SOME USES OF STRUCTURAL EQUATION MODELING IN VALIDITY STUDIES: EXTENDING IRT TO EXTERNAL VARIABLES USING SIMS RESULTS

Bengt Muthen

CSE Report No. 268

Center for the Study of Evaluation
Graduate School of Education
University of California, Los Angeles

1986

The project presented, or reported herein, was performed pursuant to a Grant from the Office of Educational Research and Improvement/Department of Education (OERI/ED). However, the opinions expressed herein do not necessarily reflect the position or policy of the OERI/ED and no official endorsement by the OERI/ED should be inferred.

ABSTRACT

This paper proposes the use of a new extension of standard Item Response Theory (IRT) modeling of dichotomous variables to include external variables. The extension requires the formulation of a model for the relationships between both categorical and continuous external variables and response items. Four important issues in item analysis are addressed simultaneously:

1. Estimation of IRT type item measurement parameters.

2. Assessment of the strengths of hypothesized antecedents to the student's latent trait level

3. Detection of item bias (differential item performance).

4. Testing and relaxation of the IRT requirements of unidimensionality and conditional independence.

The model and its underlying methodology are illustrated with data from the eighth grade US sample from the Second International Mathematics Study.

I. Introduction

The aim of this paper is to propose the use of a new extension of standard Item Response Theory (IRT) modeling of dichotomous items to include external variables. Because external variables may appear both as categorical grouping variables and as continuous variables. This requires the formulation of a model for the relationships between the external variables and the response items. Given the availability of sufficiently rich data, such extensions can yield a more informative and powerful analysis of constructs and their measurement than what has so far been possible by standard IRT.

To make the discussion concrete, we will illustrate the methodology in the context of educational achievement test data, analyzing the eighth grade US sample from the Second International Mathematics Study (SIMS, Crosswhite, Dossey, Swafford, McKnight & Cooney, 1985). The achievement testing covered topics in algebra, measurement, geometry, and arithmetic. The responses to a set of algebra items administered at the end of the eighth grade will be related to a set of external variables in the form of background variables measured at the beginning of the eighth grade. The background variables include scores on mathematics tests, family background variables, information on the student's attitude towards math, and type of math class attended in the eighth grade. This information will be brought together in a single model. The new general feature of this model is that it simultaneously addresses four important issues in item analysis:

(i) Estimation of IRT type item measurement parameters.

(ii) Assessment of the strengths of hypothesized antecedents to the student's latent trait level.

(iii) Detection of item bias (differential item performance).

(iv) Testing and relaxation of the IRT requirements of unidimensionality and conditional independence.

While the major novelty is the inclusion of external variables, there are several new specific features of the analyses to be presented. One feature is the relaxation of the conditional independence requirement for certain items that by virtue of the question format have an association that can not be described solely by their common dependence on the single trait. Another feature concerns the handling of items that have been deemed "biased", e.g., items that are sensitive to instructional coverage, but still contain valuable measurement information. Such items can be retained in the model by explicitly including parameters that describe the differential item performance. A third feature is the potential for explaining item bias by the influence of background variables. A fourth feature is a stronger test of unidimensionality obtained by checking the homogeneity of the items in relation to the background variables, not only by considering inter item associations as is customary. Finally, the modeling is capable of including several sets of items of differing content in a simultaneous analysis of several traits.

To prepare for a discussion of the general modeling approach of Section III and the data analysis in Section V, Section II briefly outlines relevant latent variable measurement modeling theory for dichotomous and continuous response variables. Section III outlines theory for the structural equation modeling that we propose for data of this kind. Section IV describes the response items and a set of interesting additional variables that are available in the SIMS data. Section V uses this modeling approach to analyze the relationship between some of the response variables of the SIMS data and a set of external variables. Section 6 concludes.

The statistically less sophisticated reader may wish to skip sections II and III and go straight to the description of the data in section IV. Before doing do, such a reader may wish to note that the modeling framework is given in Figure 1, where the relationships between the dichotomously scored y's and the latent trait η are described in an IRT fashion by two-parameter normal ogive item characteristic curves, while the relationship between η and the background variables of x is described by a standard linear regression (although values for η need not be estimated to obtain these regression coefficients).

II. Latent Variable Measurement Modeling

Let us consider dichotomous and continuous response variable models. Assume a vector of p continuous latent response variables y* that follow a standard linear measurement model in each of g groups of students (the student subscript i and the group subscript will be deleted),

$$y^* = \nu + \Lambda \eta + \varepsilon \qquad (1)$$

where η is the latent variable vector, ε is the vector of measurement errors, ν and Λ are intercept and slope (loading) measurement parameters, so that

$$E (y^*) = \nu + \Lambda \kappa \qquad (2)$$

$$V (y^*) = \Lambda \psi \Lambda' + \Theta \qquad (3)$$

where κ is the mean vector of η, ψ is the covariance matrix of η, and Θ is the covariance matrix of the measurement errors, usually assumed to be diagonal.

When modeling dichotomous response variables we have for variable j

$$y_j = 1, \text{ if } y_j^* \geq \tau_j \qquad (4)$$
$$0, \text{ otherwise}$$

When working with aggregates of items in the form of subscores or item parcels, we assume a continuous response variable,

$$y_j = y_j^* \qquad (5)$$

This is the standard confirmatory factor analysis measurement framework of Joreskog (1969), extended to a comparative multiple-group analysis in Joreskog (1971) and Sorbom (1974, 1978), extended to a multiple factor dichotomous response model by Christoffersson (1975), Muthen (1978), and Bock & Aitkin (1981), and further extended to dichotomous multiple-group analysis in Muthen & Christoffersson (1981). For an overview, see Mislevy (1986).

The generality of the above type of covariance/correlation structure framework makes it suitable for a wide range of analyses involving validity issues, see Joreskog (1978) and for instance Bohrnstedt (1983). One specific example concerns the analysis of multitraitmultimethod matrices by covariance structure methods; for a recent overview see Schmitt & Stults (1986).

Let us consider factor analytic modeling of achievement variables of the SIMS type. Our interest may be in assessing the dimensionality and strength of relation between each observed variable and construct(s). The observed variables may represent the subscores for the different content areas of algebra, measurement, geometry, and arithmetic. The subscores may be broken down in suitable item parcels so that there are several observed scores for each area. We may entertain the simplistic hypothesis of a four-factor structure, assuming that the responses within each content area are unidimensional and that the correlations between the scores from different areas can be fully explained by their dependencies on the correlated constructs. We may also study the measurement qualities and relationships among the constructs across subgroups of students. By multiple-group approaches we may then test hypothesis of invariant measurement parameters in the G groups, such as

$$\nu_1 = \nu_2 = \ldots = \nu_G = \nu \tag{6}$$

$$\Lambda_1 = \Lambda_2 = \ldots = \Lambda_G = \Lambda \tag{7}$$

If (6) and (7) are true we may next want to test the structural hypotheses

$$\kappa_1 = \kappa_2 = \ldots = \kappa_G = \kappa \tag{8}$$

$$\psi_1 \quad \psi_2 = \ldots = \psi_G = \psi \tag{9}$$

We may find that for different instructional exposure to the topics covered in the test items, invariance of ν, or Λ may not hold for certain of the item parcel scores related to certain constructs, while for other scores measurement invariance may be found. As noted by Miller & Linn (1986), the instructional coverage may be assumed to affect the construct in question homogeneously across a set of test items, so that bias does not exist at the item level. To further scrutinize such issues of validity in educational achievement data, it is useful to be able to shift the analysis from the score level down to the "micro" item level. Such an effort will be described below, although it should be kept in mind that the techniques to be discussed are equally applicable on the aggregated continuous score level

III. A Structural Model

Let y* be as in (1) and let the vector of latent constructs follow the linear structural equation system

$$\eta = \alpha + B\eta + \Gamma x + \zeta, \tag{10}$$

where α is an intercept parameter vector, B is a matrix of slopes for regressions among the η's (the diagonal elements of B are zero and $I - B$ is nonsingular), Γ is a matrix of slopes for regressions of the η's on the set of q exogenous observed x variables, while ζ is a vector of residuals. With standard assumptions it follows that

$$E\,(y^*|x) = \nu + \Lambda(I - B)^{-1}\alpha + \Lambda(I - B)^{-1}\Gamma x, \tag{11}$$

$$V\,(y^*|x) = \Lambda(I - B)^{-1}\psi(I - B)'^{-1}\Lambda' + \Theta \tag{12}$$

This model framework was described in Muthen (1983, 1984), where it was pointed out that structural models with dichotomous, ordered categorical, and continuous latent variable indicators could be fitted into the following three-part structure:

part 1: $\sigma_1 = \Delta^*\{K_\tau\tau - K_\nu[\nu + \Lambda(I - B^{-1}\alpha]\},$ (13)

(mean/threshold/reduced-form regression intercept structure)

part 2: $\sigma_2 = \text{vec}\,\{\Delta\Lambda(I - B)^{-1}\Gamma\},$ (14)

(reduced-form regression slope structure)

part 3: $\sigma_3 = K\,\text{vec}\,\{\Delta[\Lambda(I - B)^{-1}\psi(I - B)'^{-1}\Lambda' + \Theta]\Delta\}.$ (15)

(covariance/correlation/reduced-form residual correlation structure)

Here, Δ represents a diagonal matrix of scaling factors related to the covariance matrix $V\,(y^* \mid x)$ and the K matrices are designed to select various elements. This model also encompasses the LISREL formulation of Joreskog (1973, 1977) and Joreskog & Sorbom (1984). For an overview of the various types of modeling that are possible, see Muthen (1983).

The parameters of the model are estimated by minimization of the generalized least squares fitting function

$$F = 1/2\,(s - \sigma)'\,W^{-1}\,(s - \sigma) \tag{16}$$

where **s** contains the sample quantities corresponding to α, $\alpha' = (\sigma_1', \sigma_2', \sigma_3')$, and **W** is an estimate of the asymptotic covariance matrix of s. Twice the F value at the minimum gives an approximation to a large-sample chi square test of model fit to the restrictions imposed on σ. Large sample standard errors of parameter estimates are readily available. For technical details, see Muthen (1984).

3.1 Extending IRT to external variables: a MIMIC structural probit model

Of particular interest in this paper is the formulation of a special case of the above general model, namely a model with a single construct underlying a set of dichotomous items (letting $\nu = 0$),

$$y^* = \lambda\eta + \varepsilon \tag{17}$$

It is well-known that assuming a normal ε that is independent of η and has independent elements gives rise to the two-parameter normal ogive model of Item Response Theory (IRT), see e.g., Lord & Novick (1968). This specifies a probit regression of each y on η. We will now extend this IRT model to include a set of regressors x,

$$\eta = \alpha + \gamma' x + \zeta. \tag{18}$$

This model is schematically depicted in Figure 1.

The reduced-form solution for y* is

$$y^* = \lambda\,\alpha + \lambda\,\gamma' x + \lambda\,\zeta + \varepsilon \tag{19}$$

The reduced form regression intercept vector is $\lambda\,\alpha$, the reduced-form regression slope matrix is $\lambda\gamma'$ and has rank one, while the reduced-form residual covariance matrix $\lambda\,\psi\,\lambda' + \Theta$ has a single factor correlational structure. To standardize, we take $V(y^* \mid x)$ to have unit diagonal elements. We will add the multivariate probit assumption that $y^* \mid x$ is multivariate normal. Note that this does not mean that we assume normality for the y*'s or for η, but normality is merely required for the residual ζ and for ε. The distribution of η and the y*'s is actually to some extent generated by the x's.

In its continuous response form, this is the traditional so called MIMIC (multiple indicators and multiple causes) structural equation model described, e.g., in Joreskog & Goldberger (1975); see also references therein. For dichotomous response variables, this type of model has been studied in Muthen (1979, 1981, 1983, 1985), and in Muthen & Speckart (1985), where it was termed a structural probit model.

A multiple group version of the MIMIC model with dichotomous responses would seem to be particularly useful in analyzing the present set of achievement data, allowing a simultaneous analysis of several groups of students with respect to both measurement and structural properties in a single framework.

The generalized least squares estimator becomes computationally heavy with a large number of elements in σ . Exceeding much beyond, say, 250 elements gives rise to unreasonable computing demands both in terms of storage and time. While an unweighted least squares estimator, using W=I, presumably can handle at least twice this number, it would not give a chi-square model test, nor would standard errors be provided. A simultaneous multiple-group analysis would normally involve all three parts of the model. However, in a single group analysis the σ_2 and σ_3 part of the model need only be used, since such a model does not impose restrictions on σ_1. With p denoting the number of y variables and q denoting the number of x variables, there are pq elements in σ_2 and p(p-1)/2 elements in σ_3. While problems with p=5, q=30 and p=10, q=15 could easily be handled by the generalized least squares estimator, p=15 would restrict q to less than 10. Larger models could be handled by ignoring the restrictions imposed on the σ_3 part, which would use less information in the estimation but would give all the results needed. Here, p=20, q=10 could be handled with somewhat heavy but not excessive computations. In the analyses of Section 5, a single group analysis using σ_2 and σ_3 was carried out with p=8 and q=24 and a multiple-group analysis of two groups with p=8 and q=14. While the multiple-group analysis involved modest computing, the single group analysis, using 224 σ elements, involved rather heavy but not excessive computing. Still, it is clear that the analysis proposed are best suited to the detailed scrutiny of a small set of items.

IV. The SIMS Data

To illustrate the methodology in a realistic setting, we will use data from the Second International Mathematics Study (Crosswhite, Dossey, Swafford, Mcknight, & Cooney, 1985). We will be concerned with a subset of data from the population of U.S. eighth grade students enrolled in regular mathematics classes. A national probability sample of school districts was selected proportional to size; a probability sample of schools were selected proportional to size within school district; and two classes were randomly selected within each school yielding a total of about 280 schools and about 7,000 students measured at the end of Spring 1982.

The achievement test contained 180 items in the areas of arithmetic, algebra, geometry, probability and statistics, and measurement distributed among five test forms. Each student responded to a core test (40 items) and one of four randomly assigned rotated forms (34 or 35 items). All items were presented in a five category multiple choice format. In Section 6 our analysis will not include probability and statistics and will only use the core items within the other areas, 8 each for algebra, geometry, and measurement, and 16 for arithmetic. In this chapter, the responses to the eight algebra items will be of particular interest.

The instructional coverage of algebra, and the mathematics curriculum in general, is rather varied for U.S. 13 year olds. Hence, to complement the item response information for these algebra items, we will utilize a class-level variable which categorizes the mathematics classes into four types, basic or remedial arithmetic (REMEDIAL), general or typical mathematics (TYPICAL), pre-algebra or enriched (ENRICHED), and algebra (ALGEBRA). Furthermore, we will check the plausibility of our analyses by drawing from class-level, item-specific, information on teacher reports of Opportunity To Learn (OTL), where a student is regarded as having OTL if the teacher taught or reviewed the mathematics needed to answer the item correctly either during this year or prior school years.

The responses to the SIMS items discussed above were collected at the end of the eighth grade. The achievement level obtained by the student on the various aspects of the mathematics content has at that point of time been influenced by factors such as the type and amount of instruction given during the school year, initial aptitude, motivation, and interest in the topic, and a variety of socio-demographic and other variables. Regarding algebra achievement, the outcome should be strongly related to the type of class attended, since in the eighth grade the content of the algebra test would usually only be well covered in the enriched (prealgebra) or algebra classes. To a certain extent, selection into such classes takes place based on the student's seventh grade scholastic performance in mathematics, particularly the central topic of arithmetic. The participation in eighth grade algebra classes may have important consequences since this allows students to take calculus in high school, which in turn opens up possibilities to study science and mathematics topics in colleges and universities (see also Kifer, 1984).

Much could be learned if student post-test performance could be related to the mathematics course taken and to student characteristics as they entered the course. With the SIMS data we are in the fortunate position of having available a set of such external measurements from the beginning of the eighth grade. Fall 1981 "pre-test" data was gathered for a large portion of the "post-test" students measured in the spring of 1982. We will use this additional data to study both the algebra post-test item responses and a set of external variables in the framework of a model that relates the post-test algebra achievement to pre-test predictors. These additional pieces of background data will now be briefly described.

The pre-test data were gathered in the same way as the post-test data. The new set of variables to be used in our model in addition to the post-test algebra items includes pre-test scores on the core items of algebra, measurement, geometry, and arithmetic, measurements of father's and mother's education, father's occupation, ethnicity, gender, attitude measurements describing the student's interest in more education, how useful he or she thinks mathematics knowledge will be, and his or her attraction to mathematics, and finally information on class type. The measurement and scoring of these background variables is described in Table 1. The abbreviations of Table 1 will be used from now on. It is important to note that some of the variables were measured only at the post-test occasion, particularly MORED, USEFUL, ATTRACT. These three measured were taken from Delandshere (1986).

Insert Table 1 about here

The wording of the eight post-test algebra core items is given in Table 2.

Insert Table 2 about here

The sample used for analysis is the match between post- and pre-test students that have complete data on all variables except father's occupation. For this variable there was unfortunately a large portion of missing data and it was decided to retain such observations by including missing data as a special category, in addition to the dummy coded categories Low, Middle, and High. The analysis sample is, however, only a subset of the two pre- and post-test data sets and in order to judge the effects of the missing data, Table 3 gives descriptive statistics for relevant variables form each of the three data sets. For purposes of simplifying the analyses, the variables have all been transformed to a 0-1 range. The analysis sample has somewhat higher means than the other samples both on variables thought to be positively correlated with achievement and on post-test algebra performance.

Insert Table 3 about here

Although not included directly in our analysis in Section 6, we will also utilize the item-specific OTL measurements on the post-test algebra items in order to enhance our understanding of the analysis. The upper panel of Table 4 gives the percent correct on each item broken down by class type, while the bottom panel gives the corresponding OTL means.

Insert Table 4 about here

V. Analysis of the SIMS Data by a Structural Model

Let us now analyze the SIMS data using the modeling framework presented in Sections 2 and 3. It may be noted that the proposed analyses can not be handled by present IRT software, nor by present structural equation modeling software, such as LISREL. The estimation and testing of the models to be presented was carried out by an experimental version of the LISCOMP computer program (Analysis of Linear Structural Equations by a Comprehensive Measurement model), developed by the author. The program provides limited information generalized least-squares estimation of the model parameters as they appear in the three-part structure of Section 3. Standard errors of estimates and a large-sample chi-square test of fit to the restrictions on the three model parts are also provided.

We consider the MIMIC model of Figure 1. The $\mathbf{y}$ vector of response items correspond to the eight items of Table 4. The $\mathbf{x}$ vector of regressors consists of the 17 background variables given in Table 1: PREALG, PREMEAS, PREGEOM, PREARITH, FAED, MOED MORED, USEFUL, ATTRACT, NONWHITE, REMEDIAL, ENRICHED, ALGEBRA, FEMALE, LOWOCC, HIGHOCC, MISSOCC, and seven interaction terms, between NONWHITE and the three class type dummies, between PREARITH and the class type dummies, and between NONWHITE and PREARITH. In a preliminary analysis we also included interactions between sex, PREALG, and the class type dummies, but these were not found significant. The latent variable construct, post-test algebra achievement as measured by the core items, is viewed as an intervening variable in the regressions of the y's on the x's.

We have attempted to use a large set of regressors which also contains some variables that may not have a direct substantive influence on the latent variable construct. This was done for two reasons. One reason relates to the fact that our analysis sample was obtained by "list-wise deletion" of incomplete cases where judging from Table 3 the missingness appeared to be somewhat selective. If the missingness on the y's can be largely predicted by the included x's, the bias that could potentially have resulted in the parameters of the regressions may be small (c.f. Marini, Olsen, & Rubin, 1980). A second reason is related to the fact that we will also study subgroups of students in certain class types, which will involve the analysis of selective samples. For instance, Kifer (1984) noted that whites are overrepresented in algebra courses, and also that "...almost 2/3 of the students in algebra classes have pretest arithmetic scores in the top quarter of the distribution", while "...almost 2/3 of the students whose pretest arithmetic scores are in the top quarter are not in algebra classes." Hence, we have included various interaction terms among the x's involving Ethnicity, Class type, and pretest arithmetic score, again to reduce potential bias. Furthermore, Muthen (1986) found that in addition to pretest scores and demographic variables class type membership was also strongly related to the attitude variables ATTRACT and MORED.

Section V.1 deals with certain weaknesses in the actual data analysis. The reader who merely wants to view the analysis as illustrations of the potential of the new type of modeling may want to skip to section 5.2.

V.1 Analysis caveats

We may recognize some weaknesses in the forthcoming analyses related to the sampling, the temporal ordering of the variables, and the potential of measurement error and omitted variables in the set of x's, problems which may cause bias in the regressions. First of all, our analyses ignore the complications of stratified sampling and multilevel, hierarchical observations. Although we realize that these features may have non-negligible consequences, the proper methods for handling them are not available in this context. Second, the attitudinal measurements MORED, USEFUL, and ATTRACT were obtained only at the post-test occasion, causing a possible problem if attempting to view these regressors as both predictors of entrance into advanced eighth grade classes and post-test achievement. These scores presumable reflect attitudes built up both before and during the eighth grade, although they are most likely not a direct reflection of the post-test performance. Furthermore, the pre-test scores are created from a small number of items, giving rise to low reliability. Although the rotated form items could have been used, this was avoided since it would have either involved equating of observed scores or using IRT techniques with sets of items many of which may have low validity at the pre-test due to rather limited OTL. For the 16 pre-test arithmetic items, an attempt was made to avoid the influence of measurement error by instead using factor scores. These were obtained in the form of estimated θ values from a marginal maximum likelihood estimation (see Bock & Aitkin, 1981) of the 16 items with a three parameter logistic model using the computer program BILOG (Mislevy & Bock, 1984). Although reduction of measurement error would have been even more desirable for the other subtests, which involve fewer items, it was judged that the small number of items and the heterogeneous OTL measures for these subtests might not yield reliable results by IRT methods. For algebra and measurement, one item each was rejected as invalid in relation to the total 40 item score. This results in "favoring" the variable PREARITH in the search for influential regressors. However, it was thought to be important to try to measure this variable well since it may be viewed as a proxy for final seventh grade mathematics achievement, which is an important factor in deciding eighth grade curriculum.

A further measurement flaw included a 40% missingness on Father's occupation. We should also note that the Ethnicity category NONWHITE is a very heterogeneous group consisting of 741 students, broken down as 8% American Indians, 41% Blacks, 17% Chicano, 6% Latin, 9% Oriental, and 19% Other. In terms of omitted variables, parental income may be a predictor of class type but was not measured, and it would have been very valuable if more general ability measures had been available before entrance into the eighth grade instead of merely fall pretest scores. Also, measures of reading comprehension and vocabulary would have been of interest since they might play a role in "word problems".

Preliminary analyses were carried out on the post-test response items in order to investigate the presence of guessing (or non-zero lower item characteristic curve asymptote) and/or violations of unidimensionality in the algebra items. Marginal maximum likelihood estimation of the two and three parameter logistic IRT models was carried out in BILOG and unidimensionality was tested both via LISCOMP's limited information GLS procedure and via the full information estimation procedure of TESTFACT (Wilson, Wood,

& Gibbons, 1984; see also Bock, Gibbons, & Muraki, 1985), in both cases assuming zero lower asymptotes. While unidimensionality could not be rejected using these approaches, the likelihood ratio chi-square test of zero lower asymptotes obtained a value of 46 with 8 degrees of freedom. Although the large sample size of 4,320 yields a strong power for rejection and lower asymptotes may not be well estimated from such small number of items, there seems to be a possibility of some nonzero asymptotes. The influence of this on our two-parameter model would presumably be a slight underestimation of the corresponding slope (loading) and a biasing of the threshold, while structural parameters may be relatively unchanged. Anticipating the analysis discussion below, it is interesting to note that neither the difficult Item 1 nor Item 5 exhibits significant asymptotes, either when analyzing the 8 algebra items alone or together with the other core items in a 40 item analysis (39 items were actually used due to one flawed item).

V.2 A structural model for all students: Model 1

In the first step of the analysis we will consider the strongest and most restrictive model, where achievement is viewed as a unidimensional construct, so that a single latent variable intervenes in the regressions of the y's on the x's, without any direct regression paths from x's to y's. This model will be called Model I. It should be noted that in this first step of the analysis, the categorical grouping variables of class type, gender, and ethnicity are included as dummy coded variables among the set of x's. Our intention is to let the analysis of Model I, and modifications thereof, assist in generating ideas for subsequent simultaneous multiple-group analyses, where the grouping is based on such categorical variables, and where a more detailed analysis is possible. For our first analysis of the whole analysis sample of 4,320 students, the complete set of assumptions in Model I may not be entirely realistic, since we include all the different types of eighth grade classes, while Table 4 clearly shows that percent correct and OTL varies greatly and in different patterns for different items over these classes. Nevertheless, this may be useful starting point for our analysis.

Model I is an overidentified model, which imposes 188 restrictions on the reduced form regression slopes and residual correlations. The standard IRT unidimensionality assumption with conditional independence contributes 20 restrictions, since 28 reduced form residual correlations are described by 8 parameters related to the measurement part. The concept of an intervening latent variable construct in the regressions of the y's on the x's contributes the remaining 168 restrictions, since 192 reduced form regression slopes are described by merely 24 structural regression slope parameters. Hence, in terms of restrictions imposed, the content of the model is largely a result of using the external variables of x and imposing MIMIC restrictions on the regression slopes for y on x. Utilizing external variables in this way gives a more powerful assessment of measurement qualities for the y's than would be obtained by considering responses to the y's alone as in standard IRT.

The large-sample chi-square test of fit to the 188 restrictions of Model I obtained a value of 681. This represents a significantly misfitting model. However, given the power resulting from the large sample size of 4,320, the value is in our opinion small enough to warrant attempts of modifying details of this first approximation rather than rejecting it in its entirety. Throughout, we will use the chi-square test results more as descriptive measures of overall fit for a sequence of models fitted to the same data than as a rigorous hypothesis testing instrument. In terms of such a descriptive useage, some experience with structural models for dichotomous response data lead us to judge as reasonable fit a chi-square to degrees of freedom ratio scaled to a sample size of 2,000 that is less than say 1.5 (this ratio is 1.7 for Model I). We know that there may be clear substantive reasons for lack of fit in parts of Model I and we will not be satisfied with the model as it stands, but investigate the possible reasons for misfit in an attempt to arrive at a modified Model II.

The fact that Model I is strongly over-identified offers the opportunity to check the appropriateness of the various assumptions involved and to relax some restrictions if judged necessary. This would not be possible in a straightforward multivariate regression of the y's on the x's, but is the result of our notion of a single latent construct. To aid in attempts to check the fit of the various restrictions, so called modification indices will be used. They are similar to what is provided in the LISREL structural equation modeling program (Joreskog & Sorbom, 1984). Such an index reflects the expected improvement in fit if a restricted parameter, such as one set to zero, is allowed to be freely estimated. The indices to be used in this version of LISCOMP are not scaled to represent the chi-square metric as in LISREL, but are merely the first-order derivatives of the parameters. It should be noted that the use of these modification indices as a data exploration device may be dangerous. The information from the various indices for a certain model can be misleading since they may be highly correlated, the information really only pertains to freeing up one parameter at a time, the indices are only good approximations for models that are close to a well fitting one, and we may capitalize on chance in our data. Below, we will try to use these indices with care in conjunction with substantive considerations.

The modification indices for Model I are given in Table 5 below. The indices in the top part of the table gives information on which direct paths from x's to y's may need to be freed from their restriction to zero. These paths correspond to the broken arrows of Figure 1. The indices in the bottom part of the table gives information on potential violations of the conditional independence assumption of zero correlations among the residuals. In this table, the first-order derivative modification indices have reversed signs so that the present sign describes the expected direction of change from zero in a parameter. The derivative values have also been divided by 10 and rounded.

Insert Table 5 about here

Scrutinizing Table 5 in conjunction with other substantive information will lead us to Model II. Let us only consider the three largest modification indices for Model I, marked by asterisks in Table 5. Starting with Item 1's index of 17 for the ALGEBRA class dummy (comparing to the category of Typical classes), we have an indication of a positive direct "effect" of membership in algebra classes on the performance on Item 1 (c.f. Muthen, 1986). It should be kept in mind that this direct influence occurs over and above the influence of the latent achievement construct on Item 1. This implies that students with the same algebra achievement level, but belonging to different class types, may perform differently on Item 1; algebra class membership gives an advantage. Hence, we have a suggestion of "item bias", or rather instructional sensitivity in Item 1. This empirical suggestion makes substantive sense when we consider our auxiliary information. This is the only one of the algebra items that deals explicitly with "solving for x". Table 4 shows that this is the hardest of the eight items, with a large difference in proportion correct between students of typical and algebra classes, and with the largest difference in OTL between typical and algebra classes. From Table 4 we see that Items 6 and 7 have somewhat similar features, but none of these items exhibit large ALGEBRA modification indices in Table 4. It seems as if in this set of items the lack of instructional coverage in typical classes has a particularly detrimental effect on the response to Item 1.

The largest modification index for direct x to y paths in Table 5 occurs for Item 5 on the dummy variable FEMALE. This suggests a gender item bias. The negative sign would imply that, for given achievement level, Females perform worse on Item 5 than Males. We may note that this item involves a "word problem" in a way the other items do not. This potential gender difference will be further analyzed below. The largest modification index in Table 5 occurs for a correction between the measurement error of Item 6 and 8, suggesting a violation of the conditional independence for these two items in the form of a positive correlation. From the item wording of Table 2 we do in fact note that both items, and none of the others, involve a direct translation of a word problem into a mathematical formula. Hence it is possible that the correlation may indicate the presence of a specific skill, in addition to the algebra achievement construct, required for such a translation.

V.3 A structural model for all students: Model II

Let us now free up the above three parameters that were fixed to zero in Model I and consider the modified Model II. This model obtained a chi-square value of 441.59 with 185 degrees of freedom. The difference in chi-square from Model I is 240 with 3 degrees of freedom. Given the sample size we regard this outcome as an indication of a reasonable overall fit in the major parts of the model, although further adjustments could be made. Some interesting details may be noted before we consider the estimates of Model II. First, in this case the freeing up one of the three parameters at a time would by use of the largest modification indices lead to the same final result, irrespective of the order in which this was done. Second, the major results in terms of general magnitude and significance of structural coefficients remain largely unchanged when going from Model I to Model II. Third, for Model II the modification index for PREARITH x ALG has been reduced to almost zero from the Model I value of 9, the Model I value of 10 for item 8 on FEMALE has only been reduced to 8, the Model I value of 8 for Item 3 on ENRICH remains the same , and the Model I value of -8 for Item 5 on NONWHITE also remains the same. The remaining major modification indices now appear among the error correlations with a few values of about 10.

The parameter estimates for Model II are given in Table 6, where the first part of the table gives measurement parameter results and the second part gives results on structural parameters. For the measurement part we also give estimated reliabilities for each item.

Insert Table 6 about here

The estimated reliabilities are in some cases rather low, although we must bear in mind that these are item level responses. Since Item 1 and 5 are directly related to both the latent construct to be measured and one of the regressors, these two items, in relation to the other items in the set, are not homogeneous with respect to the set of regressors (c.f. Muthen, 1985). Regarding the structural parameter estimates, we find expected strong, significant influences on achievement from PREARITH and PREALG, and the other pretest scores, but also from USEFUL, ALGEBRA, FEMALE, and HIGHOCC. The significance of the last three dummy variables implies that given other regressor values being equal, membership in advanced classes rather than typical ones, being female, and having a father in the high occupation category rather than the middle one, are conditions associated with a higher level of algebra achievement as represented by the latent variable construct.

In addition to this, we find from the bottom of Table 6 that for a given value of the achievement construct, membership in algebra classes and being female, respectively, is associated with a higher level of performance on item 1 and a lower performance on item 5, respectively. From the estimated parameters and the sample mean vector and covariance matrix for x, we may also calculate the mean and variance of the latent variable construct and the proportion of variation in this construct that is accounted for by the set of regressors. We obtained a mean of 2.20, a standard deviation of 0.87, and 73% of the variation was accounted for. Using the mean and standard deviation we can translate the measurement parameter estimates to standard IRT a and b values on a $0,1,\theta$ scale (see below in relation with Table 8).

V.4 A simultaneous structural analysis of males and females in typical classes

In Muthen (1986), the above analysis is taken further by considering class type differences. Here, we will instead study in more detail the differences and similarities in measurement and structural parameters across gender. A simultaneous, two-group analysis will be carried out for students of typical classes. In these models, 14 x variables from the original set remain after eliminating class type and gender related dummies. Table 7 gives descriptive statistics for these regressors. We note that

---

Insert Table 7 about here

---

that Males have slightly higher means on variables associated with high achievement, except for USEFUL. The proportion correct for the post-test algebra items in typical classes were for Males: 0.14, 0.65, 0.50, 0.40, 0.47, 0.51, 0.37, 0.50, and for Females: 0.14, 0.69, 0.50, 0.46, 0.38, 0.53, 0.35, 0.56. The OTL values are given in Table 4 and do not vary appreciable over gender.

In the multiple-group analysis the effect of gender can be studied in more detail than was possible in the single-group analysis of Model II. In Model II, gender differences were only captured the intercepts of the achievement and the latent response variable regressions. Although interaction terms between gender and other regressors in Model II could have been accommodated in the achievement construct relation, the dummy variable approach would not for instance be able to handle gender differences in measurement slopes (loadings). Also, in a multiple-group analysis it is easier to separately deal with tests of invariance in the measurement and the structural part.

In this analysis we will apply a multiple-group version of the Figure 1 MIMIC model. Since the same measurement instrument was used for the two sexes, we will test the notion of invariance in the measurement thresholds and slopes (loadings) for the eight response items, allowing all other parameters to differ across the two groups. Based on the previous analysis results for all students, we will however allow the threshold and slope of Item 5 to vary. As a base line model we will first consider a multiple group analysis of males and females where no parameters are invariant, in order to assess the appropriateness of the MIMIC model itself. With 236 degrees of freedom, this resulted in a chi-square value of model fit of 366. This fit is judged as satisfactory. The total sample size is 2,417 broken down as 1,150 males and 1,267 females.

The addition of invariance of measurement intercepts and slopes, except for Item 5, resulted in a chi-square value of 381 with 248 degrees of freedom, yielding a non-significant chi-square increase of 15 with 12 degrees of freedom compared to the base-line model. Also adding invariance for Item 5, however, resulted in a chi-square difference test value of 33 with 2 degrees of freedom. This strong rejection of the invariance notion for Item 5 is in line with our single-group results for Model II in all class types. The parameter estimates for the multiple-group model of invariant measurement threshold and slopes, except for Item 5, is given in Table 8.

Insert Table 8 about here

From the measurement part of Table 8 we see that Item 1 has the lowest correlation with the latent achievement construct. This is in line with the low OTL value of 50% in Table 4. For Item 5, the gender difference in thresholds and loadings translates into (see Muthen & Christofffersson, 1981, equations 28 and 29) a two-parameter normal ogive a (discrimination) and b (difficulty) value on a 0,1, θ - metric of 0.81 and 0.09 for males and 0.65 and 0.51 for females. Hence, the male item characteristic curve is shifted to the left from the female curve and is steeper, thereby favoring males. The reason for this gender difference is, however, unclear. The availability of further external variables such as a reading comprehension test might possibly have been able to shed light on this matter (c.f. Muthen, 1985).

Regarding the structural slopes, the results are rather similar to those for all students in Model II of Table 6. In the present model the intercept difference in the structural relation for the latent variable construct is not significantly different from zero. However, estimating the construct mean from the estimated coefficients and the sample mean vector for the x's, we find a value of 1.81 for males while females obtain 1.88. This difference should be viewed in relation to the male standard deviation of 0.67 and the female standard deviation of 0.63. Although males seemed to have slightly higher means on important regressors in Table 7, females end up with a higher post-test achievement level. The proportion of variation in the construct accounted for by the x's is 66% for males and 68% for females.

In addition to imposing restrictions of measurement parameter invariance, it is also of interest to study the differences in the structural parameters across gender. For instance, are the possible higher level of the achievement construct for females due to the fact that females have higher slopes on important regressors (the important variable USEFUL would however be an important exception)? Adding the restriction of invariant structural slopes, yields a chi-square difference of 29 with 14 degrees of freedom, while restricting only the slopes for PREARITH to be equal across sex yields a chi-square difference value of 2 with 1 degree of freedom. There seems to be some evidence of differences in some of the slopes, although PREARTH seems to have equal predictive strength for the two sexes.

VI. Conclusions

The MIMIC structural modeling approach was found to be quite useful with the present data where there was a particular interest in post-test responses and where pre-test data was available. Using a single model framework that extends the boundaries of IRT, we were able to simultaneously deal not only with issues of measurement qualities, but also differential item performance in different subgroups and differential prediction of achievement.

Other versions of the general model of Section 3 would be relevant in other situations. The external x variables need not only appear as background variables, predicting the dichotomous y's. For instance, we may be interested in the differential predictive validity in different groups of a set of items or subtest scores for which certain constructs are hypothesized. Here, careful measurement modeling carried out on the exogenous side may lead to better predictions of a certain y criterion. The use of structural modeling in such situations does not seem to have been fully explored.

REFERENCES

Beck, R.D., & Aitkin N. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika, 46, 443-459.

Bock, R.D., Gibbons, R., & Muraki, E. (1985). Full-information item factor analysis. Final report to the ONR. National Opinion Research Center, MRC Report #85-1.

Bohrnstedt, G.W. (1983). Measurement. Handbook of Survey Research, 69-121. Academic Press.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. Psychometrika, 40, 5-32.

Crosswhite, F.J., Dossey, J.A., Swafford, J.O., McKnight, C.C., & Cooney, T.J. (1985). Second International Mathematics Study Summary Report for the United States. Champaign, Ill.: Stipes.

Delandshere, G. (1986). The effect of teaching practices on math achievement in the eighth grade. Unpublished doctoral dissertation, University of California, Los Angeles, in progress.

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

Joreskog, K.G. (1973). A general method for estimating a linear structural equation system. In A.S. Goldberger and O.D. Duncan (Eds.), Structural equation models in the social sciences. New York: Seminar Press, 85-112.

Joreskog, K.G. (1977). Structural equation models in the social sciences: Specification, estimation and testing. In P.R. Krishnaiah (Ed.), Applications of statistics. Amsterdam: North-Holland.

Joreskog, K.G., & Goldberger, A.S. (1975). Estimation of a Model with Multiple Indicators and Multiple Causes of a Single Latent Variable. Journal of the American Statistical Association, 70, 631-639.

Joreskog, K.C., & Sorbom, D. (1984). LISREL VI; Analysis of linear structural relationships by maximum likelihood and least squares methods. Scientific Software, Inc.

Kifer, E. (1984). Issues and Implications of Differentiated Curriculum in the Eighth Grade. National Conference on the Teaching and Learning of Mathematics in the United States.

Marini, M.M., Olsen, A.R., & Rubin, D.B. (1980). Maximum likelihood estimation in panel studies with missing data. In Sociological Methodology. San Francisco: Jossey Bass.

Miller, M.D., & Linn, R.L. (1986). Invariance of Item Parameters with Variations in Instructional Coverage. Journal of Educational Measurement, forthcoming.

Mislevy, R.J. (1986). Recent developments in the factor analysis of categorical variables. Journal of Educational Statistics, 11(1), 3-31.

Mislevy, R.J., & Bock, R.D. (1984). BILOG: Marginal Estimation of Item Parameters and Subject Ability under Binary Logistic Models. Chicago: International Educational Services.

Muthen, B. (1978). Contributions to factor analysis of dichotomous variables. Psychometrika, 43, 551-560.

Muthen, B. (1979). A structural probit model with latent variables, Journal of the American Statistical Association, 74, 807-811.

Muthen, B. (1981). Factor analysis of dichotomous variables: American attitudes toward abortion. In D.J. Jackson and E.F. Borgatta (Eds.), Factor analysis and measurement in sociological research: A multidimensional perspective. London: Sage Publications.

Muthen, B. (1983). Latent variable structural equation modeling with categorical data. Journal of Econometrics, 22, 43-65.

Muthen, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. Psychometrika, 49, 115-132.

Muthen, B. (1985). A method for studying the homogeneity of test items with respect to other relevant variables. Journal of Educational Statistics, 10(2), 121-132.

Muthen, B. (1986). Instructionally sensitive psychometrics: Applying structural models to educational achievement data. In preparation.

Muthen, B., & Christoffersson, A. (1981). Simultaneous factor analysis of dichotomous variables in several groups. Psychometrika, 46, 485-500.

Muthen, B., & Joreskog, K. (1983). Selectivity problems in quasi-experimental studies. Evaluation Review, 7, 139-173.

Muthen, B., & Lehman, J. (1985). Latent variable probit ANCOVA: Treatment effects in the California civil addict programme. University of California, Los Angeles. British Journal of Mathematical and Statistical Psychology, 27, 229-239.

Muthen, B., & Speckart, C. (1985). Multiple group IRT modeling: Applications to item bias analysis. Journal of Educational Statistics, 10(2), 133-142.

Schmitt, N., & Stults, D.M. (1986). Methodology Review: Analysis of Multitrait-Multimethod Matrices. Applied Psychological Measurement, 10(1), 1-22.

Sorbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. British Journal of Mathematical and Statistical Psychology, 27, 229-239.

Sorbom, D. (1978). An alternative to the methodology for analysis of covariance. Psychometrika, 43, 381-396.

Sorbom, D. (1982). Structural equation models with structural means. In K.G. Joreskog and H. Wold (Eds.), Systems under indirect observation: Causality, structure, prediction. Amsterdam: North-Holland.

Wilson, D., Wood, R., & Gibbons, R. (1984). TESTFACT: Test scoring and item factor analysis. Scientific Software, Inc.

TABLE 1

Description of External Variables

PREALG	:	Proportion of correct reponses on seven pre-test core items.
PREMEAS	:	Proportion of correct responses on seven pre-test core items.
PREGEOM	:	Proportion of correct responses on eight pre-test core items.
PREARITH	:	Estimated pre-test theta based on the three-parameter logistic model using 16 items.
FAED	:	The highest type school attended by father or male guardian. 1 = very little schooling, or no schooling at all 2 = primary school 3 = secondary school 4 = college, university or some form of tertiary education
MOED	:	As in FAED, but for respondent's mother or female guardian.
MORED	:	Responses to the question "After this year, how many more years of full-time (including university, college, etc.) education do you expect or plan to complete?" 1 = none at all (0 years) 2 = up to 2 years 3 = more than 2 years - up to 5 years 4 = more than 5 years - up to 8 years 5 = more than 8 years
USEFUL	:	Average score of four attitude items scored: Strongly disagree (1), Disagree (2), Undecided (3), Agree (4), and Strongly agree (5). These items are: 1. I can get along well in everyday life without using mathematics (Reversed).

2. A knowledge of mathematics is not necessary in most of occupations (Reversed).
3. Mathmatics is not needed in every day living (Reversed).
4. Most people do not use mathematics in their jobs (Reversed).

ATTRACT : Average scores of five attitude items. Scoring is in the same way as for USEFUL and the items are:

1. I would like to work at a job that lets me use mathematics.
2. I think mathematics is fun.
3. Working with numbers makes me happy.
4. I am looking forward to taking more mathematics.
5. I refuse to spend a lot of my own time doing mathematics (Reversed).

TABLE 1 Con't.

Description of External Variables

Ethnicity dummy coding (0 = White): 1	NONWHITE
Class type dummy coding (0 = Typical class):	
	REMEDIAL
	ENRICHED
	ALGEBRA
Gender dummy coding (0 = Male):	
	FEMALE
Father's occupation dummy coding (0 = Middle): 2	
	LOWOCC
	HIGHOCC
	MISSOCC

Notes:

1. The non-white category consists of American Indian, Black, Chicano, Latin, Oriental, and Other.

2. The LOWOCC category of Father's occupation consists of the classifications Unskilled and Semi-skilled worker, the Middle category consists of Skilled worker, clerical, sales and related, the HIGHOCC category consists of Professional and Managerial, and the MISSOCC category consists of no response and unclassifiable response.

TABLE 2

Wording for Eight Post-test Algebra Core Items

---

1. If $5x + 4 = 4x - 31$, then x is equal to

   A -35
   B -27
   C 3
   D 27
   E 35

2. If P = LW and if P = 12 and L = 3, then W is equal to

   A 3/4
   B 3
   C 4
   D 12
   E 36

3. (-2) x (-3) is equal to

   A -6
   B -5
   C -1
   D 5
   E 6

4. If $4x/12 = 0$, then x is equal to

   A 0
   B 3
   C 8
   D 12
   E 16

5. The air temperature at the foot of a mountain is 31 degrees. On top of the mountain the temperature is -7 degrees. How much warmer is the air at the foot of the mountain?

   A -38 degrees
   B -24 degrees
   C 7 degrees
   D 24 degrees
   E 38 degrees

6. A shopkeeper has x kg of tea in stock. He sells 15 kg and then receives a new lot weighing 2y kg. What weight of tea does he now have?

   A $x - 15 - 2y$
   B $x + 15 + 2y$
   C $x - 15 + 2y$
   D $x + 15 - 2y$
   E None of these

TABLE 2 Con't.

Wording for Eight Post-test Algebra Core Items

7. The table below compares the height from which a ball is dropped (d) and the height to which it bounces (b).

d	50	80	100	150
b	25	40	50	75

Which formula describes this relationship?

A $b = \frac{d}{2}$
B $b = 2d$
C $b = d/2$
D $b = d + 25$
E $b = d - 25$

8. The sentence "a number x decreased by 6 is less than 12" can be written as the inequality

A $x - 6 > 12$
B $x - 6 \geq 12$
C $x - 6 < 12$
D $6 - x \geq 12$
E $6 - x < 12$

TABLE 3

Descriptive Statistics for the Different SIMS Samples

	Pre-test Sample (N = 6517)			Post-test Sample (N= 7248)			Analysis Sample (N = 4320)		
	Mean	S.D	N	Mean	S.D	N	Mean	S.D.	N
PREALG	0.40	0.25	6353	—	—	—	0.43	0.26	4320
PREMEAS	0.49	0.25	6353	—	—	—	0.51	0.24	4320
PREGEOM	0.33	0.23	6353	—	—	—	0.35	0.23	4320
PREARITH (OBS. SCORE)	0.39	0.23	6353	—	—	—	0.52	0.26	4320
PREARITH (THETA SC.)	—	—	—	—	—	—	0.40	0.18	4320
FAED	—	—	—	0.80	0.24	6831	0.82	0.23	4320
MOED	—	—	—	0.79	0.22	6879	0.80	0.21	4320
MORED	—	—	—	0.75	0.20	6931	0.77	0.19	4320
USEFUL	—	—	—	0.71	0.19	6878	0.72	0.19	4320
ATTRACT	—	—	—	0.54	0.20	6856	0.54	0.20	4320
NONWHITE	—	—	—	0.26	0.44	6694	0.22	0.41	4320
REMEDIAL	—	—	—	0.08	0.27	7248	0.07	0.25	4320
ENRICHED	—	—	—	0.22	0.41	7248	0.25	0.43	4320
ALGEBRA	—	—	—	0.13	0.34	7248	0.13	0.33	4320
FEMALE	—	—	—	0.52	0.50	7024	0.53	0.50	4320
LOWOCC	—	—	—	0.18	0.38	7248	0.18	0.39	4320
HIGHOCC	—	—	—	0.11	0.32	7248	0.13	0.33	4320
MISSOCC	—	—	—	0.42	0.49	7248	0.39	0.49	4320
POSTALG1	—	—	—	0.21	0.41	7013	0.22	0.41	4320
POSTALG2	—	—	—	0.69	0.46	7013	0.72	0.45	4320

Table 3 (Con't.)

POSTALG3	—	—	—	0.57	0.50	7013	0.58	0.49	4320
POSTALG4	—	—	—	0.49	0.50	7013	0.51	0.50	4320
POSTALG5	—	—	—	0.45	0.50	7013	0.47	0.50	4320
POSTALG6	—	—	—	0.55	0.50	7013	0.57	0.49	4320
POSTALG7	—	—	—	0.39	0.49	7013	0.40	0.49	4320
POSTALG8	—	—	—	0.56	0.50	7013	0.59	0.49	4320
ALG OTL%	—	—	—	0.71	0.26	6914	0.72	0.26	4224

TABLE 4

Proportion Correct and Opportunity to Learn (OTL)

Proportions for the Eight Post -Test Algebra Core Items

by Class Type

	Item							
	1	2	3	4	5	6	7	8
Class Type	Proportion Correct							
REMEDIAL	0.09	0.44	0.14	0.22	0.14	0.30	0.22	0.31
TYPICAL	0.14	0.67	0.50	0.43	0.42	0.52	0.36	0.53
ENRICHED	0.22	0.81	0.73	0.63	0.55	0.63	0.46	0.68
ALGEBRA	0.65	0.90	0.90	0.81	0.71	0.85	0.58	0.84
TOTAL	0.22	0.72	0.58	0.51	0.47	0.57	0.40	0.59
	OTL Proportion							
REMEDIAL	0.21	0.61	0.43	0.41	0.65	0.09	0.16	0.20
TYPICAL	0.50	0.85	0.97	0.76	0.93	0.40	0.38	0.64
ENRICHED	0.78	0.96	0.94	0.94	0.95	0.47	0.58	0.83
ALGEBRA	0.95	0.95	1.00	0.95	1.00	0.95	0.81	1.00
TOTAL	0.61	0.87	0.93	0.80	0.92	0.46	0.47	0.70

Table 4 (cont.)

Sample Size

REMEDIA L	TYPICAL	ENRICHED	ALGEBRA	TOTAL
299	2417	1061	543	4320

TABLE 5

Modification Indices for a Structural Model.

All Students. Model I (N = 4,320)

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8
Direct Relationships Between Items and Regressions								
PREALG	2	-1	-1	2	0	-1	0	-1
PREMEAS	-1	2	-2	-1	4	-2	1	-2
PREGEOM	1	0	-1	-3	1	-1	0	2
PREARITH	-1	1	-1	-1	3	-1	0	-1
FAED	-3	0	2	1	3	-3	-1	0
MOED	-1	0	1	1	3	0	-1	-3
MORED	0	-1	1	0	-1	1	0	-2
USEFUL	-1	1	0	0	-2	1	-1	1
ATTRACT	3	1	-1	0	-2	1	0	-1
NONWHITE	3	-4	1	5	-8	3	2	-1
REMEDIAL	3	1	-5	0	-1	2	2	0
ENRICHED	-9	3	8	5	-4	-7	-1	2
ALGEBRA	*17	-4	1	0	-6	-2	-4	0
FEMALE	0	5	3	4	*-19	0	-5	10
LOWOCC	1	0	1	-1	-2	1	0	1
HIGHOCC	-1	1	-1	2	-2	-1	3	0
MISSOCC	1	-2	-1	3	1	0	-2	-1
NONW X REM	2	0	-3	1	0	0	0	0
NONW X ENR	0	-1	0	2	-3	0	1	2
NONW X ALG	2	-1	0	1	0	0	0	-1
PREARITH X REM	1	0	-1	0	0	0	0	0
PREARITH X ENR	-4	1	3	2	0	-3	0	0
PREARITH X ALG	9	-3	0	0	-3	-1	-2	0
NONW X PREARITH	1	-1	1	2	-2	1	0	0

Table 5 (Con't.)

Measurement Error Correlations

Item 2	5						
Item 3	2	-1					
Item 4	3	0	7				
Item 5	6	-12	7	-11			
Item 6	4	10	-4	4	5		
Item 7	1	2	-6	-5	13	2	
Item 8	4	-3	-8	2	-7	*21	9

* Freed parameter in Model II.

TABLE 6

Parameter Estimates for a Structural Model.
All Students. Model II (N = 4,320)

Measurement Parameter Estimates

Response Item	Thresholds Est.	Thresholds Est./S.E.	Loadings Est.	Loadings Est./S.E.	Reliabilities
Item 1	2.19	27	0.54	16	0.19
Item 2	1.23	14	0.88	22	0.41
Item 3	1.91	20	1.00 *	--	0.49
Item 4	1.76	20	0.82	23	0.37
Item 5	1.85	20	0.89	23	0.42
Item 6	1.59	19	0.82	22	0.37
Item 7	1.57	21	0.59	19	0.22
Item 8	1.34	17	0.73	21	0.32
Error correlation for Items 6 and 8			0.12	5	

* Parameter is fixed to set the metric of the latent variable construct.

Table 6 Cont'd.
Structural Parameters with the
Latent Construct as Dependent Variable

Regressor	Estimate	Estimate/S.E.
PREALG`	0.68	11
PREMEAS	0.45	7
PREGEOM	0.33	5
PREARITH	2.09	16
FAED	0.07	1
MOED	0.02	0
MORED	0.18	3
USEFUL	0.45	7
ATTRACT	0.04	1
NONWHITE	-0.02	0
REMEDIAL	0.07	1
ENRICHED	0.22	3
ALGEBRA	0.56	4
FEMALE	0.14	6
LOWOCC	0.02	1
HIGHOCC	0.12	3
MISSOCC	0.05	2
NONW X REM	0.10	1
NONW X ENR	0.19	3
NONW X ALG	-0.18	-1
PREARITH X REM	-1.45	-3
PREARITH X ENR	-0.10	-1
PREARITH X ALG	-0.54	-2
NONW X PREARITH	-0.19	-1
Item - Regressor Relations not Mediated by Latent Construct		
Item 1 on ALGEBRA	0.86	13
Item 5 on FEMALE	-0.35	-8
Latent Construct Residual Variance	0.20	13

TABLE 7

Means and Standard Deviations for Males and Females
in Typical Classes

	Male (N = 1,150)		Female (N = 1,267)	
Regressors	Mean	S.D.	Mean	S.D.
PREALG	0.38	0.23	0.37	0.23
PREMEAS	0.50	0.23	0.45	0.23
PREGEOM	0.33	0.22	0.29	0.19
PREARITH	0.37	0.17	0.36	0.15
FAED	0.81	0.23	0.79	0.23
MOED	0.80	0.20	0.78	0.21
MORED	0.74	0.20	0.74	0.19
USEFUL	0.69	0.19	0.73	0.17
ATTRACT	0.52	0.20	0.54	0.20
NONWHITE	0.21	0.41	0.23	0.42
LOWOCC	0.21	0.41	0.20	0.40
HIGHOCC	0.11	0.31	0.11	0.31
MISSOCC	0.37	0.48	0.40	0.49
NONW X PREARITH	0.06	0.13	0.07	0.14

Figure 1.

A MIMIC Structural Probit Model

TABLE 8

Parameter Estimates for a Simultaneous Structural Model Analysis of Males and Females in Typical Classes

Measurement Parameter Estimates
(Thresholds and loadings invariant over gender, except for item 5)

Response	Thresholds		Loadings		Reliabilities	
Item	Est.	Est./S.E.	Est.	Est./S.E.	Males	Females
Item 1	2.16	18.79	0.55	10.07	0.13	0.11
Item 2	1.50	9.58	1.09	14.66	0.39	0.36
Item 3	1.83	12.61	1.00*	---	0.35	0.31
Item 4	1.83	13.54	0.90	14.25	0.29	0.26
Item 5					0.40	0.30
Males	2.06	11.91	1.10	12.51		
Females	2.13	11.72	0.97	11.69		
Item 6	1.74	12.59	0.98	14.43	0.33	0.30
Item 7	1.71	14.45	0.72	12.22	0.20	0.18
Item 8	1.52	11.65	0.88	14.00	0.28	0.26

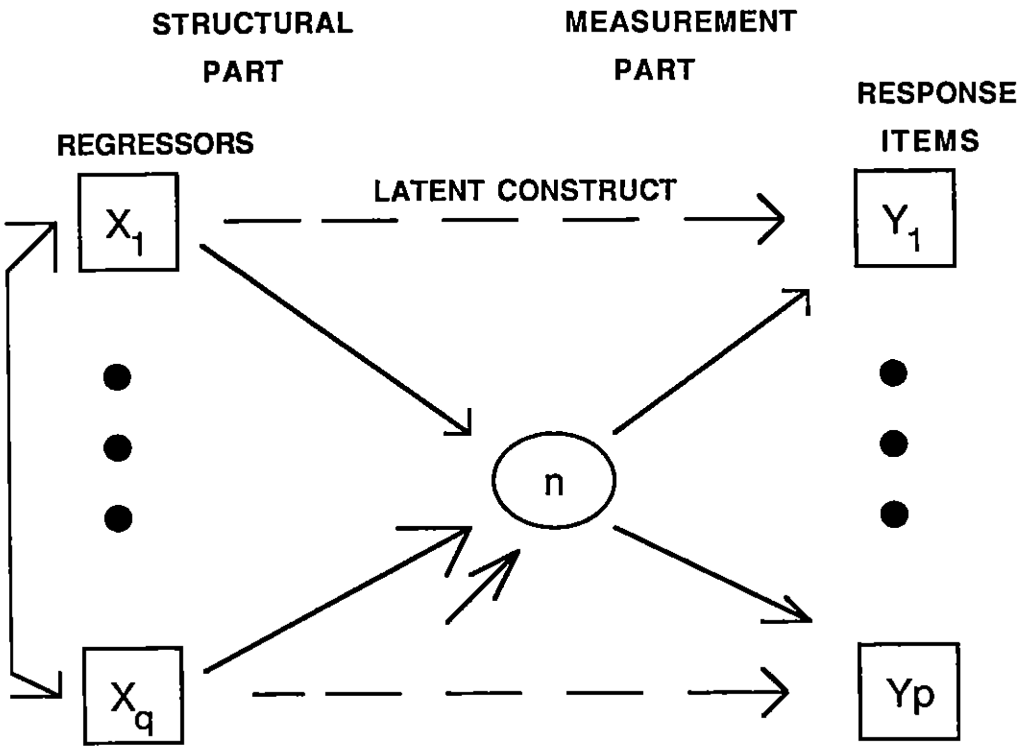
Structural Parameter Estimates

	Males (N= 1,150)		Females (N = 1,267)	
Regressors	Est.	Est./S.E.	Est.	Est./S.E.
PREALG	0.46	9	0.61	7
PREMEAS	0.51	5	0.46	5
PREGEOM	0.43	4	0.23	2
PREARITH	1.67	9	2.01	10
FAED	-0.12	-1	0.14	2
MOED	0.19	2	0.00	0
MORED	0.14	1	0.20	2
USEFUL	0.62	6	0.34	3

Table 8 (Con't.)

ATTRACT	-0.01	0	0.11	1
NONWHITE	0.10	1	0.02	0
LOWOCC	0.02	0	-0.03	-1
HIGHOCC	0.12	2	0.06	1
MISSOCC	0.12	3	-0.07	-2
NONW X PREARITH	-0.76	-3	-0.17	-1
Latent Construct Intercept	0.00 *	---	0.12	1
Latent Construct Residual	0.15	7	0.13	7

*Fixed parameter


STRUCTURAL
PART
MEASUREMENT
PART
RESPONSE
ITEMS
REGRESSORS
LATENT CONSTRUCT
X_1
Y_1
n
X_q
Yp