DOCUMENT RESUME

ED 338 678 TM 017 495

AUTHOR	Muthen, Benjt O.; And Others
TITLE	Instructional Sensitivity in Mathematics Achievement Test Items: Application of a New IRT-Based Detection Technique.
INSTITUTION	Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.
SPONS AGENCY	Office of Educational Research and Improvement (ED), Washington, DC.
REPORT NO	CSE-TR-292
PUB DATE	Jan 88
CONTRACT	G0086-003
NOTE	16p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 5-9, 1988).
PUB TYPE	Reports - Research/Technical (143) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Achievement Tests; Educational Background; Elementary Secondary Education; Grade 8; Instructional Effectiveness; Item Bias; *Item Response Theory; Junior High Schools; *Junior High School Students; Mathematical Models; Mathematics Achievement; Mathematics Instruction; *Mathematics Tests; Standardized Tests; Student Characteristics; *Test Items
IDENTIFIERS	*Instructional Sensitivity; *Opportunity to Learn; Second International Mathematics Study; United States

ABSTRACT

Item response theoretic methods are applied to the measurement of achievement of students from various instructional backgrounds. This extended item response theory (IRT) approach serves as a tool for studying instructional bias, or instructional sensitivity. The model maintains the form of an IRT model, but has parameters that quantify the extent of the effect attributed to opportunity to learn (OTL). The technique is applied to detect instructional sensitivity using the Second International Mathematics Study (SIMS) set of 40 core items for eighth-graders in the United States. The SIMS data came from about 280 schools and about 7,000 students measured at the end of spring 1982. The achievement test used contained 180 items in the areas of arithmetic, algebra, geometry, and measurement distributed among four test forms. In the SIMS data, there was considerable heterogeneity in the mathematics instruction experiences of students. The model features 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. The analysis indicates that certain test items representing early stages of learning about selected mathematical topics were particularly sensitive to specific instruction. An 18-item list of references is included. (SLD)

C·R·E·S·S·T

Center for Research on Evaluation, Standards, and Student Testing

INSTRUCTIONAL SENSITIVITY IN MATHEMATICS ACHIEVEMENT TEST ITEMS: APPLICATION OF A NEW IRT-BASED DETECTION TECHNIQUE

CSE Technical Report 292

Bengt O. Muthen
Chih-Fen Kao
Leigh Burstein

UCLA Center for Research on Evaluation, Standards, and Student Testing

INSTRUCTIONAL SENSITIVITY IN MATHEMATICS ACHIEVEMENT TEST ITEMS: APPLICATION OF A NEW IRT-BASED DETECTION TECHNIQUE

CSE Technical Report 292

Bengt O. Muthen
Chih-Fen Kao
Leigh Burstein

UCLA Center for Research on Evaluation,
Standards, and Student Testing

January, 1988

The research reported herein was conducted with partial support from the U.S. Department of Education, Office of Educational Research and Improvement, pursuant to Grant No. G0086-003. However, the opinions expressed do not necessarily reflect the position or policy of this agency and no official endorsement by this agency should be inferred.

This paper was written as part of a research project sponsored by the UCLA Center for Research on Evaluation, Standards, and Student Testing (CRESST). It was prepared for presentation at the Annual Meeting of the American Educational Research Association, New Orleans, April 1988.

Please address inquiries to: CSE Dissemination Office, UCLA Graduate School of Education, 405 Hilgard Avenue, Los Angeles, California, 90024-1521

Introduction[1]

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 (e.g., 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 and Madaus (1983), Haertle and Calfee (1983), Linn (1983), Schmidt, Porter, Schwille, Floden, Freeman (1983), Leinhardt (1983), Leinhardt and Seewald (1981), Mehrens and Phillips (1986), and Miller (1986).

The research reported here extends our developments of item response theoretic methods for achievement of heterogeneous groups of students (Muthen, 1987a,b). 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 difference 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 applicability of the methodology for research and practice within the domain of large-scale instructional testing.

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, 1988) 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 (1986) 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 and Miller-Linn 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 group criteria may lead to different conclusions.

Standard IRT techniques assume that instruction increases the item

---

[1] The authors would like to thank Michael Hollis and Suk-Woo Kim for valuable research assistance.

performance through an increase in the latent trait level, while 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. Muthen (1987a) 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. Muthen's extended IRT model may serve as a better tool for studying the instructional bias, or, as we will term it, instructional sensitivity. His model maintains the form of an IRT model, but in addition his parameters which quantify the extent of the effect attributed to OTL. Using similar modeling, Muthen (1987b) 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 which preclude the influences from auxiliary variables are challenged and examined in Muthen's model.

Muthen's model may be briefly described as follows. Building on the statistical theory of Muthen (1984), Muthen (1987b) proposed a new extension of IRT modeling that controls for student background differences by including background variables as covariates. Further extending this methodology, Muthen (1987a) 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 Muthen (1987a) technique for detecting instructionally sensitive items was illustrated with a very small set of 8 algebra items from the US sample of eighth graders in Second International Mathematics Study (SIMS), Crosswhite, Dossey, Swafford, McKnight, and Cooney (1985). The aim of this paper 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 were selected proportional to size; a proability sample of schools were selected proportional to size within school district: and two classes were randomly selected within each school yielding a total of about 280 schools and about 7,000 students measured at the end of spring 1982. The achievement test contained 180 items in the areas of arithmetic, algebra, geometry, 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 obtained as follows. For each item teachers were asked two questions regarding student opportunity to learn.

Question 1:

"During this school year did you teach of review the mathematics needed to answer the item correctly?"

1. No
2. Yes
3. No response

Question 2:

"If in this school year you did not teach or review the mathematics needed to answer this item correctly, was it mainly because?"

1. It had been taught prior to this school year
2. It will be taught later (this year or later)
3. It is not in the school curriculum at all
4. For other reasons
9. No response

Given these responses, opportunity-to-learn (OTL) level will be classified as follows:

OTL: Question 1 = 2, question 2 = 9
or question 1 = 1, of 9 and question 2 = 1
No OTL (NTL): Question 1 = 1, question 2 = 2, 3, 4, or 9
or question 1 = 9, question 2 = 2, 3, or 4

Other response combinations lead to the elimination of the observation.

The percentage distribution of OTL categories for all 40 items are given in the left-most part of Table 1 (See Appendix A) together with proportion correct.

It seems that the percentage of students having no OTL (category NTL) varies greatly across the items. With the exception of 5 items, having had OTL is most common. However, about 1/3 of the items show NTL proportions larger than 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. Students' reporting may partly reflect his/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.

In preliminary analysis, 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 of all, the prior year effect may be hard to estimate since prior year OTL is not distinctly defined, but 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 category, leading to unstable estimates. Third, use of the three-category OTL variables lead to high correlations between several items' prior year and this year 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. Plotting the sum of correct answers versus the sum of the dichotomously scored OTL, this class was found 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.

In addition to the above item-specific OTL information, instructional background information common to all items is available in the SIMS in the form classification of each mathematics class into one of four types, basic of remedial arithmetic (REMEDIAL), general of typical mathematics (TYPICAL), pre-algebra or enriched (ENRICHED), and algebra (ALGEBRA). This classification is based of 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 "preset" measurements of mathematics, family background, educational aspiration, attitudes toward mathematics, gender, ethnicity; see Table 2 and also Muthen (1987b).

The premeasurements were only collected for part of the sample. 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.

The Model

Following Muthen (1987a), detection of instructionally sensitive items among the items is achieved by estimation of the following model. A diagrammatic representation of the model is given in Figure 1 (See Appendix A). 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. Muthen (1987a) suggests the term "latent performance level." We want to study the effect of OTL on the item performance since 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 predictor for math ability, where OTL influence on math ability is specified as partial effects, holding other background variables constant.

Spring math ability is here taken 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. The majority of the background variables are assumed to only correlate because of their common influence on the math ability variable.

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. 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 which 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.

More technically, 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 class type). In our analysis we will create an OTL dummy variable for each item j, $z_j = 1$ represents OTL. The variable η, say. We specify the linear regression model.

$$(1)\quad \eta = \gamma_x' x + \gamma_z' z + \zeta$$

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 x and z.

In addition to the part of predicting η, specify an influence from the z variable for a certain item to the response for that particular item. While 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

$$(2)\quad y_j = 0, \text{ if } y^*_j < \tau_j$$
$$\qquad\quad 1, \text{ otherwise}$$

where τ_j is a threshold parameter defined on the continuous latent response variable $y^*_j = \Lambda_j \eta + B_j z_j + \varepsilon_j$.

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 the z's. 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. Muthen, 1987a). The shift in difficulty is provided by the B parameter. The parameters of this model may be translated to those of standard IRT, so that each item obtains one discrimination parameter value and, in the present case of 2 OTL categories, 2 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 $\Theta_{jj} = 1 - \Lambda^2_j \psi$. Let the mean and variance of h be denoted μ_η and $a_{\eta\eta}$, 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

(4) $a_j = \Lambda_j \Theta_{jj}^{-1/2} a_{jj}^{1/2}$,

(5) $b_{jk} = [(\tau_j - B_j z_j)\Lambda_j^{-1} - \mu_\eta] a_{\eta\eta}^{-1/2}$,

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's and the z's are also used to compute the estimated μ_η and $a_{\eta\eta}$. For each item we can then obtain 2 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, Wardop (1981), where squared probability differences are added up over the trait range -3 to +3.

Inserting (1) in (2) gives the so-called reduced-form for the regression of the y^*'s on the x's and z's. These are profit regressions, where the model imposes restrictions on the probit slopes and residual correlations. The slopes are expressed by the γ, B, and Λ parameters of the model, while the residual correlations also involve the remaining parameter ψ for the residual variance. The parameters may be estimated by fitting the model to the probit regression slopes and correlations.

Muthen (1987c) describes the LISCOMP computer program which builds on theory in Muthen (1984) and encompasses the present type of model. The technical details of our analysis will not be discussed here. The slopes and the 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 yields potentially 2,000 slopes and 780 correlations to fit the model to. This yields potentially 2,000 slopes for the regression of each y variable on these regressions involve a nonlinear maximum likelihood probit regression with 50 x variables on 3,724 observations and is therefore computationally burdensome. In order to ease the computational burden, only the slope part of the LISCOMP framework will be used to estimate the γ's, B's, and Λ's. The ψ parameter will be estimated using both the slope and correlation part through a separate analysis on a subset of about half of the items showing particularly good measurement qualities. To further simplify computations, the fitting of the model of Figure 1 is carried out by unweighted least-squares. Still, the computations are heavy in that they involve the estimation of over 100 γ's, B's, and Λ's. While the unweighted least-squares estimator does not provide standard errors of estimates, follow-up analyses on subsets of items by generalized least-squares will give indication of the magnitude of estimates needed for statistically meaningful values.

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 strongly supported the notion of unidimensionality for this set of items. A scree plot of the latent roots for the tetrachoric correlation matrix is given in Figure 2. Note that this is only a rough assessment of the dimensionality of the items since the items may correlate not only due to the trait but also due to the OTL influence.

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 indicate that estimates larger than pretest variable related to arithmetic dominates the prediction of spring math ability. This is natural since this is the area of mathematics best covered up to eighth grade and since performance on these kinds of tasks influence 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 non-pretest variables, finding mathematics useful is the most important one.

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 portion 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, Muthen (1987a) only included the OTL variables and not the other background variables used here. Given our present model, 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 the 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.

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's 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). The wording of each of the 40 items is given in Appendix B.

Table 4 shows that items 3, 6, 7, 17, 19, 21, 39 have L-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 with exception of item 17, all these items had NTL values of at least .25.

It is also interesting to note that standard IRT estimation of a and b parameters as 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 strong in that the notion of undimensionality is extended to not only explain item interrelations but also relations between items and predictors. While 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, the use of a standard IRT model ignores both student heterogeneity in the item parameters and 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 B parameters representing the direct effects of OTL on the item performance, thereby indicating instructional sensitivity in the items. The estimated B's and the corresponding 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 which for its 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 of whom 38% get the item right, while 62% get it right with OTL (this year or prior years). The B 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 given the trait and is therefore the appropriate measure of instructional sensitivity. Several items have large differences in proportion correct for OTL versus no OTL, while having negligible B effects. In order to gauge the importance of the corresponding shifts in the conditional probabilities, Figure 3 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, while rather easy once exposed to it. A harder item may show less instructional sensitivity since 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 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 L value of .43 (an estimated a value of .45).

Consider next item 39. As is seen in Appendix B, this item refers to knowledge about the coordinate system. Here, a rather large group of 30% have no OTL. In terms of proportion correct, the item seems rather easy for the OTL category (0.63) while rather hard for 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. And the item is a poor indicator of the trait (see Table 4) and is in fact the worst one. As shown in the estimated probability curves of Figure 5 the discrimination (the slope) is very small. This would mean that getting the item correct involves little of general math ability, but merely indicates the specific knowledge of the definition, a plausible explanation for this item.

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, followed by a straightforward arithmetic operation, item 16 calls for knowledge about multiplying negative integers and parentheses, and item 3 deals with the simplification and solution of a routine algebra equation. But unlike items 17 and 39 discussed above, items 38 and 16 provide good measurements of the trait.

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 both 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 matter. 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 opportunities to learn 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 would have been eliminated to avoid violation of IRT assumptions. Neither prospect is attractive.

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.

What next steps to take in response to the identification of instructionally sensitive items is unclear. An obvious possibility here 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 data base.

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 which 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 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.

References

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

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.

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

Keifer, E.E., Wolfe, R.G., & Schmidt, W.H. (1987). Understanding patterns of student growth. In L. Burstein (Ed.), *Second International Mathematics Study: Student growth and classroom processes in lower secondary school.* London: Pergamon Press.

Leinhardt, G. (1983). Overlap: Testing whether it's taught. In G.F. Madaus (Ed.), *The courts, validity, and minimum competency testing.* Hingham, MA: Kluwer-Nijhoff Publishing.

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

Linn, R.L. (1983). Testing and instruction: Links and distinctions. *Journal of Educational Measurement, 20*(2), 179-190.

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.

McKnight, C.C., Crosswhite, F.J., Dossey, J.A., Kiefer, 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,* 147-156.

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

Mislevy, R.J., & Bock, R.D. (1984). *BILOG: Marginal estimation of item parameters and subject ability under binary logistic models.* Chicago: International Educational Services.

Muthen, B. (1987a). *Using item-specific instructional information in achievement modeling.* Unpublished manuscript.

Muthen, B. (1987b). Some uses of structural equation modeling in validity studies: Extending IRT to external variables. In H. Wainer and H. Braun (Eds.), *Test validity.* Hillsdale, NJ: Erlbaum.

Muthen, B. (1987c). *LISCOMP. Analysis of linear structural equations with a comprehensive measurement model: User's guide.* Mooresville, IN: Scientific Software, Inc.

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