychological Measurement*, *7*, 81-96.

Wiener, D.N. (1948). Subtle and obvious keys for the Minnesota Multiphasic Personality Inventory. *Journal of Consulting Psychology*, *12*, 164-170.

Wilcox, R.R. (1980). Some results and comments on using latent structure models to measure achievement. *Educational and Psychological Measurement*, *40*, 645-658.

Witkin, H.A., & Berry, J.W. (1975). Psychological differentiation in cross-cultural perspective. *Journal of Cross-Cultural Psychology*, *6*, 4-87.

Wingersky, M.S., Barton, M.H., & Lord, F.M. (1982). *LOGIST User's Guide*. Princeton, N.J.: Educational Testing Service.

Wright, B.D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, *14*, 97-116.

Wright, B.D., Mead, R.J., & Draba, R. (1976). _Detecting and correcting test item bias with a logistic response model_. Research Memorandum 22, Statistical Laboratory, Department of Education, University of Chicago.

Wright, B.D., & Stone, M.H. (1979). _Best test design_. Chicago: MESA Press.