USING ITEM SPECIFIC INSTRUCTIONAL INFORMATION IN ACHIEVEMENT MODELING

Bengt Muthen

CSE Report No. 271

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

1987

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.

1. Introduction

In their recommendations for future extensions of Item Response Theory (IRT; see, e.g., Lord, 1980), Traub and Wolfe (1981) suggest that a "way of improving the linkage of achievement measurement and instruction is to obtain detailed data about the distribution (by student, class, school) of instruction and incorporate that information into the response model. We are thinking here of the so-called 'opportunity-to-learn' measures used in the IEA surveys and the exposure measures used, for example, by Fischer (1972)" (pp. 422-423). This paper represents an attempt in that direction.

Recently, one can observe an increasing concern about the match between the school curriculum and what is being tested by standardized achievement tests. See for instance Airasian and Madaus (1983), Haertle and Calfee (1983), Mehrens and Phillips (1986), Miller (1986). An interesting development is the use of "opportunity-to-learn" measures (Anderson, 1985), see e.g. Miller and Linn (1986) and Engelhard (1986).

This paper discusses methodological implications of utilizing instructional information in combination with the usual item responses. Models that expand those of standard Item Response Theory (IRT) will be considered. As an illustration, we will consider the achievement data of

the U.S. sample of the Second International Mathematics Study (SIMS), Crosswhite, Dossey, Swafford, McKnight, and Cooney (1985). In SIMS, opportunity-to-learn (OTL) response was gathered for both teachers and students corresponding to each item. Their response concerned whether or not the mathematics needed was covered during the present school year, and if not, whether it had been covered in prior years.

When analyzing achievement data from students with varying instructional background, the modeling should recognize that a heterogeneous population is at hand. Three methodological issues are of particular interest:

(1) How should the modeling capture the fact that the measurement relationship between the items and the latent trait may vary over students ?

(2) How should trait dimensionality be assessed ?

(3) How should the latent trait values be estimated ?

In IRT analyses of standardized achievement tests, it is assumed that instruction increases the item performance through an increase in the latent trait level, while the item-trait relationship remains the same; hence, no "item bias". This may be too strong an assumption when the instruction

is geared towards certain types of items in the test. If the assumption is incorrect and biased items are not removed , biased latent trait estimates are obtained. Furthermore, the factorial structure may not be the same for a group with high coverage as for a group with low coverage; indeed, with very low coverage, the validity of the item is called in question.

In this paper we will concentrate of the first question. It will be formulated as a problem of assessing "item bias", or instructional sensitivity in the items , when item specific OTL type information is available. In section 2 approaches that use traditional IRT methodology to assess measurement differences will be considered. Here, groups of students are formed based on OTL score and differences in item characteristic curves are assessed after equating one group's measurement parameters to the other group's. Section 3 demonstrates the weakness of this item bias approach by means of an artificial data analysis. In section 4 a new method is proposed which does not necessitate the creation of groups to assess item bias and avoids the problem of the standard item bias detection approach. This method generalizes IRT modeling to allow for item specific variation in measurement relations across students with varying instructional background (OTL). Item bias detection is obtained as a by-product. Section 5 applies the traditional and the new methods to the SIMS data and compares the outcomes in terms of item response curve bias.

2. Traditional item bias detection

With the availability of auxiliary information such as OTL, a conventional item bias type analysis is naturally of interest. Groups are formed based on some criterion related to OTL. IRT estimation is carried out in each group, the parameter estimates are made comparable by some form of equating, and some form of item bias index is calculated for each item (see e.g., Linn, Levine, Hastings, and Wardrop, 1981). In the present situation, however, it should be recognized that the auxiliary information of OTL is item specific. Whereas race or gender information puts a person in a group which is constant over the items, the OTL "group membership" varies with item. The problem is then how groups should be formed--or should they?

One approach to overcome the item specific nature of the OTL measure is to put in the same group students with the same OTL value ("Yes", "No", or graded in some way) for all items in some subset of the whole test and to study item response curve differences for this subset. Even with a very small number of items, however, this is likely to lead to groups with small sample sizes, jeopardizing the reliability of the IRT estimation. The analysis would also ignore a large portion of the total sample available.

Miller and Linn (1986) used an interesting alternative in analyzing SIMS data. Teacher reported OTL measures for each item were first

factor analyzed, and groups of students were then formed by cluster analysis based on the corresponding factor scores. The factors corresponded to groupings of items by content, such as the two for algebra: (1) formula, algebraic expressions, equalities and inequalities, and (2) signed number equations. This gave "curriculum clusters" that were well separated with respect to mean values of the factor variables. On comparison between item response curves for the same item across student groups, they found several items with considerable differences, i.e. evidence of "instructional bias" (see Linn & Harnisch, 1981). Miller and Linn (1986) noted that the magnitude of the biases were frequently larger than commonly eoncountered when considering student groups formed by ethnicity.

This approach has considerable merit and gave revealing results in terms of item response curve differences across clusters. It has, however, the drawback of basing the estimation of an item's parameters in a certain group (cluster) on students that may well have a wide range of OTL factor values. The approach also has the disadvantage that a student in any given group, who has a certain OTL status for an item, tends to have the same OTL status for several other items. To solve this item, the student may then be at an added advantage or disadvantage due to presence or absence of training in related topics.

There is, however, a general problem with item bias detection methods of the above type that would seem to make them inappropriate for

situations of varying instructional coverage. As we will demonstrate in the next section, this is because such situations may often be characterized as involving groups for which many or most of the items may be biased.

3. The problem of traditional item bias detection

In this section an IRT model with hypothetical population values will be described. A traditional item bias detection scheme will be applied to this hypothetical model.

Consider the situation of a set of p items y that measure a single latent trait η. Assume that a two-parameter normal ogive model holds for the items (cf. Lord, 1980),

$$(1)\quad P(y_j = 1 \mid \eta) = \Phi [\, a_j (\eta - b_j) \,] \qquad ; j = 1, 2, \ldots p$$

where Φ is the standard normal distribution function, and a and b are the usual discrimination and difficulty parameters. Conditional independence is assumed as usual.

The model above will be used to illustrate the problem of ordinary IRT bias detection in situations where the groups to be compared have different levels of instructional coverage or opportunity-to-learn. Consider for simplicity two groups, group one having latent trait mean

0 and variance 1, and group 2 having mean 1 and variance 0.5. This may represent the situation of group 2 having had more OTL for the set of items at hand than group 1, so that students of this group both have a higher trait level and are more homogeneous with respect to this trait. The two groups will be referred to as the low and the high OTL group. Consider 40 items and five different situations of varying degree of bias.

(1) Zero bias. None of the items are biased, i.e. for each item the same measurement parameters a, b hold in both groups. The item parameters vary over items as follows for item 1 - 10 (a, b): item 1 and 2: 0.98, -2.86; item 3 and 4: 0.98, -1.43; item 5 and 6: 0.98, 0.00; item 7 and 8: 0.98, 1.43; item 9 and 10: 0.98, 2.86.

These same parameter values are used for each of four sets of ten items in the total set of 40 items.

(2) 25 % bias. One of four sets of ten items shows bias, while the other three do not. For the ten biased items, the low OTL group is viewed as not having had sufficient instructional coverage in that the difficulty of each of the 10 items is perceived as higher than for the high OTL group. For the remaining three sets of ten items, the groups of students do not differ in OTL. This is captured by different b values for the two groups, while the a values remain equal. The group 1 (low OTL) b values are increased by 1.14 yielding the b values:

item 1 and 2: -1.72; item 3 and 4: -.29; item 5 and 6: 1.14; item 7 and 8: 2.57; item 9 and 10: 4.00;

(3) 50 % bias. Here, a situation similar to that of 25 % bias is considered, but twenty of the items are biased. The bias is the same as for 25 % bias except that it is applied to two sets of ten items.

(4) 75 % bias. As above, but with three sets of ten biased items.

(5) 100 % bias As above, but with all four sets of ten items being biased.

Using the hypothetical measurement parameter values above, we may study the application of a common item bias detection scheme. The first step involves expressing the item parameters, using the a, b parameterization, in a metric that corresponds to a latent trait mean and variance of 0 and 1. This metric corresponds to the one obtained for estimates obtained by standard IRT analyses. While the parameters for the low OTL group 1 then remain unchanged, the high OTL group 2 values need to be scaled.

The second step involves item response curve bias calculation, for instance in the simple way described in e.g. Linn et al (1981); see also Lord (1980). The mean and variance of the difficulty values are used to equate group 2

measurement parameter values to group 1 values. The item bias for each item is then expressed simply as the square root of the sum of squared differences between the item response curves of (1), summing at steps of 0.1 from 3 to 3 (see Linn et al, 1981). Note that we may also calculate the true bias value for each item. This is obtained by calculating the value based on the original a and b value for each item. Linn et al (1981) describe a response curve bias value of 0.2 or larger as possibly of practical importance.

Table 1 gives the results for the hypothetical bias cases of 25 % - 100 % bias. The case of no bias would simply show that the true and calculated bias values are zero for all items.

Insert Table 1 here

We note that for the cases of 25, 50 and 75 %, the difference between calculated and true bias increases with increasing bias proportion. While for 25 %, the calculated bias values are rather close to the true ones, large errors are observed for 50 % and 75 %. In all cases the calculated bias values seem to indicate bias where there is none, and underestimate bias where it exists. This reflects the fact that the detection technique operates under the assumption that no items are biased and works reasonably well only when a small proportion of items

deviate from this assumption. The fact that the technique makes bias decisions relative to the average item is clearly shown in the case of 100 % bias, where no bias is found. The fact that all items are biased is mistaken by this technique as an indication of a group difference in trait distribution.

We conclude that with strong differences in OTL between students it is quite possible that many or most items are biased, in which case the traditional detection technique is inappropriate.

4. A new IRT extension incorporating OTL

We will present a solution that avoids the problems of the traditional item bias detection scheme. Compared to such schemes, the new approach will avoid the difficulty of forming groups prior to IRT estimation and will also avoid the need for the equating step. This is achieved by allowing the difficulty parameter for each item to vary with the OTL level. In this way, item specific variation in "group membership" is allowed for and the population heterogeneity is taken into account in the model specification.

Assume the availability of OTL information for each of a set of p items y_j. Let the OTL variable connected with item j be denoted x_j. Let $\mathbf{y}^*$ be a p-vector of continuous latent response variables, such that for item j

(2) $y_j = 0$, if $y^*_j \leq \tau_j$
1, otherwise

where τ_j is a threshold parameter defined on y^*_j.

For the p- vectors $\mathbf{y}^*$ and $\mathbf{x}$, assume

(3) $\mathbf{y}^* = \boldsymbol{\lambda}\,\eta + \mathbf{B}\,\mathbf{x} + \boldsymbol{\epsilon}$,

(4) $\eta = \boldsymbol{\gamma}'\,\mathbf{x} + \zeta$,

yielding

(5) $\mathbf{y}^* = (\boldsymbol{\lambda}\,\boldsymbol{\gamma}' + \mathbf{B})\,\mathbf{x} + \boldsymbol{\lambda}\,\zeta + \boldsymbol{\epsilon}$,

where $\boldsymbol{\lambda}$ is a p - vector of measurement slopes, η is the latent trait, $\mathbf{B}$ is a diagonal p x p matrix of slopes reflecting the strength of influence of each x on the level of the corresponding y^*, $\boldsymbol{\epsilon}$ is a p - vector of measurement errors with zero expectation , $\boldsymbol{\gamma}$ is a p - vector of structural slopes describing the influence of the x's on the trait, and ζ is a residual with zero expectation. It is assumed that $\mathbf{y}^*$ conditional on $\mathbf{x}$ has a multivariate normal distribution. Assume further that $\boldsymbol{\epsilon}$ and ζ are independent of each other and of $\mathbf{x}$, and that $\boldsymbol{\epsilon}$ is also independent of η.

Let $V(\zeta) = \psi$ and $V(\epsilon) = \Theta$, where Θ is diagonal. Due to normality it suffices to consider

$$(6)\quad E(\mathbf{y}^* \mid \mathbf{x}) = (\boldsymbol{\lambda}\,\boldsymbol{\gamma}'\,\mathbf{x} + \mathbf{B})\,\mathbf{x}\,,$$

$$(7)\quad V(\mathbf{y}^* \mid \mathbf{x}) = \boldsymbol{\lambda}\,\psi\,\boldsymbol{\lambda}' + \Theta\,,$$

where we may standardize to unit conditional y^* variances, yielding diagonal Θ elements $\theta_{jj} = 1 - \lambda_j^2\,\psi$. Equations (6) and (7) describe the structure that the model imposes on the multivariate regression of $\mathbf{y}^*$ on $\mathbf{x}$. Since the y's are dichotomous, the model may be termed a multivariate structural probit model. The model imposes restrictions on the p x p slopes and on the p(p-1)/2 residual correlations of an "unrestricted" multivariate probit regression model.

Note that

$$(8)\quad E(y^*_j \mid \eta, \mathbf{x}) = \lambda_j\,\eta + \beta_j\,x_j\,,$$

$$(9)\quad V(y^*_j \mid \eta, \mathbf{x}) = \theta_{jj}\,,$$

so that by standard results on conditional means and variances,

$$(10)\quad E(y^*_j \mid \eta) = \lambda_j\,\eta + \beta_j\,E(x_j)\,,$$

$$(11)\quad V(y^{*}_{j} \mid \eta) = \theta_{jj} + \beta_{j}^{2} V(x_{j}) .$$

If x_j is normal, the distribution of y^{*}_{j} conditional on η is normal and we have

$$(12)\quad P(y_{j}=1 \mid \eta) = \Phi \{ [-\tau_{j} + E(y^{*}_{j} \mid \eta)] [V(y^{*}_{j} \mid \eta)]^{-\frac{1}{2}} \}$$

so that the standard normal ogive IRT parameters are obtained as

$$(13)\quad a_{j} = \lambda_{j} [\theta_{jj} + \beta_{j}^{2} V(x_{j})]^{-\frac{1}{2}} ,$$

$$(14)\quad b_{j} = [\tau_{j} - \beta_{j} E(x_{j})] \lambda^{-1} .$$

First note that for β_j's = 0, or no OTL x's present, this is the standard two-parameter normal ogive IRT model. With OTL x's present, each β_j is presumably positive. In standard IRT, incorrectly ignoring the OTL information (or assuming $\beta_j = 0$), (13) and (14) show that we obtain non-invariance of the standard item parameters when a certain set of items is administered to populations with varying OTL distribution; populations with larger OTL variance and higher OTL mean tend to have lower item discrimination and lower item difficulty, respectively. The present extended IRT model avoids such problems by incorporating this item non-invariance directly into the model. Note that the measurement parameters of $\boldsymbol{\tau}$ and $\boldsymbol{\lambda}$ may still be invariant.

If x_j is not normal, $P(y_j = 1 \mid \eta)$ is no longer a normal ogive, although it does represent a monotonically increasing curve. We shall be particularly interested in cases where x_j is dichotomous, denoting presence or absence of sufficient OTL according to some subjective criterion. In such cases, it is useful to consider the normal ogives at the two different x_j values 1 and 0. Using (8) and (9), we then have

$$(15) \quad a_j = \lambda_j \theta_{jj}^{-\frac{1}{2}},$$

$$(16) \quad b_j = (\tau_j - \beta_j x_j) \lambda_j^{-1},$$

From (8) and (16) we note that the β_j difference in conditional y^*_j mean at $x_j = 1$ versus $x_j = 0$ may alternatively be seen as a difference in item difficulty; item j is perceived in two different versions, with and without OTL.

We will show that this model formulation is a special case of a general structural model proposed by Muthen (1984). This general model extends traditional structural equation modeling with continuous variables to situations with categorical and other non-normal measurements, such as the dichotomous ones here. A simplified version of Muthen's general model assumes (using notation similar to that of (2) and (3))

$$(17) \quad \mathbf{y}^* = \Lambda_g \eta_g + \epsilon_g,$$

(18) $\eta_g = B_g \eta_g + \Gamma_g x + \zeta_g$,

yielding

(19) $E (y^* \mid x) = \Lambda_g (I - B_g)^{-1} \Gamma_g x$,

(20) $V (y^* \mid x) = \Lambda_g (I - B_g)^{-1} \Psi_g (I - B_g)'^{-1} \Lambda_g + \Theta_g$.

Here, the subscript g is used to denote quantities corresponding to the general model of Muthen (1984) as opposed to the specific model proposed above. Where quantities are the same, the subscript has been omitted.

To see that the proposed model fits into the general framework, let η_g' $= (y^{*\prime}, \eta)$, $\zeta_g' = (\epsilon', \zeta)$, and let

(21) $y^* = \Lambda_g \eta_g$,

(22) $\eta_g = B_g \eta_g + \Gamma_g x + \zeta_g$,

with

(23) $\Lambda_g = [\, I_{p \times p} \;\; 0 \,]$,

(24) $\Theta_g = 0_{p \times p}$,

$$(25)\quad \mathbf{B}_g = \begin{bmatrix} \mathbf{0}_{p \times p} & \boldsymbol{\lambda} \\ \mathbf{0}' & 0 \end{bmatrix},$$

$$(26)\quad \boldsymbol{\Gamma}_g = \begin{bmatrix} \mathbf{B} \\ \boldsymbol{\gamma}' \end{bmatrix},$$

$$(27)\quad \boldsymbol{\Psi}_g = \begin{bmatrix} \boldsymbol{\Theta} & \text{symm.} \\ \mathbf{0} & \boldsymbol{\psi} \end{bmatrix}.$$

We may note that since the y variables are dichotomous, the diagonal elements of V ($\mathbf{y}^*$ I $\mathbf{x}$) may again be standardized to unity. This means that only the off-diagonal elements of V ($\mathbf{y}^*$ I $\mathbf{x}$), and therefore $\boldsymbol{\Theta}$, enter into the analysis and that the diagonal elements of $\boldsymbol{\Theta}$ may be fixed to any value.

For the statistical background of this technique, the reader is referred to Muthen (1984) . Estimation is carried out by limited information generalized least squares, and a large sample chi-square test of model fit as well as standard errors of estimates are provided. Parameters may be of three kinds: free to be estimated, fixed to a certain value, and constrained to be equal to other parameters. The analyses to be presented have been carried by the LISCOMP program, which builds on the theory of Muthen (1984), see Muthen (1987).

It may be noted that when needed, the proposed model may be easily generalized in the model framework of Muthen (1984). For instance, the $\mathbf{B}$ matrix need not have all off-diagonal elements fixed at zero, the $\boldsymbol{\Theta}$ matrix need not be diagonal, and there may be more than one η.

5. Applications

For illustrative purposes, consider now the application of both the traditional bias detection technique and the proposed approach to some achievement items from the Second International Mathematics Study (SIMS). We will analyze a set of eight dichotomously scored algebra core items described in Table 2.

Insert Table 2 here

The sample consists of 4,129 U.S. eighth grade students (Crosswhite, Dossey, Swafford, McKnight, and Cooney, 1985).

5.1 Traditional IRT bias detection

Consider first LISCOMP estimation of the standard two-parameter normal ogive IRT model of section 4 with no x's present, adding the assumption of a normally distributed trait η. Similar estimates would be obtained by standard IRT analysis. LISCOMP gives a large sample chi-square test of model fit and allows for violations of the conditional independence assumption in the form of correlated ϵ residuals. The standard model of uncorrelated residuals resulted in a chi-square of 61.6 with 20 degrees of freedom. The number of degrees of freedom is the number of restrictions imposed on the correlations among the y^*'s

(see e.g. Muthen, 1978). Relaxing the model restrictions somewhat, a strong improvement in fit was obtained when allowing the residuals for items 5 and 7 to correlate, resulting in a chi-square of 46.4 with 19 degrees of freedom. Given the large sample size, this is regarded as a satisfactory fit.

In line with the student grouping approaches discussed in Section 2, this model was then used for groups based both on OTL and on type of mathematics class. The OTL measures were obtained from the teachers, where for each item the teacher responded to the question: "During this school year did you teach or review the mathematics needed to answer the item correctly ?"

The answers were No (scored 0) and Yes (scored 1). A similar question was directed to the students. There is clearly a question of reliability of both these reports. The student response may be affected by the perceived difficulty of the item, and the teacher response concerns the class as a whole, where a claim of coverage may be irrelevant for the student who was absent. We have chosen to work with the teacher response since we feel it may be the least unreliable; the fact that this measurement is not on the student level is here ignored.

A low and a high OTL group was created by splitting the students based on the sum of the item OTL scores at ≤ 6 versus higher, resulting in sample sizes of 2,101 versus 2,028. As an alternative, students were also divided into two groups based on type of mathematics class.

Remedial and Typical classes were contrasted with Enriched and Algebra classes, yielding 2,592 versus 1,537 students.

In the low OTL group, the model with 19 degrees of freedom obtained a chi-square value of 26.6, while in the high OTL group, 47.9 was obtained. For the "low" class types, the chi-square 30.2 was obtained, whereas the "high" class types obtained the value 36.5. The estimation was carried out with trait mean of zero and trait variance one. The estimated τ and λ values can be translated to the IRT a and b values by setting $\beta = 0$ in (13) and (14). The response curve bias index may then be computed as usual.

The left-most part of Table 3 ("Traditional") gives the resulting item bias values for each item given the two ways of dividing students into groups. These results are given both for the standard model with uncorrelated residuals (Model I; 20 d.f.) and the model allowing the free residual correlation (Model II; 19 d.f.).

Insert Table 3 here

We note that the least amount of bias is observed for the first three items. There seems to be little difference between bias values calculated from Model I versus Model II, and using the two ways of

creating the student groups.

5.2 Applying the new IRT approach to the SIMS data

Applying the new approach to the SIMS algebra core items, a chi-square test of model fit gave the value 223.5 with 68 degrees of freedom. The number of degrees of freedom is obtained as the total number of restrictions imposed on the p x p regression slopes of E ($\mathbf{y}^*$ I $\mathbf{x}$) and the p(p-1)/2 residual correlations of V ($\mathbf{y}^*$ I $\mathbf{x}$) (see Muthen, 1984). A strong improvement in fit was obtained when allowing the residuals for items 5 and 7 to correlate and a further improvement was obtained when allowing correlation between the residuals for items 6 and 8. For simplicity, we chose as our final model the one with only errors 5 and 7 free to correlate, resulting in a chi-square with 208.8 with 67 degrees of freedom. Let this model be denoted Model III. The estimates from Model III are given in Table 4.

Insert Table 4 here

The most interesting result concerns the estimated $\boldsymbol{\beta}$ values on the diagonal of $\mathbf{B}$ representing the effect of each of the OTL variables x on the corresponding response variable. Note that this is an effect over and above that of η, so that we are describing the effect of OTL for given

achievement trait value. For the first four items we have strong positive effects of OTL while the last four exhibit insignificant OTL effects.

The estimated β values and their standard errors give a succinct way of assessing item bias, or instructional sensitivity, in each item. However, for comparison it may be of interest to study the corresponding item response curve bias values. These may be computed from the estimated Model III by (15) and (16). The bias values are given in Table 3 in the Model III column. We note that the results contradict those for Model I and Model II using the traditional approach of dividing the students into groups. The difference is particularly strong for the last four items. The difference is possibly due to the deficiency of the traditional approach discussed in section 3.

References

Airasian, P.W. & G.F. Madaus (1983). Linking testing and instruction. Journal of Educational Measurement, 20, 103-118.

Anderson, L.W. (1985). Opportunity to learn. In Husen, T. & Postlethwaite, T.N. (eds.). The International Encyclopedia of Education, Oxford:Pergamon Press.

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.

Engelhard, G. (1986). Curriculum-based estimates of student achievement. Paper presented at the annual meeting of the Psychometric Society in Toronto, Canada.

Haertel, E. & Calfee, R. (1983). School achievement: Thinking about what to test. Journal of Educational Measurement, 20, 119-132.

Linn, R.L. & Harnisch, D.L. (1981). Interactions between item content and group membership. 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.

Lord, F.M. (1980). Applications of Item Response Theory to Practical Testing Problems. Hillsdale, N.J.: Erlbaum.

Mehrens, W.A.. & S.E. Phillips (1986). Detecting impacts of curricular differences in achievement test data. Journal of Educational Measurement, 23, 185-196.

Miller, M.D. (1986). Time allocation and patterns of item response. Journal of Educational Measurement, 23, 147-156.

Miller, M.D. & R.L. Linn (1986). Invariance of item parameters with variations in instructional coverage. Accepted for publication in the Journal of Educational Measurement.

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

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

Muthen, B. (1987). LISCOMP. Analysis of Linear Structural

Equations with a Comprehensive Measurement Model. Users' Guide. Scientific Software, Inc. Mooresville, Ind.

Traub, R.E. & Wolfe, R.G. (1981). Latent trait theories and the assessment of education achievement. In D.C. Berliner (Ed.), Review of research in education.

TABLE 1

Item Bias in Artificial Data

(Entries are the square root of the sum of squared response curve difference)

25% Bias

Biased Set (10 items)

Item	Calculated Bias	True Bias	Difference
1,2	.32	.39	-.07
3,4	.35	.45	-.10
5,6	.34	.45	-.11
7,8	.29	.41	-.12
9,10	.18	.27	-.10
Average:	.30	.39	-.09

Unbiased Sets (3 x 10 items)

1,2	.07	.00	.07
3,4	.10	.00	.10
5,6	.11	.00	.11
7,8	.12	.00	.12
9,10	.10	.00	.10
Average:	.10	.00	.10

50% Bias

Biased Sets (2 x 10 items)

Item	Calculated Bias	True Bias	Difference
1,2	.24	.39	-.15
3,4	.25	.45	-.20
5,6	.23	.45	-.23
7,8	.18	.41	-.23
9,10	.10	.27	-.17
Average:	.20	.39	-.19

Unbiased Sets (2 x 10 items)

1,2	.15	.00	.15
3,4	.20	.00	.20
5,6	.23	.00	.23
7,8	.23	.00	.23
9,10	.17	.00	.17
Average:	.20	.00	.20

75% Bias

Biased Set (3 x 10 items)

Item	Calculated Bias	True Bias	Difference
1,2	.13	.39	-.26
3,4	.13	.45	-.32
5,6	.11	.45	-.34
7,8	.09	.41	-.32
9,10	.05	.27	-.23
Average:	.10	.39	-.29

Unbiased Set (10 items)

1,2	.26	.00	.26
3,4	.32	.00	.32
5,6	.34	.00	.34
7,8	.32	.00	.32
9,10	.23	.00	.23
Average:	.29	.00	.29

100% Bias

Biased Sets (2 x 10 items)

Item	Calculated Bias	True Bias	Difference
1,2	.00	.39	-.39
3,4	.00	.45	-.45
5,6	.00	.45	-.45
7,8	.00	.41	-.41
9,10	.00	.27	-.27
Average:	.00	.39	-.39

TABLE 2

Wording for Eight Posttest 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. $(-2) \times (-3)$ 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 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

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 $b = d^2$
   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

Item Bias Values for Various SIMS Models

Item	Traditional				New
	Model I		Model II		Model III
	OTL	Class	OTL	Class	
1	.06	.06	.06	.06	.15
2	.17	.12	.17	.12	.11
3	.13	.18	.13	.17	.30
4	.26	.19	.25	.18	.13
5	.19	.20	.22	.21	.02
6	.23	.22	.22	.22	.01
7	.22	.18	.23	.18	.01
8	.21	.25	.21	.24	.03

TABLE 4

Estimates from Model III

Measurement Parameters

Item	τ		λ		β	
	Est.	t-ratio	Est.	t-ratio	Est.	t-ratio
1	1.45	22.31	0.77	17.50	0.33	6.22
2	0.07	0.97	0.86	23.43	0.22	3.72
3	0.93	10.79	1.00	0.00	0.64	8.93
4	0.66	9.51	0.86	23.19	0.28	5.46
5	0.52	6.41	0.85	22.93	-0.03	-0.49
6	0.31	4.78	0.94	24.36	-0.03	-0.65
7	0.60	11.53	0.67	18.04	-0.02	-0.38
8	0.23	3.87	0.80	21.08	0.05	1.12

	γ		ψ	
x variable	Est.	t-ratio	Est.	t-ratio
			0.43	18.18
1	0.20	6.09		
2	-0.15	-3.28		
3	-0.01	-0.22		
4	0.03	0.72		
5	0.13	2.22		
6	0.23	7.41		
7	0.17	6.12		
8	0.33	9.28		

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 endorsment by OERI/ED should be inferred.