PSYCHOMETRIKA—VOL. 54, NO. 3, 385–396
SEPTEMBER 1989

USING ITEM-SPECIFIC INSTRUCTIONAL INFORMATION IN ACHIEVEMENT MODELING

BENGT O. MUTHÉN

GRADUATE SCHOOL OF EDUCATION, UNIVERSITY OF CALIFORNIA, LOS ANGELES

The problem of detecting instructional sensitivity ("item basis") in test items is considered. An illustration is given which shows that for tests with many biased items, traditional item bias detection schemes give a very poor assessment of bias. A new method is proposed instead. This method extends item response theory (IRT) by including item-specific auxiliary measurement information related to opportunity-to-learn. Item-specific variation in measurement relations across students with varying opportunity-to-learn is allowed for.

Key words: auxiliary information, opportunity-to-learn, item response theory, structural modeling, LISCOMP.

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, there has been 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), Phillips and Mehrens (1988), and Miller (1986). An interesting development is the use of "opportunity-to-learn" measures (Anderson, 1985), see, for example, Miller and Linn (in press), 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 to include such auxiliary information will be considered. Use of auxiliary information was also studied by Mislevy (1987), focussing on the precision gain in estimating latent trait values, and Muthén (1987a), focussing on the estimation of effects of background variables on item responses.

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

This paper was presented at the 1987 AERA meeting in Washington, DC. This research was supported by grant OERI-G-86-003 from the Office of Educational Research and Improvement, Department of Education. The author thanks Michael Hollis and Chih-fen Kao for valuable research assistance, and appreciates valuable comments made by an anonymous reviewer.

Requests for reprints should be sent to Bengt O. Muthén, Graduate School of Education, 405 Hilgard Avenue, University of California, Los Angeles, Los Angeles, CA 90024-1521.

0033-3123/89/0900-9518$00.75/0

#30

group with high coverage as for a group with low coverage (see also Birenbaum & Tatsuoka, 1983); indeed, with very low coverage, the validity of the item is called in question.

This paper concentrates on the question of how the modeling should capture the fact that the measurement relationship between the items and the latent trait may vary over students. This question will be formulated as a problem of assessing "item bias", or instructional sensitivity in the items, when item-specific opportunity-to-learn type information is available.

In section 2 traditional IRT methodology to assess measurement differences will be considered and the weakness of this item bias approach demonstrated by means of an artificial data analysis. In section 3 a new method is proposed which does not necessitate the traditional 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 and item bias detection is obtained as a by-product. In section 4 the traditional and new methods for mathematics achievement data are applied and the outcomes in terms of item response curve bias are compared.

2. Traditional Item Bias Detection

With the availability of auxiliary information such as opportunity-to-learn (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 rescaling, and some form of item bias index is calculated for each item (see e.g., Linn, Levine, Hastings, & Wardrop, 1981). An example is Miller and Linn (1986), who analyzed the same mathematics achievement data as in section 4. They found evidence of "instructional bias" (Linn & Harnisch, 1981), and noted that the magnitude of the biases were frequently larger than commonly encountered when considering student groups formed by ethnicity.

There is a general problem with item bias detection methods of the traditional type that would seem to make them inappropriate for situations of strongly varying instructional coverage. As demonstrated below, this is because such situations may often be characterized as involving groups for which many or most of the items may be biased.

Consider an IRT model with hypothetical population values. 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 (Lord, 1980),

$$P(y_j = 1|\eta) = \Phi[a_j(\eta - b_j)]; \qquad j = 1, 2, \cdots p \tag{1}$$

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 in IRT modeling.

The model above will be used to illustrate the problem of traditional IRT bias detection in situations where the groups to be compared have different levels of instructional coverage or OTL. 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. Parameter values are chosen as commonly occurring values in the τ, λ metric (see section 3) and then transformed to the a, b metric.

1. Zero bias. None of the items are biased, that is, for each item the same measurement parameters a, b hold in both groups. The item parameters vary over items as follows for Item 1 through 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 for the high OTL group. 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. For the remaining three sets of ten items, the groups of students do not differ in OTL.

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 usually obtained for IRT estimates. 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, for example, Linn et al. (1981); see also Lord (1980). The mean and variance of the difficulty values are used to rescale 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.). 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. 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.

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

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

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.

3. A New IRT Extension Incorporating OTL

In the present situation it is advantageous to recognize 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?

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 rescaling 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

$$y_j = \begin{cases} 0, & \text{if } y_j^* \leq \tau_j \\ 1, & \text{otherwise,} \end{cases} \tag{2}$$

where τ_j is a threshold parameter defined on y_j^*.

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

$$\mathbf{y}^* = \boldsymbol{\lambda}\eta + \mathbf{B}\mathbf{x} + \boldsymbol{\varepsilon}, \tag{3}$$

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

yielding

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

where $\boldsymbol{\lambda}$ is a p-vector of measurement slopes, η is the latent trait, $\mathbf{B}$ is a diagonal $p \times p$ matrix of slopes reflecting the strength of influence of each of the px variables on the level of the corresponding y^*, $\boldsymbol{\varepsilon}$ 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 $\boldsymbol{\varepsilon}$ and ζ are independent of each other and of $\mathbf{x}$, and that $\boldsymbol{\varepsilon}$ is also independent of η.

It may be noted that the latent trait variable η is viewed by (4) as a latent performance level. This then deviates somewhat from the traditional treatment of latent ability as invariant, innate ability, but may be more suited to the study of how achievement responses vary due to instructional differences.

Let $V(\zeta) = \psi$ and $V(\boldsymbol{\varepsilon}) = \boldsymbol{\Theta}$, where $\boldsymbol{\Theta}$ is diagonal. Due to normality it suffices to consider

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

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

where we may standardize to unit conditional y^* variances, yielding diagonal $\boldsymbol{\Theta}$ elements $\theta_{jj} = 1 - \lambda_j^2\psi$. Equations (6) and (7) describe the restrictions that the model imposes on the multivariate regression of $\mathbf{y}^*$ on $\mathbf{x}$. The model imposes restrictions on the $p \times p$ slopes and on the $p(p-1)/2$ residual correlations of an "unrestricted" multivariate probit regression model (see also Ashford & Sowden, 1970), and may therefore be termed a multivariate structural probit model (Muthén, 1979). Note that

$$E(y_j^*|\eta, \mathbf{x}) = \lambda_j\eta + \beta_j x_j, \tag{8}$$

$$V(y_j^*|\eta, \mathbf{x}) = \theta_{jj}, \tag{9}$$

so that by standard results on conditional means and variances,

$$E(y_j^*|\eta) = \lambda_j\eta + \beta_j E(x_j), \tag{10}$$

$$V(y_j^*|\eta) = \theta_{jj} + \beta_j^2 V(x_j). \tag{11}$$

If x_j is normal, the distribution of y_j^* conditional on η is normal and we have

$$P(y_j = 1|\eta) = \Phi\{[-\tau_j + E(y_j^*|\eta)][V(y_j^*|\eta)]^{-1/2}\} \tag{12}$$

so that the normal ogive IRT parameters are obtained as

$$a_j = \lambda_j[\theta_{jj} + \beta_j^2 V(x_j)]^{-1/2}, \tag{13}$$

$$b_j = [\tau_j - \beta_j E(x_j)]\lambda^{-1}. \tag{14}$$

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 noninvariance 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, $P(y_j = 1|\eta)$ is no longer a normal ogive, although it does represent a monotonically increasing curve. Cases where x_j is dichotomous are particularly interesting, 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

$$a_j = \lambda_j\theta_{jj}^{-1/2}, \tag{15}$$

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

From (8) and (16) 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.

This model formulation will be shown to be a special case of a general structural model proposed by Muthén (1984); see also Muthén (1983). This general model extends traditional structural equation modeling with continuous variables to situations with categorical and other nonnormal measurements, such as the dichotomous ones here. A

simplified version of Muthén's general model assumes (using notation similar to that of (2) and (3))

$$\mathbf{y}^* = \boldsymbol{\Lambda}_g\boldsymbol{\eta}_g + \boldsymbol{\varepsilon}_g, \tag{17}$$

$$\boldsymbol{\eta}_g = \mathbf{B}_g\boldsymbol{\eta}_g + \boldsymbol{\Gamma}_g\mathbf{x} + \boldsymbol{\zeta}_g, \tag{18}$$

yielding

$$E(\mathbf{y}^*|\mathbf{x}) = \boldsymbol{\Lambda}_g(\mathbf{I} - \mathbf{B}_g)^{-1}\boldsymbol{\Gamma}_g\mathbf{x}, \tag{19}$$

$$V(\mathbf{y}^*|\mathbf{x}) = \boldsymbol{\Lambda}_g(\mathbf{I} - \mathbf{B}_g)^{-1}\boldsymbol{\Psi}_g(\mathbf{I} - \mathbf{B}_g)'^{-1}\boldsymbol{\Lambda}_g + \boldsymbol{\Theta}_g. \tag{20}$$

Here, the subscript g is used to denote quantities corresponding to the general model of Muthén (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 $\eta_g' = (\mathbf{y}^{*\prime}, \eta)$, $\boldsymbol{\zeta}_g' = (\boldsymbol{\varepsilon}', \zeta)$, and let

$$\mathbf{y}^* = \boldsymbol{\Lambda}_g\boldsymbol{\eta}_g, \tag{21}$$

$$\boldsymbol{\eta}_g = \mathbf{B}_g\boldsymbol{\eta}_g + \boldsymbol{\Gamma}_g\mathbf{x} + \boldsymbol{\zeta}_g, \tag{22}$$

with

$$\boldsymbol{\Lambda}_g = [\mathbf{I}_{p\times p}\mathbf{0}], \tag{23}$$

$$\boldsymbol{\Theta}_g = \mathbf{0}_{p\times p}, \tag{24}$$

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

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

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

Note that since the y variables are dichotomous, the diagonal elements of $V(\mathbf{y}^*|\mathbf{x})$ may again be standardized to unity. This means that only the off-diagonal elements of $V(\mathbf{y}^*|\mathbf{x})$, 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 Muthén (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 out by the LISCOMP program, which builds on the theory of Muthén (1984), see Muthén (1987b).

It may be noted that when needed, the proposed model may be easily generalized in the model framework of Muthén (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.

4. 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; Crosswhite, Dossey, Swafford, McKnight, & Cooney, 1985). We will analyze a set of eight dichotomously scored algebra core items described in Table 2.

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

4.1 Traditional IRT Bias Detection

Consider the LISCOMP estimation of the two-parameter normal ogive IRT model of section 3 with no OTL variables (x's present, adding the assumption of a normally distributed trait η. 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 latent response variables (the y*'s; see e.g., Muthén, 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.

This model was then used for student 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 and the answers do not correlate very highly with the teachers'. 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. The teacher response was selected since 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.).

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.

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) 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 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 ≥ 12
   C x - 6 < 12
   D 6 - x ≥ 12
   E 6 - x < 12

4.2 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 $p \times p$ regression

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

slopes of $E(y^*|\mathbf{x})$ and the $p(p-1)/2$ residual correlations of $V(\mathbf{y}^*|\mathbf{x})$ (see Muthén, 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, the final model was chosen as 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.

The most interesting result concerns the estimated β 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 the effect of OTL for given achievement trail value is described. For the first four items we have strong positive effects of OTL while the last four exhibit insignificant OTL effects. This article will not go into substantive interpretations of the results; the interested reader is instead referred to the more extensive analyses of Muthén, Kao, and Burstein (1988).

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. 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 2.

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		

5. Discussion

A method has been proposed that enables the estimation of different versions of each item-trait relationship, where the versions correspond to different levels of opportunity-to-learn. This also results in an assessment of instructional sensitivity for each item.

The detection of instructionally sensitive items is of interest not only because

ignoring such items would bias traditional IRT results. The analysis also provides a more detailed way to learn about item response in relation to instruction and this may benefit both instruction and test construction. Instructionally sensitive items are not suitable for attempts at measuring stable and general traits since the instruction does not affect the item response only through the trait, but also directly.

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 T. Husen & T. N. Postlethwaite (Eds.), The international encyclopedia of education. Oxford: Pergamon Press.

Ashford, J. R., & Sowden, R. R. (1970). Multivariate probit analysis. *Biometrics, 26*, 535–546.

Birenbaum, M., & Tatsuoka, M. (1983). The effect of scoring systems based on the algorithm underlying the student's response patterns on the dimensionality of achievement test data of the problem solving type. *Journal of Educational Measurement, 20*, 17–26.

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, IL: Stipes.

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

Fisher, G. H. (1972). A measurement model for the effect of mass-media. *Acta Psychologica, 36*, 207–220.

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, NJ: Erlbaum.

Mehrens, W. A., & Phillips, S. E. (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., & Linn, R. L. (1988). Invariance of item parameters with variations in instructional coverage. *Journal of Educational Measurement, 25*, 205–219.

Mislevy, R. J. (1987). Exploiting auxiliary information about examinees in the estimation of item parameters. *Applied Psychological Measurement, 11*, 81–91.

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika, 43*, 551–560.

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

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

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

Muthén, B. (1987a). Some uses of structural equation modeling in validity studies: Extending IRT to external variables, In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 213–238). Hillsdale, NJ: Lawrence Erlbaum.

Muthén, B. (1987b). LISCOMP. *Analysis of linear structural equations with a comprehensive measurement model*. Users' guide. Mooresville, IN: Scientific Software.

Muthén, B., Kao, C-F., & Burstein, L. (1988). *Instructional sensitivity in mathematics achievement test items: Application of a new IRT-based detection technique*. Los Angeles: University of California, Los Angeles, Graduate School of Education. (Accepted for publication in the *Journal of Educational Measurement*)

Phillips, S. E., & Mehrens, W. A. (1988). Effects of curricular differences on achievement test data at item and objective levels. *Applied Measurement in Education, 1*, 33–51.

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*.

Manuscript received 9/3/87
Final version received 3/21/88