Journal of Educational Measurement
Spring 1991, Vol. 28, No. 1, pp. 1–22

Instructionally Sensitive Psychometrics: Application of a New IRT-Based Detection Technique to Mathematics Achievement Test Items

Bengt O. Muthén, Chih-Fen Kao, and **Leigh Burstein**
University of California—Los Angeles

Achievement modeling is carried out in groups of students characterized by heterogeneous instructional background. Extensions of item response theory models incorporate variables reflecting different amounts of opportunity-to-learn (OTL). The effects of these OTL variables are studied with respect to their influence on both the latent trait and the item performance directly. Such direct effects may reflect instructionally sensitive items. U.S. eighth-grade mathematics data from the Second International Mathematics Study are analyzed. Here, the same test is taken by students enrolled in typical instruction and students enrolled in elementary algebra classes. It is shown that the new analysis provides a more detailed way to examine the influence of instruction on responses to test items than does conventional item response theory.

Standardized achievement testing in most American schools today involves a heterogeneous group of students. One major source of this heterogeneity at a given grade level is the difference in instructional experiences of students (McKnight et al., 1987). It is little wonder that the match between the school curriculum and what is tested continues to be of concern (e.g., Airasian & Madaus, 1983; Haertle & Calfee, 1983; Linn, 1983; Schmidt, Porter, Schwille, Floden, & Freeman, 1983; Leinhardt, 1983; Leinhardt & Seewald, 1981; Mehrens & Phillips, 1986; Miller, 1986).

The research reported here extends our developments of item response theoretic methods for achievement of heterogeneous groups of students (Muthén, 1988, 1989). Within this framework, the present study expands on efforts to disentangle the influences of ascriptive instructional backgrounds as they impact estimation of the parameters of the achievement measurement model. The emphasis here is on how one might model the effects of differences in instructional backgrounds of students on the resulting achievement latent trait and observed item difficulties. This work is being reported at a relatively early phase of the inquiry in order to call attention to what we view to be a potentially fruitful psychometric method for examining achievement test data obtained from students with varying instructional backgrounds. It is hoped that presentation of the research at this stage will stimulate discussion about the

The research for this article, presented at the annual meeting of the American Educational Research Association, Boston, 1990, was supported by grant OERI-G-86-003 from the Office of Educational Research and Improvement, Department of Education. The authors thank Michael Hollis and Suk-Woo Kim for valuable research assistance.

#35

applicability of the methodology for research and practice within the domain of large-scale assessment.

Item Response Theory (IRT) is a common tool for the study of item bias. Under the IRT model, invariance of measurement parameters is assumed to hold for different subgroups. Deviations from this assumption are viewed as *item bias.* To detect bias, the group membership of the examinees is identified, and the estimated curves describing the probability of a correct answer for a given ability level are compared across groups. A large area between curves is an indication of *IRT item bias.*

As suggested by Linn and Harnisch (1981), *instructional bias* may be mistaken as bias due to ethnicity. Recent studies have changed the traditional focus on ethnic and gender biases in achievement tests to instructional bias. For instance, Lehman (1986) studied algebra items for eighth-grade students. Gender and opportunity-to-learn (OTL; Anderson, 1985) in the classroom were used as grouping variables. Relative to gender, OTL was found to be a much more important cause of item bias. Miller and Linn (1988) used an alternative approach to the study of instructional bias. Based on OTL and item content, cluster analysis was carried out to create curriculum clusters. When comparing item response curves for the same item across clusters, they found strong evidence of instructional bias. The magnitude of the instructional bias was claimed to be larger than that usually found with different ethnic groups.

The Lehman (1986) and Miller-Linn (1988) approaches build on grouping test takers. The grouping may depend on the sample distribution. There is also 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. Different grouping criteria may lead to different conclusions.

Standard IRT techniques assume that instruction increases the item performance through an increase in the latent trait level, whereas the item-trait relationship remains the same. This assumption is usually too strong for groups of students with widely different content coverage. Certain classes may have obtained more extensive instruction for specific content areas so that the performance on the corresponding item types is relatively better than on the majority of the items for the average student. This is the cause of instructional item bias. Muthén (1989) pointed out the psychometric problem of traditional IRT-based item bias detection schemes, showing a misestimation of bias in the plausible situation of many items' showing instructional bias. Muthén's extended IRT model may serve as a better tool for studying the instructional bias, or, as we will term it, the *instructional sensitivity.* His model maintains the form of an IRT model but in addition has parameters that quantify the extent of the effect attributed to OTL. Using similar modeling, Muthén (1988) also considers other educational and social student background information as predictors of item response. As Mislevy (1987) indicated, "what IRT models miss are these systematic differences among examinees performing at the same general level" (pp. 261–262). The assumptions of IRT that preclude the influences from auxiliary variables are challenged and examined in Muthén's model.

Muthén's model may be briefly described as follows. Building on his own statistical theory (1984), Muthén (1988) proposed a new extension of IRT modeling that controls for student background differences by including background variables as covariates. Further extending this methodology, Muthén (1989) proposed a method for explicitly including item-specific information on instructional differences, allowing for OTL effects on performance not only through an increase in trait level but also directly. This model parameterization essentially allows for several difficulty levels for each item corresponding to different instructional classifications. In this way, the deficiency of traditional IRT bias detection techniques is avoided. The instructional heterogeneity of the students is taken into account, and any differential instructional effects on the item difficulty parameters can be directly estimated.

The Muthén (1989) technique for detecting instructionally sensitive items was illustrated with a very small set of eight algebra items from the U. S. sample of eighth graders in the Second International Mathematics Study (SIMS; Crosswhite, Dossey, Swafford, McKnight, & Cooney, 1985). The aim of this article is to apply the technique to detect instructional sensitivity in a more realistic setting, using the SIMS set of 40 core items for U.S. eighth graders. This set contains items covering algebra, arithmetic, geometry, and measurement. By this analysis, it is hoped that types of items that are particularly susceptible to instructional sensitivity in this context can be discerned. Such items may be less suitable to activities of broad assessment of more stable traits but may be of primary interest for achievement assessment. The achievement measurement process can be improved by better understanding the link between item types and instruction in this way. Furthermore, item analysis by standard IRT techniques would ignore instructionally sensitive items and result in biased estimates of measurement parameters.

The Data

In brief, the SIMS data features are as follows. A national probability sample of school districts was selected proportional to size; a probability sample of schools was selected proportional to size within the school district; and two classes were randomly selected within each school yielding a total of about 240 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, and measurement distributed among four 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 the analysis that follows, a key piece of instructional information was OTL. Teachers rated items for student OTL. For this study, OTL was defined as whether the math needed to answer this item correctly was taught this year or in prior years.

The percentage distribution of OTL categories for all 40 items are given in Table 1 together with proportion correct for each test item. It seems that the percentage of students having no opportunity-to-learn (NTL) varies greatly across the items. With the exception of five items, having had OTL is most common. However, about 1/3 of the items show NTL proportions larger than

Table 1

Proportions and Effect Comparisons for OTL Categories

Item	Overall proportion correct	Proportions for OTL categories[a]		OTL effects beta coefficients D values
		OTL	NTL	
1	.43	.79 .47	.21 .26	.25 .12
2	.60	.97 .61	.03 .53	-.01 -.01
3	.21	.62 .28	.38 .09	.38 .19
4	.33	.87 .34	.13 .26	-.04 -.02
5	.44	.93 .45	.07 .30	.28 .14
6	.55	.72 .55	.28 .54	-.08 -.04
7	.66	.31 .68	.69 .66	.15 .08
8	.89	.83 .89	.17 .88	.05 .03
9	.52	.86 .53	.14 .48	-.01 -.01
11	.31	.60 .35	.40 .26	-.04 -.02
12	.44	.90 .44	.10 .40	.15 .08
13	.71	.88 .73	.12 .59	-.02 -.01
14	.61	.85 .63	.15 .53	-.00 -.00
15	.32	.90 .32	.10 .28	.04 .02

(table continues)

Table 1 (continued)

Item	Overall proportion correct	Proportions for OTL categories[a]		OTL effects beta coefficients D values
		OTL	NTL	
16	.58	.94 .60	.06 .16	.41 .20
17	.59	.87 .62	.13 .38	.59 .32
18	.51	.80 .56	.20 .29	.16 .08
19	.33	.24 .39	.76 .32	.08 .04
20	.77	.98 .77	.02 .60	-.37[b] -.17[b]
21	.34	.40 .39	.60 .30	.10 .04
22	.59	.87 .64	.13 .26	.22 .10
23	.47	.81 .51	.19 .30	.02 .01
24	.57	.93 .58	.07 .36	.05 .02
25	.46	.94 .47	.06 .34	-.05 -.02
26	.62	1.00 .62	--[c] --	-- --
27	.57	.47 .64	.53 .50	-.01 -.01
28	.62	.91 .63	.09 .49	-.02 -.01
29	.75	.90 .77	.10 .60	.04 .02

(table continues)

0.33. It is also seen that the proportion correct varies greatly over the different OTL categories. These are clear indications of the student heterogeneity.

The use of the dichotomously scored, teacher-reported OTL in our model is noteworthy. Mehrens and Phillips (1986) used textbook series and school personnel ratings to study the influence of the match between what was taught and what was tested for reading and math scores in Grades 3 and 6. As Leinhardt and Seewald (1981) pointed out, the two most common approaches to the measurement of overlap between what is tested and what is taught are instructional-based and curriculum-based measurement.

In the SIMS, student-reported item-specific OTL is also available. Both teacher- and student-reported OTL is presumably fraught with error. Teachers' reporting may not be relevant for a student who was absent from or did not understand the instruction. A student's reporting may partly refelct his or her perception of the item difficulty. The two ways of reporting are not highly correlated (Lehman, 1986). We feel that the teacher-reported OTL is more trustworthy.[1]

Table 1 (continued)

Item	Overall proportion correct	Proportions for OTL categories[a] OTL	NTL	OTL effects beta coefficients D values
30	.40	.48	.52	-.05
		.45	.36	-.03
31	.62	1.00	--[c]	--
		.62	--	--
32	.45	1.00	--[c]	--
		.62	--	--
33	.50	.95	.05	.06
		.51	.33	.03
34	.39	.96	.04	-.41
		.40	.19	-.18
35	.59	.71	.29	.24
		.65	.44	.13
36	.56	.93	.07	.22
		.57	.38	.10
37	.37	.85	.15	.03
		.40	.23	.01
38	.51	.97	.03	.74
		.52	.23	.34
39	.54	.70	.30	.67
		.63	.33	.35
40	.47	.53	.47	.21
		.52	.41	.11

[a]Entries are proportion of students and proportion correct.

[b]Estimate is not dependable due to small category proportion (<=0.02).

[c]No students are available in the NTL category.

In addition to the above item-specific OTL information, instructional background information common to all items is available in the SIMS in the form of a classification of each mathematics class into one of four types: basic or remedial arithmetic (REMEDIAL), general or typical mathematics (TYPICAL), pre-algebra or enriched (ENRICHED), and algebra (ALGEBRA). This classification is based on teacher questionnaire data and on information on textbooks used.

In the SIMS data, there is also available a set of background variables for each student measured during the fall of eighth grade. These variables include pretest measurements of mathematics, family background, educational aspiration, attitudes toward mathematics, gender, and ethnicity (see Table 2 and Muthén, 1988).

The premeasurements were collected for only part of the sample. But the analysis considers a total number of 3,724 students who had complete observation on both fall and spring measurements in this set. This analysis sample involves 198 classes.

Table 2

Description of Background Variables

PREALG: Proportion of correct responses 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 education (including university, college, etc.) 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 occupations (Reversed).
3. Mathematics is not needed in every day living (Reversed).
4. Most people do not use mathematics in their jobs (Reversed).

(table continues)

Table 2 (continued)

ATTRACT: Average scores of five attitude items. Scoring is 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).

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.

The Model

Following Muthén (1989), detection of instructionally sensitive items among the set of items is achieved by estimation of the following model. A diagrammatic representation of the model is given in Figure 1. The model will first be described in words and then statistically.

The mathematics trait in the spring of eighth grade is an unobserved continuous variable that is measured by, or in other words, predicts, the set of test items. This trait will alternatively be called *math ability* or *achievement level,* although a more careful distinction is no doubt desirable when discussing a trait for students with varying OTL. Muthén (1989) suggests the term *latent performance level.* We want to study the effect of OTL on the item performance because it is possible that having OTL enhances the specific skills needed to solve the corresponding item correctly. Adding these variables as predictors, the modeling has to recognize that math ability in the spring is an endogenous variable relative to the OTL variables. The OTL variables predict the item performance but also determine a part of the math ability level itself. To correctly model the prediction of spring math ability, it then becomes necessary to specify a more comprehensive set of predictors for math ability, where item OTL influences on math ability are specified as partial effects, holding other background variables constant.

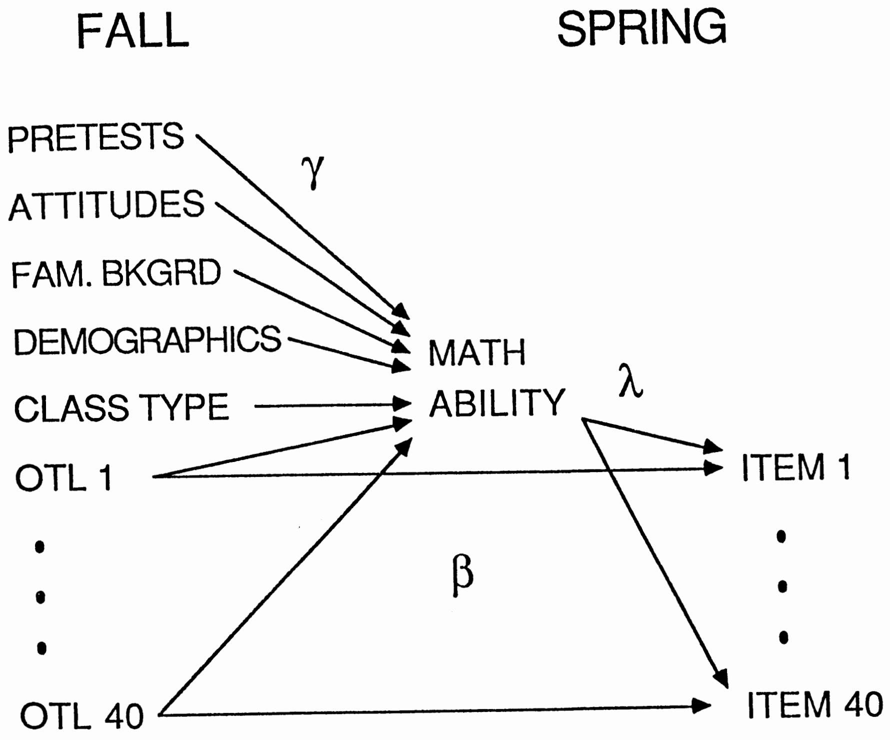

FIGURE 1. *Model for assessing instructional item sensitivity*

Spring math ability is taken here to be predicted by fall pretests, attitudes, family background, demographics, class type, and OTL. These predictors influence the math ability variable and thus, indirectly, also the performance on the test items. All the background variables are assumed to have direct effects on math ability. However, the majority of the background variables is assumed to have only indirect effects on items.

The OTL variables, however, are also allowed to influence the corresponding test items directly, although not all items are expected to have such effects. Any such effect would be an influence of OTL over and above that which is transferred via the math ability. Hence, the probability of a correct response for students with different OTL would be different even if they have the same math ability. This effect implies item bias due to instructional sensitivity in the item at hand in terms of measuring the math ability trait. This can be stated as OTL not influencing math ability homogeneously across the set of test items.

It is interesting to note that bias due to instructional sensitivity in the items is assessed here without resorting to traditional item bias detection schemes that necessitate a classification of students into groups with different OTL values. The present analysis avoids the arbitrariness of such groupings in a situation where group membership obviously varies across items. The model also presents a wealth of other relevant information on the achievement process.

Statistically, the model may be presented as follows. An IRT model is specified for measuring the trait by the set of items. In this analysis, a two-parameter normal ogive response curve model is chosen for this measurement part (e.g., Lord, 1980). Let us consider the influence of the item-specific OTL variables, z say, and the student background variables, x say (premeasurements, attitudes, demographics, and clsss type). In our analysis, we will create an OTL dummy variable for each item j, z_j, representing the two categories of the OTL levels described in the section entitled *The Data.* Here, $z_j = 1$ represents OTL. The variables of $\mathbf{z}$ and $\mathbf{x}$ are assumed to influence the latent trait variable η, say. We specify the linear regression model

$$\eta = \gamma_x'\mathbf{x} + \gamma_z'\mathbf{z} + \zeta \qquad (1)$$

where x and z are vectors of variables and ζ is a normally distributed residual with zero mean, variance ψ, and where ζ is independent of $\mathbf{x}$ and $\mathbf{z}$.

In addition to predicting η, we specify an influence from the z variable for a certain item to the response for that particular item. Whereas each item's z variable influences the item response through the η variable, this part of the model concerns the direct influence from the z to the item, over and above that which goes through η. It is convenient to express the direct influence of the z variables on the items using a latent response variable formulation, where

$$y_j = 0, \text{ if } y_j^* < \tau_j \qquad (2)$$

1, otherwise where τ_j is a threshold parameter defined on the continuous latent response variable y_j^*;

$$y_j^* = \lambda_j\eta + \beta_j z_j + \epsilon_j\,. \qquad (3)$$

The latent response variable may be viewed as the specific skill needed to solve the corresponding item correctly; when the latent response variable exceeds a threshold, the item is correctly answered. We assume that ϵ_j is a residual with mean zero that is independent of η and z. By adding the assumption that ϵ_j has a normal distribution, the standard normal ogive model of IRT is obtained, except that OTL is allowed to have direct influence on the item.

In effect, this specification allows items to have different difficulty for different OTL levels (cf. Muthén, 1989). The shift in difficulty is provided by the β parameter. The parameters of this model may be translated into those of standard IRT so that each item obtains one discrimination parameter value and, in the present case of two OTL categories, two difficulty parameter values. The formulas for the translation are as follows. The conditional variance of y_j^* given the x and z variables is standardized to 1, resulting in a residual (ϵ) variance $\theta_{jj} = 1 - \lambda_j^2\,\psi$. Let the mean and variance of η be denoted μ_η and $\sigma_{\eta\eta}$, respectively. It can then be shown that the two-parameter normal ogive parameters a (discrimination) and b (difficulty) for item j can be written as

$$a_j = \lambda_j\theta_{jj}^{-1/2}\sigma_{\eta\eta}^{1/2}, \qquad (4)$$

$$b_{jk} = [(\tau_j - \beta_j z_k)\lambda_j^{-1} - \mu_\eta]\sigma_{\eta\eta}^{-1/2}, \qquad (5)$$

In these formulas, the trait has been standardized to mean zero and variance one. The estimated values of a and b may be obtained by inserting model parameter estimates in (4) and (5), where the sample means, variance, and covariances for the x and the z are also used to compute the estimated μ_η and $\sigma_{\eta\eta}$. For each item, we can then obtain two estimated item characteristic curves and compute differences between these curves. In this paper, we will choose to use the simple index (called D) discussed by Linn, Levine, Hastings, and Wardrop (1981), where squared probability differences are added up over the trait range −3 to +3.

Inserting (1) in (3) gives the so-called reduced-form for the regression of the y^* on the x and z. These are probit regressions, where the model imposes restrictions on the probit slopes and residual correlations. The slopes are expressed by the λ, β, and γ parameters of the model, whereas the residual correlations also involve the remaining parameter ψ for the residual variance. The parameters may be estimated by fitting the model to these probit regression slopes and residual correlations.

Muthén (1987) describes the LISCOMP computer program which builds on theory in Muthén (1984) and encompasses the present type of model. The technical details of our analysis will not be discussed here. Instead, the steps of the analysis will be outlined. The probit slopes and the probit residual correlations correspond to different model parts in the LISCOMP framework and can be analyzed together or separately. In the present case, there are 40 y variables (items) and a total of about 50 x and z variables. This is a large model; the full model was fitted by using the probit slopes only to keep the computations manageable.

In the first step, LISCOMP was used to estimate the probit slopes by specifying a dichotomous variable type for the y, inducing probit modeling, and requesting slope (LISCOMP model, Part 2) statistics only. In the second step, the estimated slopes were read as sample statistics to which the λ, β, and γ parameters were fitted by structural equation modeling using the appropriate LISCOMP structural model part (Part 2). Unweighted least-squares estimation was used to simplify the computations in this step. To make this estimation independent of x variable scale, the slopes of Step 1 were computed for x variables transformed to the 0–1 range. In Step 3, an estimate of the residual variance ψ in the ability variable was obtained by analysis of a subset of about half of the items showing particularly good measurement properties. Here, Part 3 of the LISCOMP structural model was fitted to sample probit residual correlations.

Analysis Results

Preliminary analyses were performed by standard IRT techniques. Using the two-parameter logistic model and marginal maximum likelihood estimation provided by the BILOG program (Mislevy & Bock, 1984), it was revealed that Item 10 was very hard and had deficient measurement properties. The subsequent analyses were performed with only 39 test items. An item factor analysis including a scree plot strongly supported the notion of unidimensionality for this set of items.[2]

The estimation of the influence of the background variables on the ability will be discussed first. Next, the estimates of the measurement parameters relating the item responses to the ability will be presented. Finally, we will turn to the estimates of primary concern in this paper—namely, those representing the effect of instructional sensitivity.

Relating the Ability to Background Variables

The estimates from the regression of the trait on the background variables are given in Table 3. Although standard errors of estimates are not provided for this model, generalized least-squares estimation on a subset of items indicates that estimates larger than .10 are most likely statistically significant. It is seen that the pretest variable related to arithmetic dominates the prediction of spring math ability. This is natural because this is the area of mathematics best covered up to eighth grade and because performance on these kinds of tasks influences the selection of students into more advanced math classes where they get further training that enhances their ability. One may note that the prearithmetic variable correlates 0.76 with the posttest sum of correct answers. Among nonpretest variables, finding mathematics useful is the most important one.

Table 3

Effects of Background Variables on Ability

Variables	Effect Estimates
Pretests	
PREALG	.45
PREMEAS	.75
PREGEOM	.45
PREARITH	2.53
Class type	
REMEDIAL	-.22
ENRICHED	.18
ALEGEBRA	.10
Attitudes	
MORED	.31
USEFUL	.53
ATTRACT	.19
Demographics	
FEMALE	.03
NONWHITE	-.16
Family Background	
FAED	.08
MOED	-.05
LOWOCC	.05
HIGHOCC	.08
MISSOCC	.04

The γ-parameter estimates for the effect of OTL variables on the math ability will not be presented here. Overall, the effects are negligible. The prediction of math ability by fall measurements is quite successful in that the estimated proportion of variation in math ability explained by the various background variables is 76%.

When using the SIMS data to illustrate the approach to assessing instructional item sensitivity, Muthén (1989) included the OTL variables only and not the other background variables used here. If we assume that our present model including such background variables is true, omitting these other background variables would lead to biased estimates of the item parameters and their instructional sensitivity. However, we have found that such biases are small for these data, probably due to the rather small correlations between the OTL variables and the other background variables. This is a useful finding for situations where pretests, or other early performance measures, are not available, and OTL is expected to correlate little with such measures. In situations where OTL correlates more highly with socioeconomic background variables, the study of OTL effects is more complex, because inclusion of such background variables may make the OTL effect appear weaker.

Relating the Items to the Ability

The measurement of the trait η is reflected in the λ parameters representing the slopes (factor loadings) in the regressions of the latent response variables y^* on the trait η. The estimates of these are given in Table 4, which also contains the estimated values of the threshold τ and of the corresponding IRT parameters, one a and two b for each item calculated as in (4) and (5). Table 4 also contains the corresponding estimates of IRT parameters a and b as obtained by standard analysis, here carried out by marginal maximum likelihood in the BILOG program (Mislevy & Bock, 1984).

Table 4 shows that Items 3, 6, 7, 17, 19, 21, and 39 have λ values less than or equal to .45 and are not good measurements of the math ability trait. It is interesting to note that six of these seven items have geometric or spatial content and that, with the exception of Item 17, all of these items had NTL values of at least .25.

It is also interesting to note that standard IRT estimation of a and b parameters, compared to our approach, gives results that are rather similar for a but quite different for b. Two explanations may be offered for this. One is that our results come from a model that extends the standard IRT to background variables, giving a fuller description of the trait where it is determined not only by item performance but also by predictors thereof. In statistical terms, the model is stronger in that the notion of unidimensionality is extended to not only explain item interrelations but also relations between items and predictors. Because this is largely a matter of using more information for estimation, the second reason relates to bias in the standard IRT estimation due to use of the wrong model. Under a model that allows for direct OTL influence on the items,

Table 4

Measurement Parameter Estimates

			Model-based IRT parameters			Standard IRT parameters	
				b			
Item	Threshold	Loading	a	OTL	NTL	a	b
1	2.52	.57	.60	1.26	1.70	.65	.33
2	1.91	.84	.94	-.45	-.46	.70	-.47
3	2.30	.43	.44	1.70	2.58	.52	1.79
4	2.32	.66	.71	.84	.77	.56	.92
5	2.11	.57	.60	.49	.98	.60	.28
6	1.08	.40	.42	.14	-.05	.44	-.31
7	.91	.44	.46	-.99	-.66	.67	-.77
8	.71	.57	.61	-1.57	-1.48	1.00	-1.78
9	1.78	.66	.71	-.02	-.04	.71	-.11
11	2.75	.66	.71	1.47	1.41	.78	.83
12	1.72	.60	.64	-.12	.13	.58	.32
13	2.08	.71	.76	.24	.21	.85	-.89
14	1.68	.70	.75	-.32	-.32	.73	-.49
15	1.31	.49	.52	-.15	-.06	.42	1.22
16	1.77	.72	.78	-.84	-.28	.80	-.33
17	1.47	.43	.45	-.67	.66	.59	-.45
18	2.49	.53	.55	1.66	1.96	.63	-.04
19	1.72	.41	.42	1.29	1.48	.44	1.05
20	1.64	.80	.89	-.22	-.67	1.06	-1.07

(table continues)

the use of a standard IRT model ignores both student heterogeneity in the item parameters and the fact that in addition to the trait the OTL influence also causes dependency among the items.

Instructional Sensitivity

Of greatest interest in this paper are the estimated β parameters representing the direct effects of OTL on the item performance, thereby indicating instructional sensitivity in the items. The estimated β and the corresponding

Table 4 (continued)

			Model-based IRT parameters			Standard IRT parameters	
				b			
Item	Threshold	Loading	a	OTL	NTL	a	b
21	2.66	.45	.47	2.91	3.13	.53	.89
22	2.50	.78	.85	.22	.49	.83	-.39
23	2.23	.71	.77	.38	.41	.84	.11
24	1.95	.95	1.09	-.72	-.67	.86	-.28
25	1.83	.88	.99	-.59	-.64	.75	.16
26	.32	.52	.55	-2.08	--[a]	.45	-.76
27	1.91	.59	.62	.53	.50	.68	-.30
28	2.00	.75	.82	-.06	-.08	.72	-.52
29	1.14	.65	.69	-1.02	-.96	.93	-1.04
30	1.54	.49	.51	.51	.41	.48	.56
31	2.44	1.08	1.28	-.45	--[a]	.97	-.46
32	1.90	.89	1.01	-.60	--[a]	.83	.21
33	1.86	.58	.61	.38	.49	.61	-.00
34	3.05	1.00	1.17	.72	.31	.77	.45
35	1.14	.46	.48	-.76	-.25	.58	-.45
36	1.20	.76	.83	-1.42	-1.13	-.68	-.28
37	1.98	.89	1.01	-.54	-.50	.83	.51
38	2.83	.85	.95	-.28	.57	.71	-.06
39	.60	.34	.35	-2.90	-.95	.49	-.21
40	.91	.52	.52	-1.38	-.98	.54	.16

[a] No student is available in the NTL category

measures of distance between the probability curves (item characteristic curves) are given in the rightmost part of Table 1. The implications of the estimates in this part of the table are best understood by a discussion of the items that show substantial instructional sensitivity.

Consider first Item 17. This is a geometry item, containing pictures of angles, that for the correct solution requires knowledge of the definition of an acute angle. From Table 1, we note that 13% have had no OTL for this item but that

38% get the item right, whereas 62% get it right with OTL (this year or prior years). The β estimate for OTL is positive reflecting the extra advantage, over and above what the trait level would predict, of having OTL versus not having OTL. Note that while the proportion correct for an item is an estimate of marginal probability, the β effect corresponds to a change in conditional probability given the trait and is therefore the appropriate measure of instructional sensitivity. Several items have large differences in proportion correct for OTL versus NTL but have negligible β effects. In order to gauge the importance of the corresponding shifts in the conditional probabilities, Figure 2 shows the standardized probability curves over the trait range -3 to $+3$.

For an average trait value of 0, the extra advantage of OTL is estimated as an approximate increase of 0.15 for the probability of a correct answer. The corresponding curve distance (D value) is .32.

A working hypothesis for a particularly strong reason for instructional sensitivity is that the item is definitional in nature and represents early learning on the topic of angles. It is therefore rather hard for students who have not been exposed to it but rather easy for students who have been exposed to it. A harder item may show less instructional sensitivity because, even with OTL, many students may get it wrong. An item such as Number 17 may be less valuable as an indicator of a more general trait than as an indicator of exposure in a certain limited area. From Table 4, we note that Item 17 is among the group of items that we identified as having rather poor measurement qualities, with an estimated λ value of .43 (an estimated a value of .45).

Consider next Item 39. This item shows a point in a coordinate system and asks for its coordinates. Here, a rather large group of 30% have no OTL. In terms of proportion correct, the item seems easy for the OTL category (0.63) but hard for the no OTL (0.33). There is a substantial difference between the estimated probability curves for OTL versus no OTL (0.35). Like Item 17, this instructional sensitivity in Item 39 seems to correspond to definitional learning such that the item becomes quite easy when the student is exposed to this knowledge. But the item is a poor indicator of the trait (see Table 4)—in fact, the worst one. The estimated probability curve shows a small discrimination (slope). This would mean that getting the item correct involves little general math ability and only indicates a specific knowledge of the definition, a plausible explanation for this item.

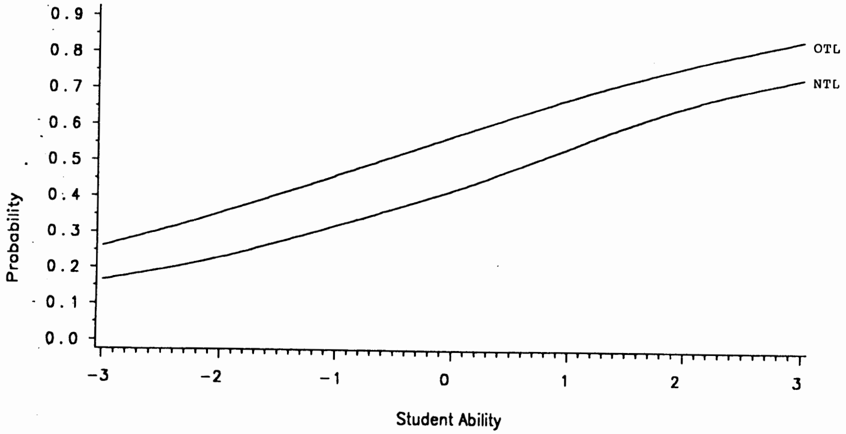

FIGURE 2. *Estimated probability curves for Item 17*

Other items also show substantial instructional sensitivity and may further support the hypothesis of introductory definitional content. To solve Item 38, a student needs to know the definition of percentage and to apply a straightforward arithmetic operation; Item 16 calls for knowledge about multiplying negative integers and parentheses. But unlike Items 17 and 39 discussed above, Items 38 and 16 provide good measurements of the trait.

Conclusions

The proposed methodology represents a new way to study the instructional sensitivity of achievement items. Given sufficiently rich data, instructionally sensitive items can be detected while at the same time gaining information about the achievement process through the estimation of a comprehensive model that goes well beyond those of standard IRT analytical methods for examining achievement test data.

The exact nature of the benefits to be gained from estimating the effects of instructional opportunities on the latent ability depends on the specific empirical context in which the methodology is employed. Naturally, the heterogeneity of the pool of achievement items and of the student population tested matters. What also matters is adequacy of the specification of the model of achievement and of the measurement of instructional opportunities and other characteristics.

In the present case, there was considerable heterogeneity in the mathematics instruction experiences of students; some students were still enrolled in remedial instruction dominated by arithmetic operations with integers and common and decimal fractions when others were enrolled in elementary algebra classes. The set of test items broadly spanned topics typically covered by the end of elementary algebra instruction. Against this backdrop, the model examined here featured parameters estimating the influence of student background and OTL content pertinent to each specific test item on a single latent mathematics ability trait and the effects of the mathematics ability trait and the item-specific OTL on the difficulties of test items.

Under these modeling conditions, item-specific OTL had limited impact on the latent variable representing mathematics ability once student background variables (which included pure mathematics performance) were controlled. However, for selected test items, there were strong direct effects of latent mathematics ability. In other words, the general, presumably more stable achievement trait, was insufficient to account for performance on these items. According to standard IRT analysis methods, either the IRT results would be biased by the inclusion of items or items would have to be eliminated to avoid violation of IRT assumptions. Neither prospect is attractive.

In our opinion, concerns about IRT bias when ignoring effects of OTL and other background variables should not be exaggerated. Preliminary studies of ability estimation (Muthén & Short, 1988) indicate that ability estimates are little affected by ignoring such heterogeneity. We would expect that equating would be even less adversely affected. The issue is not so much what goes wrong in conventional IRT, but what is overlooked and not uncovered in conventional IRT.

Clearly, the present analysis provides a more detailed way to examine the influence of instruction on responses to test items, a matter of considerable interest in developing achievement tests and interpreting test results. In the present case, certain test items representing early stages of learning about selected mathematical topics were particularly sensitive to specific instruction. Individual differences represented within the single latent mathematics ability did not adequately account for performance differences on these items.

This methodology has also been successfully applied to all rotated forms of the eighth-grade SIMS test (Kao, 1990), identifying further instructionally sensitive items using OTL composites. What next steps to take in response to the identification of instructionally sensitive items is unclear. A possibility for generalization is to consider employing a multidimensional latent achievement model to represent the domain of test items. Incorporating specific latent factors representing instructionally important curriculum segments within the psychometric model is both theoretically and practically desirable. Presumably, differential instructional exposure should then influence the specific factors. Under such conditions, any residual direct effects of OTL on item performance represent teaching to the specifics of the test, a typically undesirable instructional strategy. We are currently exploring the possibility of applying models with multidimensional latent achievement traits with the SIMS database (e.g., Gold, 1990; Kim, 1990; Muthén, 1990).

Given psychometric methodology that can better tie test item performance to both ability and instruction, the proper measurement and measurement modeling of instruction is highlighted. The above analyses utilized a class-level, and rather crude, OTL variable reported by the teacher. It is recognized that the mixture of student-level responses and class-level OTL information creates multilevel, or hierarchical observations, a problem that we were forced to ignore in our analyses. With few classes in an OTL category, measurement error in the teacher-reported OTL may have strong biasing effects. The class-level information may also be incorrect for a given student. Student-level OTL is available, but it may contain even more measurement error. Further substantive research needs to find ways to properly combine information of several kinds in order to provide more reliable and informative instructional student background.

Notes

[1] In preliminary analyses, we considered using three-category OTL measurements corresponding to OTL this year, OTL prior year(s), and no OTL. However, this approach was abandoned in favor of using dichotomous OTL for the following conceptual and technical reasons. First, the prior year effect may be hard to estimate because prior year OTL is not distinctly defined. It may refer to OTL more than a year ago as well as OTL late in the previous year. Second, many items showed low percentages for the prior year OTL, leading to unstable estimates. Third, use of the three-category OTL variables leads to high correlations between several items' prior year and this year's OTL measurements, resulting in multicollinearity among the predictors.

Preliminary analyses also found probable misreporting by a teacher. For two items, the no OTL category was made up of 24 students from one class who all got the items right. We plotted the sum of correct answers versus the sum of the dichotomously scored OTL and found this class to be a distinct outlier with very high performance and rather low OTL. For these two reasons, this class was deleted from the analyses to be presented.

[2] Note that this is only a rough assessment of the dimensionality of the items, because the items may correlate not only due to the trait but also due to the OTL influence.

Appendix
The 40 Core Items

Item #	Question	Content Classification	Behavioral level
1	2 meters + 3 millimeters is equal to	M	I
2	1/5 is equal to	A	II
3	If 5x + 4 = 4x - 31, then x equal to	A	I
4	Four 1-liter bowls of ice cream were set out at party	F	III
5	Which is closest appro. to area of	M	II
6	Area of the shaded figure to nearest square unit, is	M	III
7	Diagram shows a cardboard cube which has been cut along	G	III
8	Lenght of ab is 1 unit. The best estimate	M	II
9	On above scale reading indicated by the arrow is	M	III
10	A solid plastic cube with edges with 1 cm long weighs 1 gram.	M	IV
11	On a number line two points a and b are given. The coordinate	G	II
12	A painter is to mix green and yellow paint in the ratio of	P	III
13	If p = lw and if p = 12 and 1 = 3, then w is equal to	A	I
14	A model boat is built to scale so that it is 1/10 as long as	P	III
15	The value of 0.2131 x 0.02958 is approx.	F	II
16	(-2) x (-3) is equal to	A	I
17	Which of the indicated angles is acute?	G	I
18	If 4x/12 = 0, then x is equal to	A	I
19	The length of the circumference of the circle with center at	G	IV
20	In a discus-throwing competition, the winning throw was 61.60	F	III
21	In the above diagram, triangles ABC and DEF are congruent	G	III

(continued)

Appendix (continued)

22	(Triangle with 2 angles given) X is equal to	G	II
23	A square is removed from the rectangle shown. What is the	M	III
24	Cloth is sold by the square meter. If 6 square meters of	P	III
25	The air temperature at the foot of a mountain is 31 degrees	A	III
26	0.40 x 6.38 is equal to	F	I
27	A shopkeeper has x kg of tea in stock. He sells 15 kg and	A	III
28	In the figure the little squares are all the same size and	F	II
29	The distance between two towns is usually measured in	M	I
30	The table below compares the height from which a ball is	A	II
31	2/5 + 3/8 is equal to	F	I
32	7 3/20 is equal to	F	I
33	In a school of 800 pupils, 300 are boys. The ratio of the	P	II
34	What is 20 as a percent of 80	P	I
35	The sentence 'A number x decreased by 6 is less than 12' can	A	II
36	30 is 75% of what number?	P	I
37	Which of the points a, b, c, d, e, on this number line	F	II
38	20% of 125 is equal to	P	I
39	What are the coordinates of point p?	G	I
40	Triangles PQR and STU are similar. How long is SU?	G	III

KEY:

Content Classification
F = Fractions
P = Ratio proportion percent
A = Algebra
G = Geometry
M = Measurement
I = Integers
S = Statistics
A = Algebra
G = Geometry
M = Measurement
I = Integers
S = Statistics

Behavioral Level
I = computation
II = comprehension
III = application
IV = analysis

References

Airasian, P. W., & Madaus, G. F. (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* (Vol. 6, pp. 3682–3686). 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, IL: Stipes.

Gold, Karen F. (1990). *Applications of hierarchical confirmatory factor models: Assessment of structure and integration of knowledge exhibited in achievement data.* Unpublished doctoral dissertation, University of California, Los Angeles.

Haertle, E., & Calfree, R. (1983). School achievement: Thinking about what to test. *Journal of Educational Measurement, 20,* 119–132.

Kao, Chih-Fen (1990). *An investigation of instructional sensitivity in mathematics achievement test items for U.S. eighth grade students.* Unpublished doctoral dissertation, University of California, Los Angeles.

Kim, Suk-Woo (1990). *Gender and OTL effects on mathematics achievement for U.S. SIMS 12th grade students.* Unpublished doctoral dissertation, University of California, Los Angeles.

Lehman, J. D. (1986). *Opportunity to learn and differential item functioning.* Unpublished doctoral dissertation, University of California, Los Angeles.

Leinhardt, G. (1983). Overlap: Testing whether it is taught. In G. F. Madaus (Ed.), *The court, validity and minimum competency testing* (pp. 115–132). Boston: Kluwer-Nijhoff.

Leinhardt, G., Seewald, A. M. (1981). Overlap: What's tested, what's taught? *Journal of Educational Measurement, 18,* 85–96.

Linn, R. L. (1983). Curriculum validity: Convincing the courts that it was taught without precluding the possibility of measuring it. In G. F. Madaus (Ed.), *The court, validity and minimum competency testing* (pp. 115–132). Boston: Kluwer-Nijhoff.

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

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

Lord, F. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

McKnight, C. C., Crosswhite, F. J., Dossey, J. A., Kifer, E., Swafford, J. O., Travers, K. J., & Cooney, T. J. (1987). *The underachieving curriculum: Assessing U.S. school mathematics from an international perspective.* Champaign, IL: Stipes.

Mehrens, W. A., & Phillips, S. E. (1986). Detecting impacts of curricular differences in achievement test data. *Journal of Educational Measurement, 23,* 147–158.

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). Recent developments in item response theory with implications for teacher certification. *Review of Research in Education, 14,* 239–275.

Mislevy, R. J., & Bock, R. D. (1984). *BILOG II. Item analysis and test scoring with binary logistic models. User's guide* [Computer program]. Mooresville, IN: Scientific Software.

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

Muthén, B. (1987). *LISCOMP. Analysis of linear structural equations with a comprehensive measurement model. User's guide* [Computer program]. Mooresville, IN: Scientific Software.

Muthén, B. (1988). 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 Associates.

Muthén, B. (1989). Using item-specific instructional information in achievement modeling. *Psychometrika, 54,* 385–396.

Muthén, B. (1990, April). *Differences in change for groups of individuals and items: A latent variable approach.* Paper presented at the annual meeting of the American Educational Research Association, Boston.

Muthén, B., & Short, L. (1988). *Estimation of ability by IRT models allowing for heterogeneous instructional background.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Schmidt, W. H., Porter, A. C., Schwille, J. R., Floden, R. E., & Freeman, D. J. (1983). Validity as a variable: Can the same certification test be valid for all students? In G. F. Madaus (Ed.), *The court, validity and minimum competency testing* (pp. 133–151). Boston: Kluwer-Nijhoff.

Authors

BENGT O. MUTHÉN is Professor, Graduate School of Education, UCLA, Los Angeles, CA 90024-1521. *Degree:* PhD, University of Uppsala, Sweden. *Specializations:* analysis of categorical data and structural equation modeling.

CHIH-FEN KAO is a Graduate Student, Graduate School of Education, UCLA, Los Angeles, CA 90024-1521. *Degree:* PhD, UCLA. *Specializations:* educational measurement and psychometrics.

LEIGH BURSTEIN is Professor, Graduate School of Education, Center for Research on Evaluation, Standards, and Student Testing, UCLA, Los Angeles, CA 90024-1521. *Degrees*: BA, Michigan State; PhD, Stanford University. *Specialization*: quantitative methodology for educational research and policy analysis.

Journal of Educational Measurement
Spring 1991, Vol. 28, No. 1, pp. 23–35

Gender Differences in Multiple-Choice Tests: The Role of Differential Guessing Tendencies

Gershon Ben-Shakhar and **Yakov Sinai**
Hebrew University, Jerusalem, Israel

The present study focused on gender differences in the tendency to omit items and to guess in multiple-choice tests. It was hypothesized that males would show greater guessing tendencies than females and that the use of formula scoring rather than the use of number of correct answers would result in a relative advantage for females. Two samples were examined: ninth graders and applicants to Israeli universities. The teenagers took a battery of five or six aptitude tests used to place them in various high schools, and the adults took a battery of five tests designed to select candidates to the various faculties of the Israeli universities. The results revealed a clear male advantage in most subtests of both batteries. Four measures of item-omission tendencies were computed for each subtest, and a consistent pattern of greater omission rates among females was revealed by all measures in most subtests of the two batteries. This pattern was observed even in the few subtests that did not show male superiority and even when permissive instructions were used. Correcting the raw scores for guessing reduced the male advantage in all cases (and in the few subtests that showed female advantage the difference increased as a result of this correction), but this effect was small. It was concluded that although gender differences in guessing tendencies are robust they account for only a small fraction of the observed gender differences in multiple-choice tests. The results were discussed, focusing on practical implications.

A great deal of research has been devoted during the past 3 decades to the issue of gender differences in cognitive abilities. In their extensive review of the literature, Maccoby and Jacklin (1974) concluded that three consistent differences between males and females can be detected during puberty. Females outscore males in tests of verbal ability. The female advantage is revealed throughout the whole range of verbal abilities, and it averages about .25 standard deviation (*SD*). Males perform better than females on tests of both visual-spatial and mathematical ability. The gender differences in these two areas are between .4 and .5 *SD* in high school ages.

The conclusions drawn by Maccoby and Jacklin (1974) were criticized by some researchers (e.g., Block, 1976; Sherman, 1978), but there is little doubt that some consistent correlations between gender and performance exist on various multiple-choice tests. Hyde (1981) conducted a meta-analysis of gender cognitive differences, using the set of studies reviewed by Maccoby and Jacklin

We thank the Hadassah Counselling Institute and the National Institute for Testing and Evaluation for providing the data for this article and Maya Bar-Hillel, Michal Beller, Naomi Gafni, and the anonymous reviewers for their helpful comments on an earlier version.