11

Instructionally Sensitive Psychometrics: Applications to The Second International Mathematics Study

Bengt O. Muthén

CRESST and Graduate School of Education
University of California, Los Angeles
Los Angeles, CA

This chapter discusses new psychometric analyses that improve capabilities for relating performance on achievement test items to instruction received by the examinees. The modeling discussion will be closely tied to the SIMS data for U.S. eighth-grade students.

Item Response Theory (IRT) is a standard psychometric approach for analyzing a set of dichotomously scored test items. Standard IRT modeling assumes that the items measure a unidimensional latent trait, a hypothetical or unobserved characteristic (for example,

mathematical ability). Latent trait models describe the relationship between the observable test performance and the unobservable traits or characteristics that are assumed to influence performance on the test. This particular kind of latent trait model is used to assess the measurement qualities of each item and to give each examinee a latent trait score (the examinee's standing on the latent trait). However, as will be shown, IRT modeling is limited in ways that are a hindrance to properly relating achievement responses to instructional experiences. Taking IRT as a starting point, this chapter summarizes some work on a set of new analytic techniques that give a richer description of achievement-instruction relations.

Six topics that expand standard IRT and specifically deal with effects of varying instructional opportunities will be discussed:

1. *Variation in latent trait measurement characteristics.* This relates to the classic IRT concern of "item bias," here translated as the advantage or disadvantage to OTL in getting an item right.
2. *Multidimensional modeling.* Inclusion of narrowly defined, specific factors closely related to instructional units in the presence of a general, dominant trait.
3. *Modeling with heterogeneity in levels.* Analyses that take into account that achievement data often are not sampled from a single student population but one with heterogeneity of performance levels.
4. *Estimation of trait scores.* Deriving scores based on both performance and background information for both general and specific traits.
5. *Predicting achievement.* Latent trait modeling that relates the trait to student background variables.
6. *Analyzing change.* Relating change in general and specific traits to opportunity to learn.

The SIMS data will be used throughout to illustrate the new methods. All analyses will be carried out within the modeling framework of the LISCOMP computer program (Muthén, 1984, 1987).

The first part of this chapter describes the data to be analyzed. The second part describes general features of the psychometric problem. The third part presents a descriptive analysis of the achievement–instruction relation for the SIMS data and sets the stage for later modeling. The final sections of the chapter discuss methods appropriate for Topics 1–6 listed above.

THE SIMS DATA

We will concentrate our analysis on the U.S. eighth-graders (for whom there are about 4,000 observations from both fall and spring) sampled from about 200 randomly sampled classrooms varying in size from about 5 to 35 students. For the part of the sample that we will be concerned with, the core test was administered both during the fall and the spring to all students in the study while the rotated forms varied in their use pattern. We will be particularly concerned with analyses of the 40 core items but will also report on analyses of the four rotated forms. The rotated form analyses will be presented as a cross-validation of findings for the core items. The SIMS data provide a uniquely rich set of data with which to study instructionally sensitive psychometrics.

It is well known that eighth-grade mathematics curricula vary widely for students in the United States. Part of this information *opportunity to learn* (OTL) for the topics covered by each test item. As noted in previous chapters, for each item on the cognitive test, teachers were asked two questions:

Question 1. During this school year did you teach or review the mathematics needed to answer the item correctly?

1. No
2. Yes
3. No response

Question 2. If in the 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
5. No response

Using these responses, OTL level is defined as:

No OTL: Question 1 = *1*; Question 2 = 2,3,4, or 5
Prior OTL: Question 1 = *1* or *3*; Question 2 = *1*
This Year OTL: Question 1 = *2*; Question 2 = 5 (other response combinations had zero frequencies)

In most analyses to follow, *Prior OTL* and *This Year OTL* will be combined into a single OTL category.

For the U.S. eighth-grade mathematics students, information was also collected in order to make a distinction between "tracks" or class types, yielding a categorization into *remedial*, *typical*, *enriched* (or *pre-algebra*), and *algebra* classes. This classification was based on the SIMS teacher questionnaire data. Other teacher-related information was also collected, as well as student background information on family, career interests, and attitudes. Some of this additional information will also be used in some of the analyses to follow.

THE GENERAL PROBLEM

In general, psychometric modeling assumes independent and identically distributed observations (*i.i.d.*) from some relevant population. This assumption is also made in IRT. Because of the varying curricula and instructional histories of the students in a study like SIMS, the assumption of identically distributed observations is not realistic to describe either relationships between what is measured (achievement responses) and what the measurements are attempting to capture (the traits) or how traits vary with relevant covariates such as instructional exposure and student background.

The distribution of responses for various values of the latent trait cannot be expected to be identical for a student who has had no specific instruction on the item topic and a student who has had instruction. The trait distribution cannot be expected to be the same for students in *enriched* classes as for students in *typical* classes. The students are naturally sampled from heterogeneous populations. Increased homogeneity can be obtained by dividing the students into groups based on instructional experiences; however, such groupings may have to be very detailed to achieve their purpose and any simple grouping may be quite arbitrary.

A more satisfactory approach is to use modeling that allows for heterogeneity, using parameters that vary for varying instructional experiences. Such modeling also accomplishes the goal of instructionally sensitive psychometrics by explicitly describing achievement response-instructional experiences relations.

DESCRIPTIVE ANALYSES

It is informative to consider descriptively how achievement responses within SIMS vary with instructional exposure. This forms a basis for our subsequent modeling efforts. We will first study this in

terms of univariate achievement distributions using the posttest core items. We will also study the change in univariate responses from pretest to posttest.

Univariate Response

Consider first the univariate responses for the posttest. The proportion correct for these items is presented in Table 11.1, broken down by the class-type categories *remedial*, *typical*, *enriched*, and *algebra* and by the OTL categories *No OTL*, *This Year OTL*, and *Prior OTL*. From the totals it is seen that both class type and OTL have a strong effect on proportion correct.

For most items the proportion correct is higher for enriched and algebra classes than for remedial and typical classes. For almost all items the proportion correct increases when moving from No OTL to This Year OTL to Prior OTL. The reason why Prior OTL gives higher proportion correct than This Year OTL is partly because Prior OTL is more common for enriched and algebra classes to which we presume students of higher achievement levels have been selected.

OTL appears to also have an overall positive effect on proportion correct when controlling for class type, at least for typical classes. Also, when controlling for OTL, class type seems to still have a strong effect. These univariate relationships are informative but confound effects of instructional exposure with effects of student achievement level. For example, the higher proportion correct for a certain item for students with Prior OTL may be solely due to such students having a higher achievement level on the whole test. It would be of interest to know if students with the same achievement level perform differently on a certain item for different instructional exposure.

To explore this possibility, we may consider the total score on the posttest as the general mathematics achievement level of each student. Then, for each general achievement level, we could study the variation of proportion correct for each item as a function of instructional exposure. We have carried this out using the dichotomous version of OTL, combining Prior OTL with This Year OTL into a single OTL category.

For each value of the achievement variable we then have a proportion correct for a No OTL and an OTL group and can study whether OTL makes a difference. Conversely, for each of the two OTL categories we will present the distribution of the achievement variable in order to study whether having OTL for an item implies that these students have a higher general achievement level. These plots are given in Figures 11.1-11.4

TABLE 11.1
Percentage of Students and Percentage Correct for Selected Core Items by OTL and Class Type

	Total*		No OTL			This Year OTL			Prior OTL		
Item	*PR*	*PO*	*ST*	*PR*	*PO*	*ST*	*PR*	*PO*	*ST*	*PR*	*PO*
ME01											
TOT	35	43	21	22	26	59	36	47	20	44	48
REM	11	18	33	7	8	60	12	23	7	21	21
TYP	30	38	24	21	27	64	34	43	12	28	43
ENR	42	52	17	25	24	71	48	63	12	29	29
ALG	61	64	6	64	64	5	39	50	89	62	65
AR02											
TOT	47	60	3	34	53	89	45	59	8	74	78
REM	12	21	9	17	33	91	11	20	0	0	0
TYP	42	57	3	34	40	97	42	57	0	0	0
ENR	58	74	4	46	86	90	57	73	6	74	81
ALG	74	75	0	0	0	43	73	71	57	74	78
AL03											
TOT	9	21	38	8	9	61	10	28	1	3	19
REM	15	9	78	15	8	22	13	13	0	0	0
TYP	8	14	49	7	9	50	8	18	2	3	19
ENR	8	21	16	12	11	84	7	23	0	0	0
ALG	16	64	7	0	19	94	17	68	0	0	0
ME06											
TOT	49	55	28	48	54	59	48	55	13	52	59
REM	20	31	41	23	35	45	21	31	14	11	22
TYP	47	52	27	48	53	65	48	52	8	42	47
ENR	52	61	32	51	60	65	52	62	2	82	68
ALG	66	73	10	83	80	28	68	75	62	63	72
ME08											
TOT	89	89	17	89	88	58	88	88	25	93	92
REM	67	61	34	62	55	58	69	64	8	76	67
TYP	89	89	17	94	93	66	88	89	18	89	88
ENR	93	93	16	90	91	59	93	93	26	96	94
ALG	98	97	14	96	100	12	96	98	74	99	97
ME09											
TOT	42	52	14	41	48	56	38	50	30	50	59
REM	16	18	27	18	19	58	15	18	15	21	15
TYP	37	48	14	41	49	62	36	47	23	38	49
ENR	48	64	11	42	53	63	46	65	27	56	65
ALG	67	73	12	76	78	2	56	33	85	66	73
AL16											
TOT	23	58	6	9	16	92	24	60	2	37	88
REM	9	14	52	10	9	48	7	20	0	0	0
TYP	18	50	3	6	11	97	18	52	0	0	0
ENR	28	74	2	17	89	94	28	73	4	34	94
ALG	53	89	0	0	0	94	53	89	6	41	77
GE17											
TOT	47	59	13	39	38	72	46	62	15	59	63
REM	24	24	41	22	15	48	25	26	10	29	46

TABLE 11.1 (cont.)
Percentage of Students and Percentage Correct for Selected Core Items by OTL and Class Type

	Total*		No OTL			This Year OTL			Prior OTL		
Item	*PR*	*PO*	*ST*	*PR*	*PO*	*ST*	*PR*	*PO*	*ST*	*PR*	*PO*
TYP	42	56	11	42	37	82	43	60	8	35	40
ENR	53	68	12	44	44	80	55	72	8	53	68
ALG	76	80	10	61	85	18	78	93	72	78	77
GE19											
TOT	23	33	76	23	32	23	22	38	1	52	57
REM	10	19	0	10	19	0	0	0	0	0	0
TYP	22	30	72	22	29	28	21	33	0	0	0
ENR	25	39	71	25	35	29	25	49	0	0	0
ALG	39	49	89	38	48	0	0	0	11	52	57
GE21											
TOT	20	34	60	20	30	37	21	39	3	23	39
REM	16	16	97	16	17	3	25	13	0	0	0
TYP	18	30	60	17	29	39	20	33	1	22	11
ENR	20	39	46	20	34	52	20	44	2	6	33
ALG	34	50	65	33	45	18	44	71	17	28	49
GE22											
TOT	37	59	13	26	26	80	37	64	7	62	67
REM	21	18	79	23	19	17	9	11	4	30	40
TYP	33	55	8	28	26	90	33	58	2	29	37
ENR	40	71	6	20	15	92	40	75	2	59	59
ALG	70	81	9	47	82	44	70	85	47	73	78
AL25											
TOT	42	46	7	28	34	92	42	47	2	70	59
REM	12	15	28	8	13	72	13	16	0	0	0
TYP	38	42	7	36	40	92	37	43	2	68	44
ENR	48	55	3	40	60	97	49	55	0	0	0
ALG	69	67	0	0	0	94	69	66	6	73	86
AR34											
TOT	24	39	4	16	19	90	22	39	7	45	53
REM	10	15	19	14	16	81	9	14	0	0	0
TYP	19	34	4	17	22	96	19	34	0	0	0
ENR	29	54	0	0	0	97	29	54	3	39	35
ALG	44	53	0	0	0	43	43	50	57	45	55

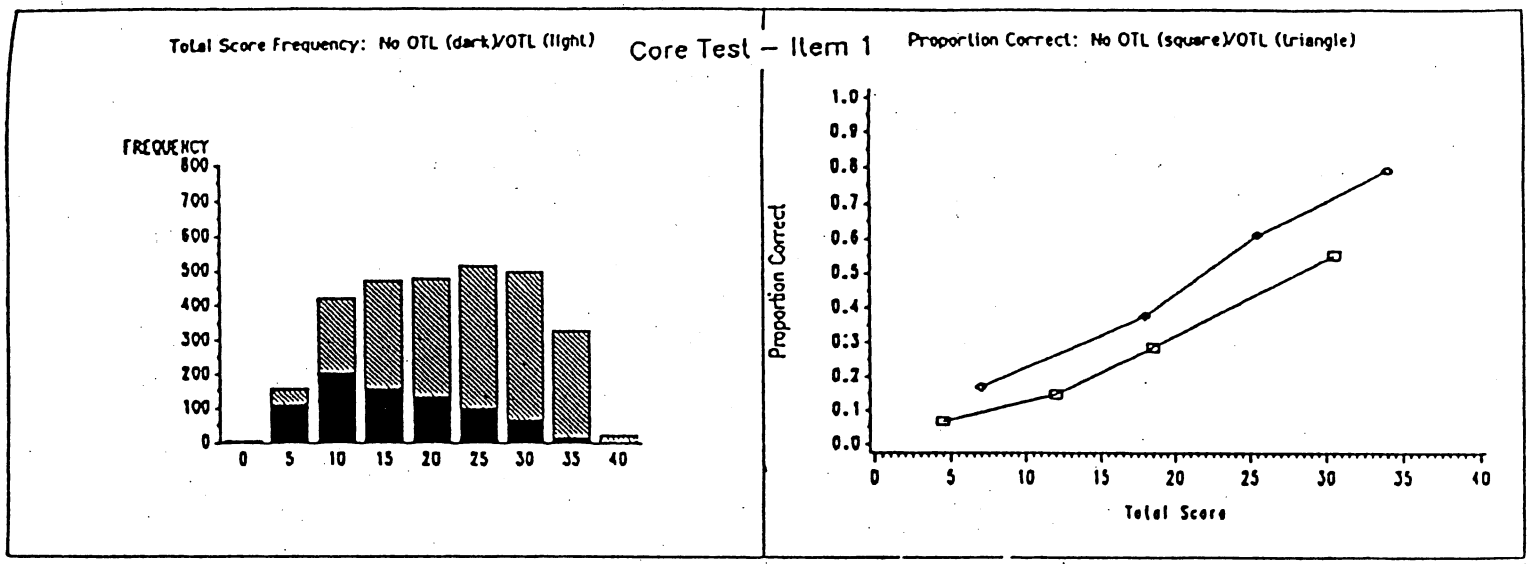
*Percentages of students by class type are:
REM = Remedial: 7.1 (*N* = 268), *TYP* = Typical: 57.6 (*N* = 2148)
ENR = Enriched: 24.4 (*N* = 909), *ALG* = Algebra: 10.7 (*N* = 399)

ST = Percentage students
PR = Percentage correct for pretest
PO = Percentage correct for posttest

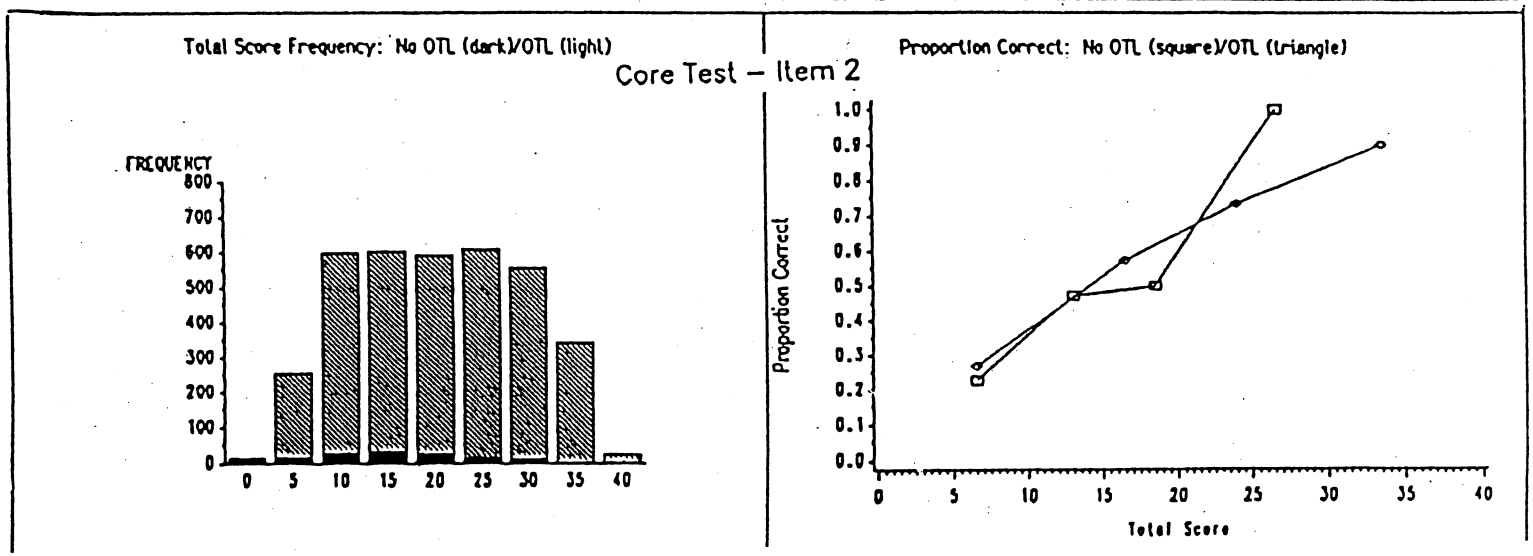
ME = Measurement
AR = Arithmetic
AL = Algebra
GE = Geometry

Figure 11.1 describes Items 1, 2, and 3. The left-most panel shows the total score distribution given No OTL and OTL, respectively. We note that the score distributions have different means with the OTL distribution having a somewhat higher mean, supporting the notion that students who receive OTL perform better as measured by

Core Test – Item 1



Core Test – Item 2



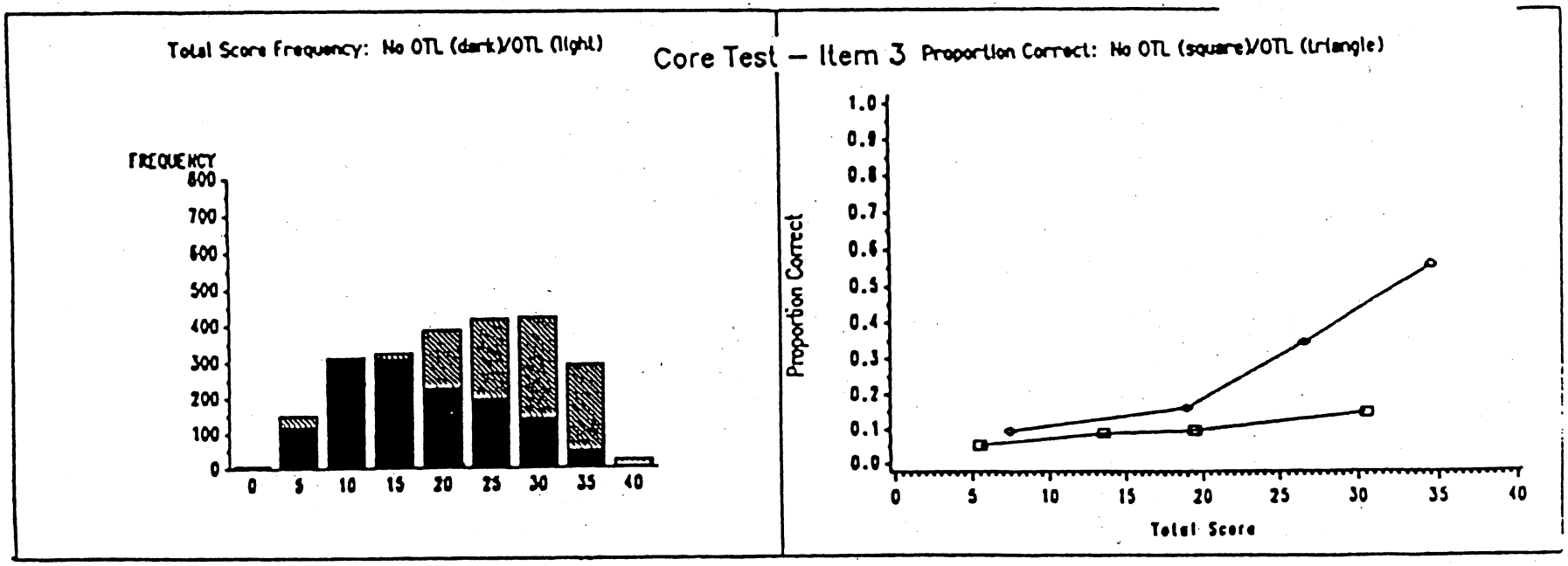


FIGURE 11.1 Score Distributions for Core Items 1, 2, and 3

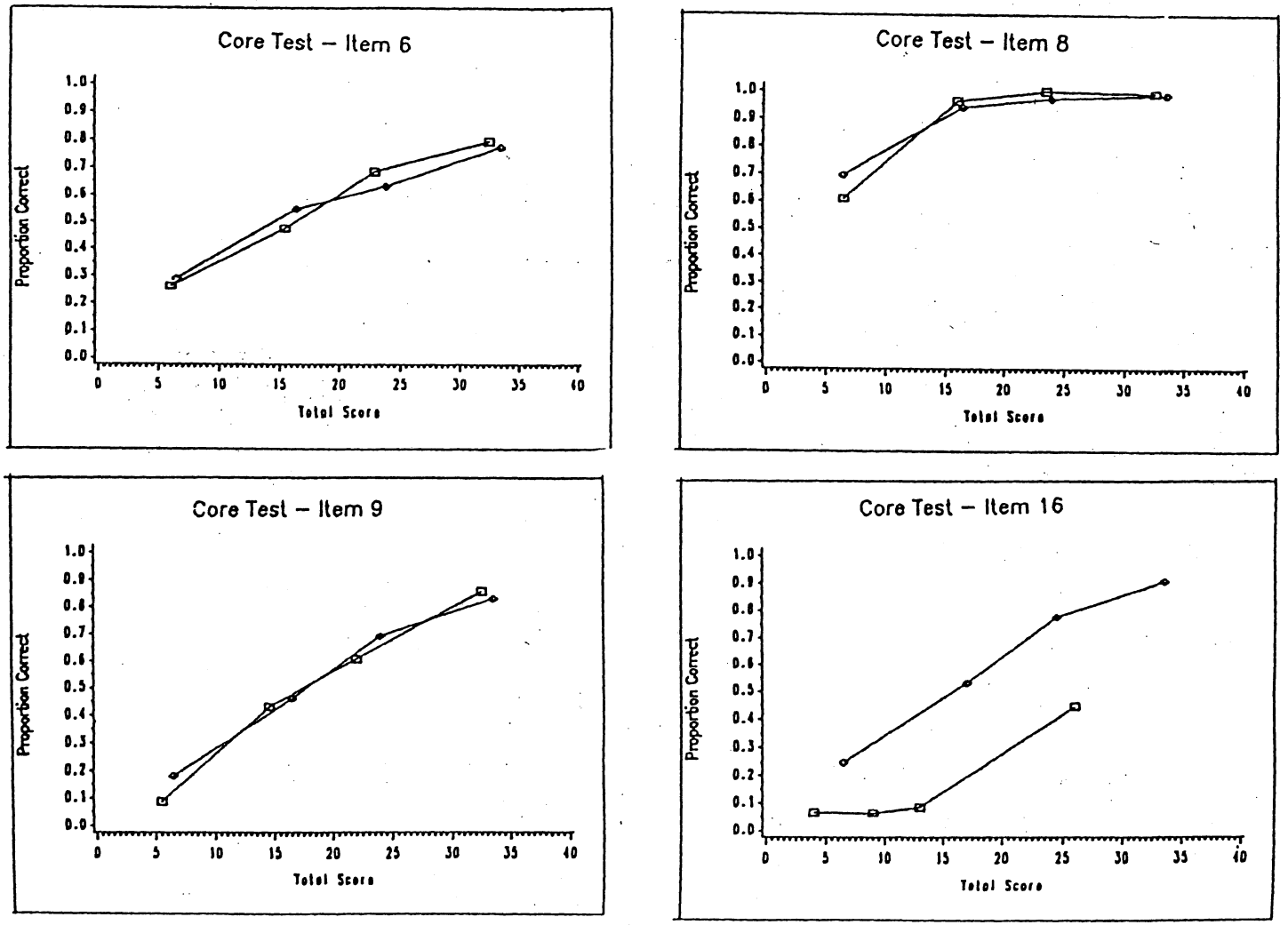


FIGURE 11.2 Score Distributions for Core Items 6, 8, 9, and 16

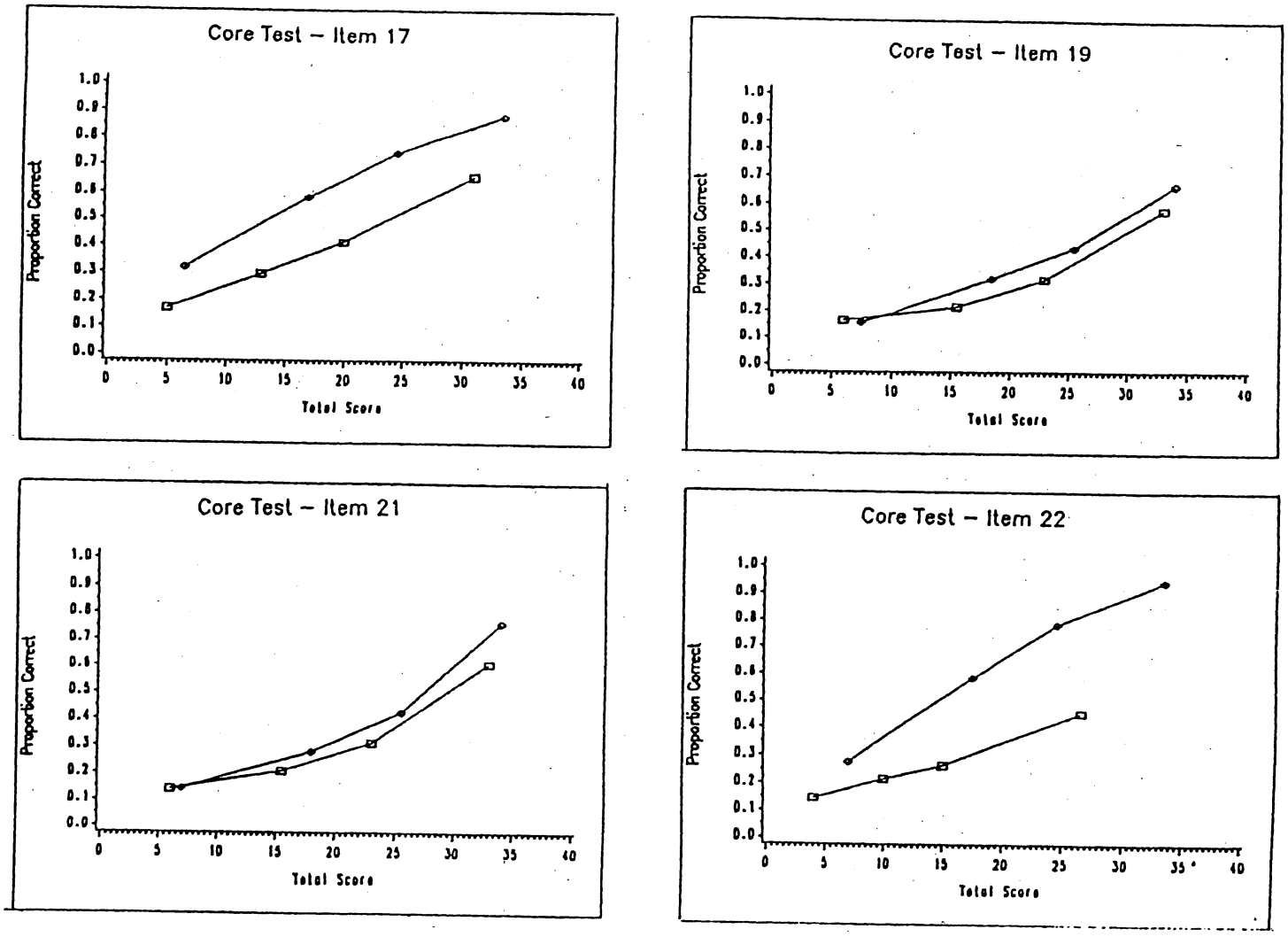
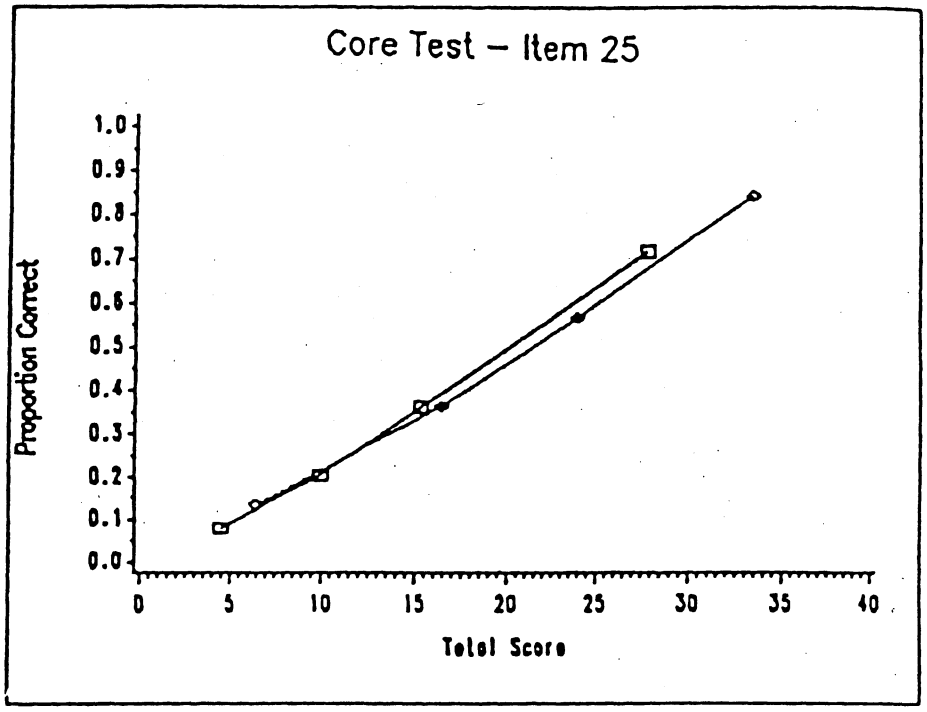


FIGURE 11.3 Score Distributions for Core Items 17, 19, 21, and 22



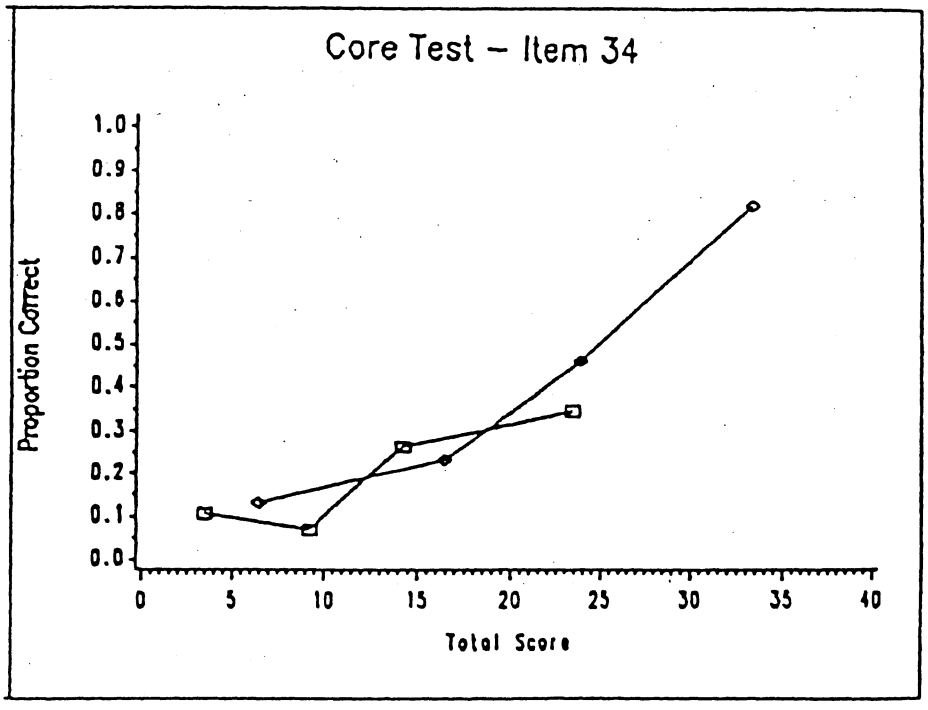


FIGURE 11.4 Score Distributions for Core Items 25 and 34

this test. We also note that the variances of the two distributions are about the same. The score distributions shown are representative of all core items.

The right-most part of Figure 11.1 and Figures 11.2-11.4 contain curves showing the proportion correct for a given total score for the two OTL categories. For each item and both OTL categories, proportion correct increases with total score indicating that for both OTL categories the item is a good indicator of the general achievement variable that the total score represents. It is particularly noteworthy that this is true also for the No OTL category and that the No OTL and OTL curves most often are very close. The students who, according to their teachers, have not been taught the mathematics needed to answer the item correctly still appear to have a high probability of answering the item correctly and this probability increases with increasing total score. This may indicate that students can, to a large degree, draw on related knowledge to solve the item. It may also indicate unreliability in the teachers' OTL responses. However, the differences in score distributions for the core items show that the OTL measures have consistent and strong relations to the total score. Instead of unreliability there may be a component of invalidity involved in the teachers' responses, where OTL may to some extent be confounded with average achievement level in the class and/or the item's difficulty.

The score distributions show that OTL is correlated with performance. Our hypothesis is that OTL helps to induce an increased level of general achievement and that, in general, it is this increased level that increases the probability of a correct answer, not OTL directly. In this way, moving from the No OTL status to the OTL status implies a move upwards to the right along the common curve for No OTL and OTL.

There are some exceptions to the general finding of common curves for the No OTL and OTL categories. For example, Items 3 and 17 show a large positive effect of having OTL. Several other items with sizeable numbers of students in the two OTL categories also show positive effects. This means that for these items, the added advantage of having OTL is not fully explained by a corresponding increase in total score. OTL directly affects success in answering the item correctly. From Table 11.1 we find that for the three items listed, the proportion correct increases strongly when moving from the No OTL category to the OTL categories.

However, Table 11.1 cannot be counted on for finding items with direct OTL effects of this kind since several other items also show

strong increases in proportion correct due to OTL. For example, in Table 11.1, Item 25 shows substantial increases in proportion correct moving from the No OTL to the OTL categories, but the curves shown in Figure 11.4 are essentially the same. We will return to the interpretation of this type of effect in Section 4. Note also that with the exception of Item 3 any OTL effect appears to be such that the two curves are approximately parallel, implying that the OTL effect is constant across achievement levels. For Item 3 the OTL advantage increases with increasing achievement level, perhaps because it is a difficult item.

Change of Univariate Responses

The SIMS core items also provide the opportunity to study changes in proportion correct for each item from the fall to the spring testing. This change can be related to OTL. For each item we may distinguish between three groups of students: those who did not have OTL before the pretest or before the posttest (the No OTL group); those who had OTL before the pretest (Prior OTL); and those who did not have OTL before the pretest but did have OTL before the posttest (This Year OTL).

The change for the No OTL group gives an indication of change due to learning on related topics. The change for the Prior OTL group gives an indication of effects related to practice, review, and, perhaps, forgetting. The change for the group having This Year OTL reflects the direct exposure to the topic represented by the item. These changes can also be studied in Table 11.1. Table 11.1 shows that where changes occur, they are largely positive for each OTL category, with the largest changes occurring for students in the category of This Year OTL as expected. They may be taken to support the dependability of the teacher-reported OTL measure.

VARIATIONS IN LATENT TRAIT MEASUREMENT CHARACTERISTICS

The study of the univariate achievement responses above showed that the set of core test items served as good indicators of the total test score. We may hypothesize that this test score is a proxy for a general mathematics achievement variable as measured by the combined content of the set of core items. However, the total test score is a fallible measure and what we are interested in are the relationships between the items and the true score and estimates of the true scores. This is a situation for which Item Response Theory (IRT) has been proposed as being appropriate (see, for example, Lord, 1980).

The curves of Figures 11.1 to 11.4 are, in IRT language, empirical item characteristic curves, which as theoretical counterparts have conditional probability curves describing the probability correct on an item given a latent trait score. We will now describe the IRT model and how it can be extended to take into account instructional heterogeneity in its measurement characteristic.[1]

In this part of the chapter we investigate descriptively whether the proportion correct for a given total test score varied across OTL groups. In IRT language this is referred to as investigating *item bias* or, using a more neutral term, *differential item functioning*. Standard IRT assumes that the item functions in the same way for different groups of individuals. Concerns about item bias due to instructional heterogeneity have recently been raised in the educational measurement literature. (See Chapter 8.) A variety of bias detection schemes related to IRT have been discussed in the literature. Conflicting results have been found in empirical studies. For example, Mehrens and Phillips (1986, 1987) found little differences in the measurement characteristics of standardized tests due to varying curricula in schools; however, Miller and Linn (1988), using the SIMS data, found large differences related to opportunity to learn, although these differences were not always interpretable.

---

[1] In formulas the IRT model may be briefly described as follows: Let y^* be a p vector of continuous latent response variables that correspond to specific skills needed to solve each item correctly. For item j,

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

where 0 denotes the incorrect answer, 1 denotes the correct answer, and τ_j is a threshold parameter for item j corresponding to its difficulty. Assume also that the latent response variable y^*_j is a function of a single continuous latent trait η and a residual ϵ_j,

$$y^*_j = \lambda_j \eta + \epsilon_j \qquad (2)$$

where λ_j is a slope parameter for Item j, interpretable as a factor loading. With proper assumptions on the right-hand-side variables, this gives rise to the two-parameter normal ogive IRT model. For each item there are two parameters, τ_j and λ_j. The conditional probability of a correct response on Item j is

$$P(y_j = 1|\eta) = \Phi\,[(-\tau_j + \lambda_j \eta)\,\phi^{-1/2}] \qquad (3)$$

where ϕ is the variance of ϵ_j. This means that the threshold τ_j determines the item's difficulty, that is the horizontal location of the probability curve, and the loading λ_j determines the slope of the probability curve.

Muthén (1989b) pointed out methodological problems in assessing differential item functioning when many items may be biased. He suggested a new approach based on a model which extends the standard IRT. The analysis is carried out by the LISCOMP program (Muthén 1987). This approach is particularly suitable to the SIMS data situation with its item specific OTL information.[2]

The model disentangles the effects of OTL in an interesting way. It states that OTL has a direct effect on the general achievement trait. Here we are interested in finding positive effects of instruction. Through the expected increase in the general achievement trait, such effects also have an indirect positive effect on the probability of a correct item response.

In addition to the indirect effect of OTL for an item, there is also the possibility of a direct OTL effect on an item. Any direct effect indicates that the specific skill needed to solve the item draws not only on the general achievement trait but also on OTL. The size of the OTL effect on the general achievement trait indicates the extent to which the trait is sensitive to instruction. The size of the OTL

---

[2]Let x be a vector of p OTL variables, one for each achievement item. The x variables may be continuous, but assume for simplicity that x_j is dichotomous with $x_j = 0$ for *No OTL* and $x_j = 1$ for *OTL*. Consider the modification of Equation (2)

$$y^* = \lambda\eta + B\,x + \epsilon \qquad (4)$$

where in general we restrict B to be a diagonal $p \times p$ matrix. The diagonal element for Item j is denoted β_j. The OTL variables are also seen as influencing the trait η,

$$\eta = \gamma\,x + \zeta \qquad (5)$$

where γ is a p-vector of regression parameter slopes and ζ is a residual. It follows that

$$P(y_j = 1|\eta, x_j) = \Phi\,[(-\tau_j + \beta_j x_j + \lambda_j\eta)V(y^*_j|\eta)^{-1/2} \qquad (6)$$

In effect, then, the β_j coefficient indicates the added or reduced difficulty in the item due to OTL. Equivalently, using equation (4), we may see this effect as increasing y^*_j, the specific skill needed to solve Item j.

We note that this model allows for differential item functioning in terms of difficulty but not in terms of the slope related parameter λ_j. This is in line with the data analysis findings of Section 3.1 where little difference in slopes of the conditional proportion correct curves was found across OTL groups (Item 3 was an exception; we assume that this item will be reasonably well fitted by a varying difficulty model). More general modeling is in principle possible, but the data features do not seem to warrant such an extra effort.

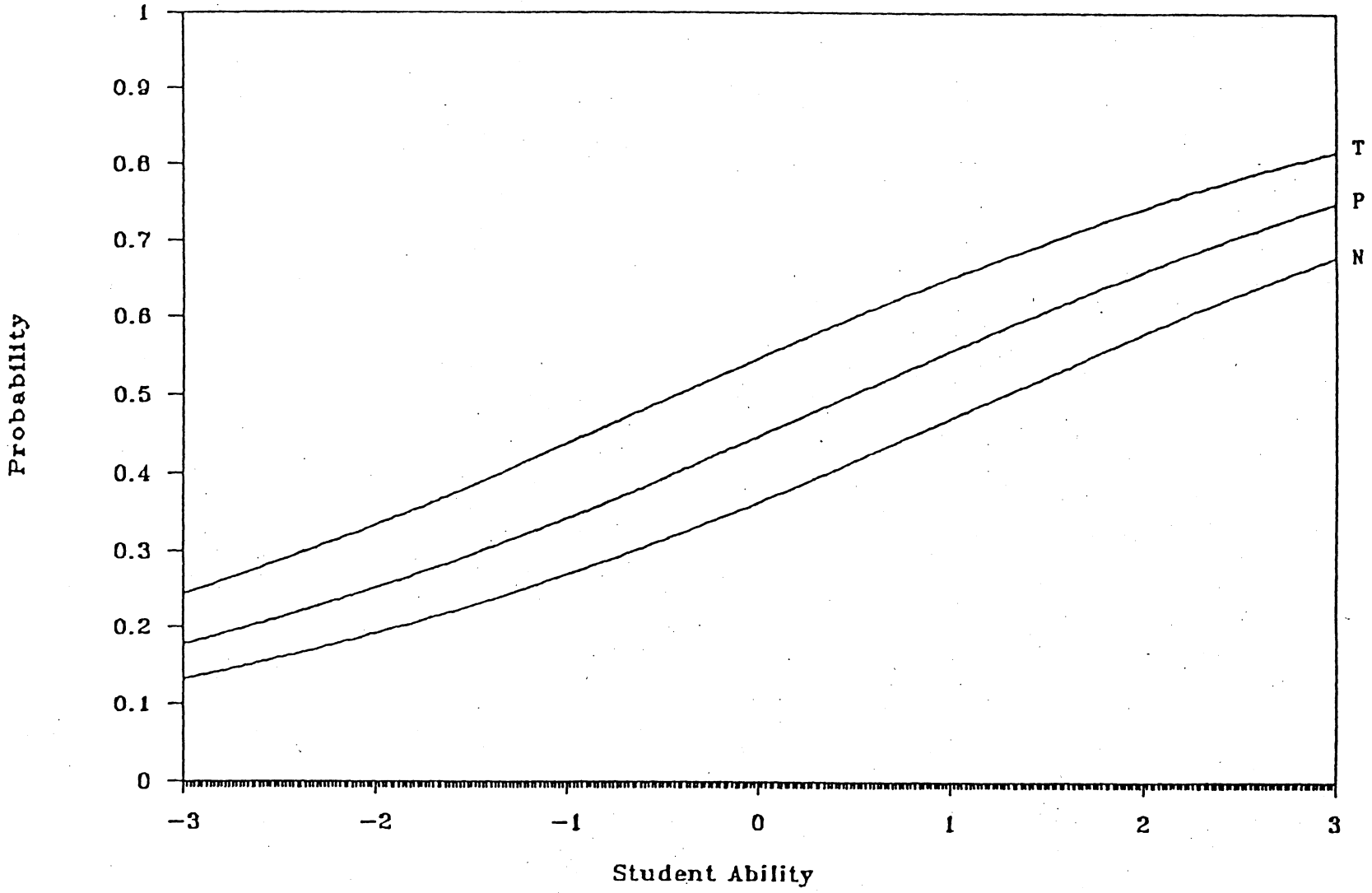
effect on the probability of a correct response indicates the amount of exposure sensitivity, or instructional oversensitivity, in the item.

While positive effects on the general achievement trait correspond to a positive educational outcome, possible direct effects on items are of less educational interest in that they demonstrate effects of teaching that influence very narrow content domains. From a test construction point of view, items that show such exposure sensitivity are less suitable for inclusion in standardized tests since they are prone to "item bias" in groups of examinees with varying instructional history. If such item bias goes undetected, IRT analysis is distorted. However, in the modeling presented here exposure sensitivity is allowed for and the analysis does not suffer from the presence of such effects.

Muthén, Kao, and Burstein (1991) present examples of analysis of exposure sensitivity using the dichotomous OTL groupings. However, we will first consider an example where the OTL categories *No OTL*, *This Year OTL*, and *Prior OTL* were used. Figure 11.5 shows the estimated item characteristic curves for Item 17, which has to do with acute angles. Since there are three OTL categories, there are three curves corresponding to three difficulty values. Since the curves for both This Year OTL and Prior OTL are above the No OTL curve, the direct effects of OTL on the probability of a correct response are positive for these two OTL groups. Exposure to the concept of acute angles produces a specific skill, which has the same effect as a reduced item difficulty, and this skill is not included in the general achievement trait.

It is interesting to relate this finding to the percentage correct on Item 17 broken down by OTL group as given in Table 11.1. Percentage correct increases dramatically from the No OTL category to the OTL categories, but the percentage correct is slightly higher for Prior OTL than for This Year OTL. For Item 17 the Prior OTL students may do better than This Year OTL students, but Figure 11.5 shows that the recency of OTL gives an advantage for students at the same achievement trait level. Comparing the estimated item characteristic curves of Figure 11.5 with the empirical curves of Figure 11.3 we find a large degree of similarity but also differences. The estimated curves represent more correct and precise estimates of these curves.

Muthén et al. (1991) found substantial exposure sensitivity in Items 3 (solving for x), 16 (product of negative integers), and 17 (acute angles), 38 (percentages), and 39 (coordinate system). While Items 3, 17, and 39 provided rather poor measurements of the achievement trait as indicated by their estimated factor loadings,


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Probability
−3
−2
−1
0
1
2
3
Student Ability
T
P
N

that was not the case for the other two. We hypothesized that the exposure sensitivity corresponded to early learning of a definitional nature.

Further analyses of the rotated form items, carried out by Kao (1990), supported this hypothesis. For example, the rotated forms showed exposure sensitivity for items covering square root problems. Overall, about 15 to 30 percent of the items exhibit mild exposure sensitivity, while only about 10 to 15 percent exhibit strong exposure sensitivity. We may note that these percentages are considerably lower than Miller and Linn's (1988) findings using related parts of the SIMS data and standard IRT methodology. The effects of OTL on the achievement trait will be discussed in later sections.

MULTIDIMENSIONAL MODELING

Standard IRT modeling assumes a unidimensional trait. For a carefully selected set of test items, this is often a good approximation. However, in many achievement applications, it is reasonable to assume that sets of items draw on more than one achievement trait.

Although of great substantive interest, models with many minor factors are very hard to identify by the means of analysis that are commonly used. For instance, assume, as we will for the SIMS data, that a general achievement factor is the dominant factor in that it influences the responses to all items. Assume further that, in addition to this general factor, there are several specific factors, uncorrelated with the general factor, that influence small sets of items with a common, narrow content. It is well known that such models with continuous data cannot be easily recovered by ordinary exploratory factor analysis techniques involving rotations.[3] This problem carries over directly to dimensionality analysis of dichotomous items using tetrachoric correlations.

---

[3]Muthén (1978) presented a method for the factor analysis of dichotomous items, where the model is

$$y^* = \Lambda\eta + \epsilon \quad (7)$$

$$V(y^*) = \Lambda\Psi\Lambda' + \Theta \quad (8)$$

where Λ is a $p \times m$ factor loading matrix, Ψ is a factor covariance matrix, and Θ is a diagonal matrix of residual variances. In line with item analysis tradition (see Lord and Novick, 1968), Muthén fitted the model to a matrix of sample tetrachorics. For an overview of factor analysis with dichotomous items, see Mislevy (1986).

Consider as an illustration of the problem an artificial model for 40 dichotomous items. Assume that one general factor influences all items and eight specific factors each influence a set of five items. Let the general factor loadings be 0.5 and 0.6 while the specific factor loadings are 0.3 and 0.4. Let the factors be standardized to unit variances and let the factors be uncorrelated. The eigenvalues of the corresponding artificial correlation matrix are shown in Figure 11.6.

Such a "scree plot" is used for determining the number of factors in an item set. The number of factors is taken to correspond to the first break point in the plot where the eigenvalues level off. If the first eigenvalue is considerably larger than the others and the others are approximately equal, this is usually taken as a strong indication of unidimensionality. Figure 11.6 clearly indicates unidimensionality despite the existence of the eight specific factors. There would be no reason to consider solutions of higher dimensionality. As a comparison, Figure 11.7 shows the eigenvalues for the tetrachoric correlation matrix for the 39 core items of the SIMS data. The two eigenvalue plots are rather similar.

Models similar to the artificial one considered above have been studied by Schmid and Leiman (1957). They pointed out that the situation with one general factor and k specific factors uncorrelated with the general factor could also be represented as a k-factor model with correlated factors. Hence, they used the term *hierarchical factor analysis*.

The usefulness of hierarchical factor analysis has recently been pointed out by Gustafsson (1988a, 1988b). He sought to circumvent the difficulties of exploratory factor analysis by formulating confirmatory factor analysis models. Hypothesizing a certain specific factor structure in addition to a general factor, the confirmatory model enables the estimation of factors with very narrow content. (Applications of this type of modeling to the SIMS data are being considered by the author in collaboration with Burstein, Gustafsson, Webb, Kim, Novak, and Short.)

In line with our previous modeling, we may consider a simplified version of the confirmatory model.[4] In this simplified version of the

[4]We may write a simple version of this model as

$$y^*_j = \lambda_{G_j}\,\eta_G + \lambda_{S_j}\,\eta_{S_k} + \epsilon_j \qquad (9)$$

where y^* is the latent response variable for Item j (cf. the Section 3 model), η_G is the general achievement factor, η_{S_k} is the specific factor for Item j, and ϵ_j is a residual. The three right-hand side variables are taken to be uncorrelated. This means that the items belonging to a certain specific factor correlate not only due to the general factor but also due to this specific factor.

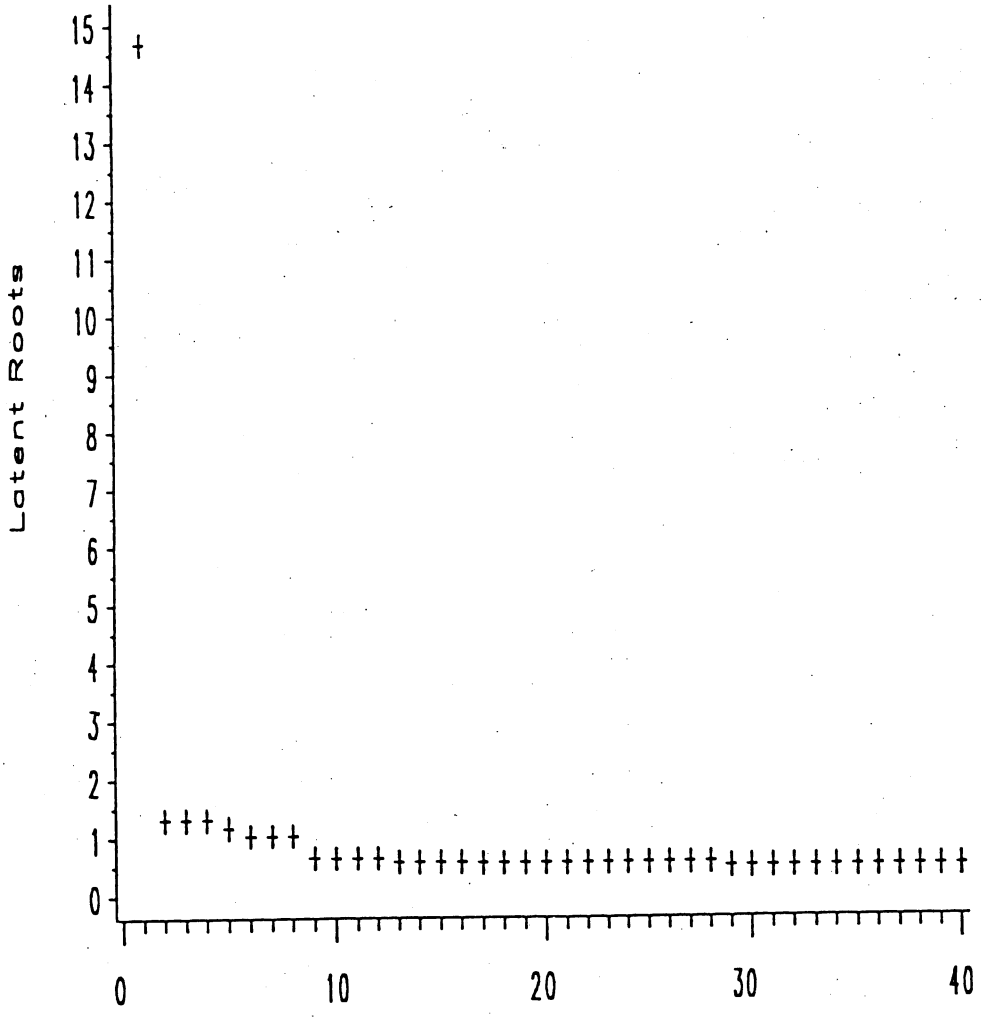


FIGURE 11.6 Scree Plot for Tetrachoric Correlations with Artificial Model for 40 Items

model, it is assumed that each item measures only one specific factor. For identification purposes we assume that each specific factor is measured by at least two items. The general factor is assumed to influence each item to a different degree, while the specific factor has the same influence on all items in the corresponding set.

The multidimensional confirmatory factor analysis model allows an interesting variance component model interpretation. The model implies a decomposition of the latent response variable variances into a general factor component, a specific factor component, and

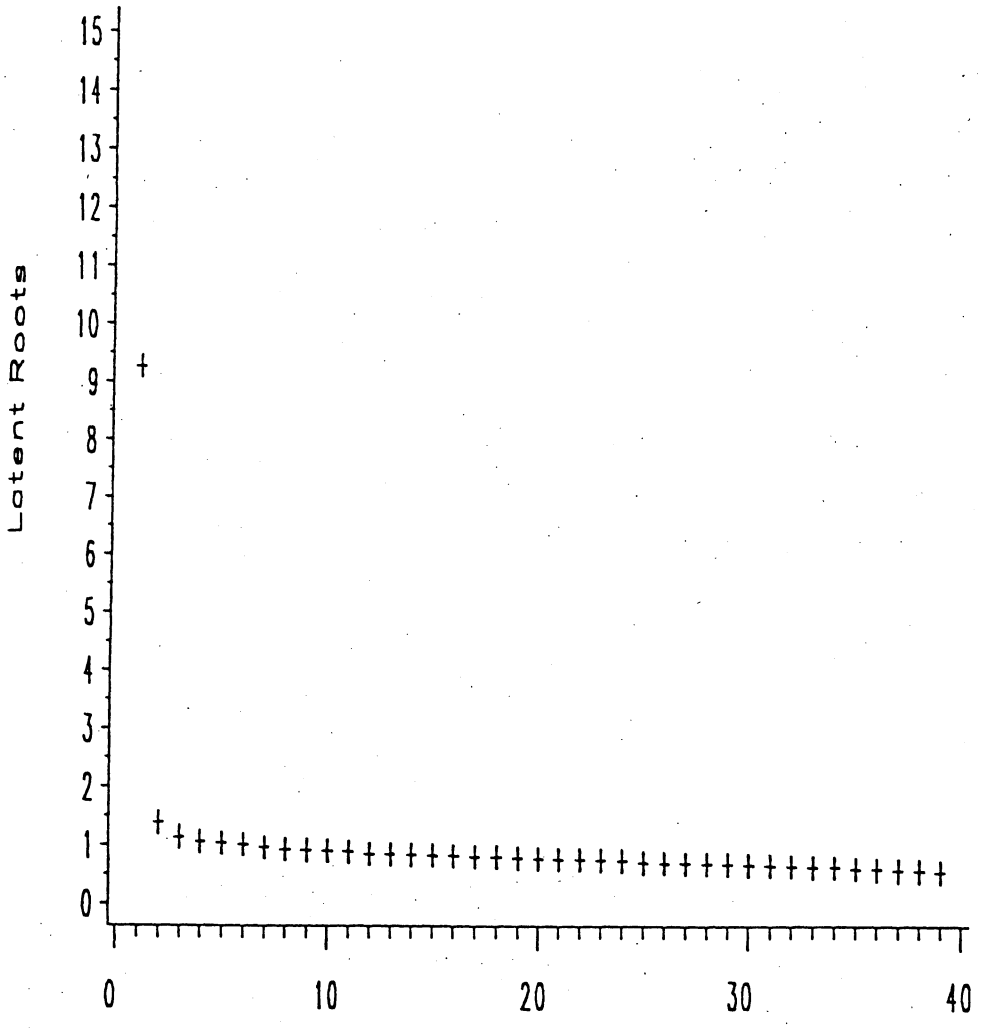


FIGURE 11.7 Scree Plot of Latent Roots for 39 Items Based on Tetrachorics

an error component.[5] The relative sizes of the general and the specific components are of particular interest. The specific component can also be interpreted as the average correlation remaining between items belonging to specific factor k when holding the general factor constant. The model can be estimated by confirmatory factor

[5]The variance component model is estimated by standardizing the general factor variance to unity, while letting the specific factor variances be free parameters. The decomposition is:

$$V(y^*_j) = \lambda_{G_j}^2 + \psi_{S_k} + \theta_j \qquad (10)$$

where ψ_{Sk} is the variance of the specific factor k. Since the items are dichotomous, the variances of the y^*'s are standardized to one by restrictions on the θ_j's.

analysis techniques for dichotomous items using the LISCOMP computer program (see Muthén, 1978, 1987).

The SIMS items of the core and the rotated forms were classified into subsets corresponding to specific factors defined both by content and procedure. Examples of the narrow item domains that were considered are: arithmetic with signed numbers (Core Items 3, 16, 25), percent calculations (Core Items 2, 34, 36, 38), estimation skills (size, distance; Core Items 6, 8, 9), and angular measurements (Core Items 17, 19, 21, 22).

The analysis steps are as follows: For a given hypothesized set of specific factors, a confirmatory factor analysis can be performed. The initial model may then be refined in several steps. An inappropriate combination of items for a specific factor gives rise to a low or negative variance component estimate for this specific factor. Modifications may be assisted by inspection of model misfit indices.

For this model a useful index is related to the loadings of the specific factors that are fixed to unity in the baseline model. The sign and size of the derivatives of these loadings are of interest. A positive value for a certain item indicates that if the loading is free to be estimated, the estimated value will be smaller than one. In effect, this allows the estimate of the variance component for the specific factor at hand to increase. This is because the specific variance component is related to the average correlation of the specific factor items, conditional on the general factor, where the decrease in the factor loading for a certain item means that the contribution from this item is weighted down. Thus, modifying the initial analysis, items that obtain very low or negative specific factor loadings are candidates for exclusion from the set assigned to this specific factor. This modification process may be performed in several iterations. In the analyses performed for the SIMS data, this procedure appeared to produce substantively meaningful results in that the items that were signaled out clearly had features that distinguished them from the others in the set.

Table 11.2 gives the estimated variance components for core items corresponding to three of the specific factors. It is seen that the variance contribution from the specific factors can be as large as 50 percent of that of the general factor and are therefore of great practical significance. This is particularly so since the sets of items for a specific factor correspond closely to instructional units. Analyses of the rotated forms replicated most of the specific factors found for the core.

The confirmatory factor analysis procedure described is a cumbersome one, involving many iterations and many subjective decisions. An attempt was therefore made to find an approach that would involve fewer steps and a more objective analysis. It was

TABLE 11.2
Variance Components for Selected Items from the SIMS Population A Test Core*

Item	General Factor	Specific Factors: Percent	Specific Factors: Estimate	Specific Factors: Angular Measurement
AR02	33 (24)	9 (9)		
AR24	39 (32)	9 (9)		
AR36	32 (27)	9 (9)		
AR38	35 (26)	9 (9)		
ME06	20 (14)		9 (10)	
ME08	38 (27)		9 (10)	
ME09	38 (29)		9 (10)	
GE17	28 (17)			11 (12)
GE19	17 (12)			11 (12)
GE21	24 (17)			11 (12)
GE22	43 (30)			11 (12)

*The estimate when controlling for mean level heterogeneity is given in parentheses (See Section 5).

reasoned that if the influence of the general factor could be removed from the item correlations, the remaining correlations would be due to the specific factors alone. Such residual correlations could then be factor analyzed by regular exploratory techniques, at least if nesting of specific factors within each other was ignored. Given a proxy for the general factor, the residual correlations could be obtained by bivariate probit regressions of all pairs of items on the proxy using the LISCOMP program.

An attempt was first made to approximate the general factor for the posttest core items with the posttest total score. However, this produced almost zero residual correlations. Instead, the pretest total score was used for the posttest items. An exploratory factor analysis of these residual correlations, using an orthogonal rotation by Varimax, resulted in 11 factors with eigenvalues greater than one.

The interpretation of these factors showed an extraordinary high degree of agreement with the specific factors previously obtained. The best agreement was obtained for factors that had obtained the largest variance component estimates. The exploratory analysis also suggested a few items to be added to the specific factors as defined earlier. The agreement of these two very different approaches is remarkable and it is interesting that the pretest score appears to be a better proxy for the general factor at the posttest occasion than the posttest score. This may indicate that the general factor is a rela-

tively stable trait related to the achievement level before eighth grade instruction (we note from Table 11.1 that This Year OTL is the most prevalent category). In contrast, controlling for posttest score may control for a combination of the general factor and specific factors.

It is interesting to note that analyses of the core items administered at the pretest gave very similar results in terms of specific factors identified by the confirmatory approach. This indicates stability of the specific factors over the eighth grade. Attempting to compute residual correlations for exploratory factor analysis again gave near zero values when controlling for the total score, the pretest in this case, and this approach had to be abandoned.

MODELING WITH HETEROGENEITY IN LEVELS

The factor analysis described in the previous section was performed under the regular assumption of identically distributed observations. That is, all students are assumed to be sampled from the same population with one set of parameters. However, we have already noted that the students have widely varying instructional histories and that the homogeneity of student populations is not a realistic assumption. This is a common problem in analysis of educational data and has been given little attention. We may ask how this heterogeneity affects our analysis and if it can be taken into account in our modeling.

Muthén (1989a) considers covariance structure modeling in populations with heterogeneous mean levels. He considers the effect of incorrectly ignoring the heterogeneity, and proposes a method to build the heterogeneity into the model. The method is directly applicable to the multidimensional factor analysis model considered in the previous section and can also be carried out within the LISCOMP framework.[6] This modeling has two important outcomes. The

---

[6]Consider the model of Equation (7)

$$y^* = \Lambda\eta + \epsilon \qquad (11)$$

In the previous section we made the usual standardization of $E(\eta_i) = 0$ for all observations i and assumed $V(\eta_i) = \Psi$. However, we know that it is unrealistic to assume that, for example, students from different class types have the same factor means levels. We may instead want to assume that the means vary with class type such that for Student i in Class c we have $E(\eta_{i_c}) = \alpha_c$. As pointed out in Muthén (1989a) this may be accomplished by considering in addition to (11) the equation

dimensionality analysis can be carried out without distortion due to the differences in factor mean levels across class types, and the factor mean levels can be estimated.

A mean-adjusted analysis was carried out on the SIMS core items using the multidimensional factor model from Table 11.2. Factor mean differences were allowed for class type using three dummy variables and also gender. We will concentrate our discussion of the results on the factor structure.

Despite large mean differences across class type for the general achievement factor, a factor structure very similar to the previous one emerged. The same specific factors showed large and small variances, respectively. Hence, the potential for a distorted structure is not realized in these data. The results are presented in parentheses in Table 11.2. It is seen that the variance contributions to the general factor are considerably reduced as compared to the first approach.

$$\eta_{i_c} = \Gamma x_c + \zeta_{i_c} \tag{12}$$

where x_c represents a vector of class type dummy variable values for Class c, Γ is a parameter matrix, and ζ_{ic} is a residual vector for Student i in Class c. We assume that conditional on class-type membership the factor means vary while the factor covariance matrix remains constant.

$$E(\eta_{i_c}|x_c) = \Gamma x_c \tag{13}$$

$$V(\eta_{i_c}|x_c) = \Psi \tag{14}$$

The modeling also assumes that the matrices Λ and Θ are constant across class types, so that

$$E(y^*|x_c) = \Lambda\Gamma x_c \tag{15}$$

$$V(y^*|x_c) = \Lambda\Psi\Lambda' + \Theta \tag{16}$$

It is interesting to note that the assumption of constancy of the conditional covariance matrix $V(y^*/x_c)$ is in line with the findings of constancy of the homogeneity of correlations.

The structure imposed on the parameter matrices of (15) and (16) may correspond to an exploratory or a confirmatory factor analysis model. Muthén (1989a) points out that the conditional covariance matrix of (16) is not in general the same as the marginal covariance matrix $V(y^*)$. In our context this means that even when we have the same factor analysis structure in the different class types this covariance structure does not hold in the total group of students. The approach outlined here essentially provides a mean-adjusted analysis of pooled covariance matrices assumed to be equal in the population. In our situation the analysis effectively is carried out on pooled tetrachoric correlation matrices.

The reduction in variance contribution from the general factor is natural, since holding class type constant reduces the individual difference in the general achievement trait due to selection of students. If the inference is to the mix of students encountered in the SIMS data, the unreduced variation in the trait is the correct one; but this variation is not representative for a student from any given class type. It is also interesting to note that the specific factor variances are not similarly reduced by holding class type constant, presumably indicating that these specific skills are largely unrelated to the student differences represented by class type.

ESTIMATION OF TRAIT SCORES

So far in this chapter, we have considered various factor analysis models for the achievement responses. Assuming known or well-estimated parameter values for these models, it is of interest to estimate each student's score on the factors of these models. For the standard, unidimensional IRT model, estimation of the trait values is a standard task which may be carried out by maximum likelihood, Bayes' model (maximum a posteriori), or expected a posteriori estimators (see for example, Bock & Mislevy, 1986).

However, the instructionally sensitive models we have considered for the SIMS data have brought us outside this standard situation in the following three respects:

1. In line with Section 4, we want to consider factor score estimation that takes into account that certain items have different difficulty levels depending on the students' OTL level.
2. In line with Section 5, we want to consider factor scores for both the general achievement factor and the specific factors in the multidimensional model.
3. In line with Section 6, we want to consider factor scores estimation that takes into account differences in student achievement level.

We note that (1) and (3) are quite controversial since these points raise the issue of estimating achievement scores based not only on the student's test responses but also on his or her instructional background. Bock (1972) has argued that prior information on groups should not be used in comparisons of individuals across groups. Nevertheless, it would seem that students who have had

very limited OTL on a set of test items will be unfairly disadvantaged in comparison with students with substantial exposure. The aim may instead be to obtain achievement scores for given instructional experiences.

Point (2) is of considerable interest. While a rough proxy for the general achievement score is easily obtainable as the total test score, adding of items corresponding to specific factors would involve only a few items resulting in a very unreliable score. As a contrast, estimating the specific factor scores draws on the correlated responses from all other items.

Muthén and Short (1988)[7] considered an example of the situation of (1) and (3). They generated a random sample of 1,000 observations from a model with 40 items measuring a unidimensional trait. Observations were also generated from 40 OTL variables and five other background variables. All background variables were assumed to influence the trait while the first 20 OTL variables had direct effects on their corresponding items, giving rise to exposure sensitivity in these items.

Among other results, Muthén and Short considered differences in factor score estimates using the above method and the traditional IRT method. In Table 11.3 comparisons of the two corresponding score distributions are presented by quartiles, broken down in two parts—students with a high total sum of OTL and students with a low sum. The table demonstrates that for students of the low OTL group, estimated scores are on the whole higher with the new method, corresponding to an adjustment for having had less exposure, while for the high OTL group the estimated scores are on the whole lower for the new method.

Other work by Muthén and Short has investigated Situation (2) and the precision with which scores for specific factors can be estimated. Once the estimated factor scores have been calculated

---

[7]The following estimation procedure was discussed in Muthén and Short (1988) and handles all three cases above. For various density and probability functions g, consider the posteriori distribution of the factors of η,

$$g(\eta|y,x) = \phi(\eta|x)g(y|\eta,x)|g(y|x) \quad (17)$$

Here, the first term on the right-hand side represents a normal prior distribution for η conditional on x, where as before x represents instructional background variables such as OTL and class type. In line with Section 5, the factor covariance matrix may be taken as constant given$x$, while the factor means may vary with x. The second term on the right-hand side represents the product of the item characteristic curves, which may vary in difficulty across OTL levels as discussed in Section 3.

TABLE 11.3
Trait Estimates by Traditional and New Approaches*

Low OTL Group					
	Traditional				
NEW	25%	50%	75%	100%	Total
	136	6	0	0	142
25%	−1.323	−0.610			−1.293
	−1.255	−0.724			−1.233
	10	125	5	0	140
50%	−0.783	−0.361	0.037		−0.375
	−0.624	−0.338	−0.119		−0.351
	0	13	111	7	131
75%		−0.094	0.309	0.827	0.297
		0.058	0.316	0.691	0.311
	0	0	6	124	130
100%			0.691	1.282	1.255
			0.834	1.308	1.286
Total	146	144	122	131	543
	−1.286	−0.347	0.317	1.257	
	−1.212	−0.318	0.324	1.275	

High OTL Group					
	Traditional				
NEW	25%	50%	75%	100%	Total
	99	9	0	0	108
25%	−1.306	−0.578			−1.245
	−1.349	−0.743			−1.298
	5	94	12	0	111
50%	−0.726	−0.340	0.049		−0.315
	−0.581	−0.366	−0.119		−0.349
	0	3	110	5	118
75%		−0.167	0.345	0.870	0.355
		0.022	0.322	0.640	0.327
	0	0	6	114	120
100%			0.653	1.386	1.349
			0.782	1.334	1.306
Total	104	106	128	119	457
	−1.278	−0.355	0.332	1.364	
	−1.312	−0.389	0.302	1.305	

*Entries are: Frequency
Mean value by the traditional approach
Mean value by the new approach

they may conveniently be related to various instructional variables and may also be studied for change from pretest to posttest.

PREDICTING ACHIEVEMENT

Given the explorations outlined in the previous sections, we may attempt to formulate a more comprehensive model for the data. Muthén (1988) proposed the use of structural equation modeling for this task. He discussed a model that extends ordinary structural modeling to dichotomous response variables, while at the same time extending ordinary IRT to include predictors of the trait. He studied part of the SIMS data using a model that attempted to predict a unidimensional algebra trait at the posttest using a set of instructional and student background variables from the pretest.

The set of predictors used and their standardized effects are given in Table 11.4. While pretest scores have strong expected effects,

TABLE 11.4
Structural Parameters with the Latent Construct as Dependent Variable

Regressor	Estimate	Estimate/S.E.
PREALG	0.68	11
PREMEAS	0.45	7
PREGEOM	0.33	5
PREARITH	2.09	16
FAED	0.07	1
MOED	0.02	0
MORED	0.18	3
USEFUL	0.45	7
ATTRACT	0.04	1
NONWHITE	−0.02	0
REMEDIAL	0.07	1
ENRICHED	0.22	3
ALGEBRA	0.56	4
FEMALE	0.14	6
LOWOCC	0.02	1
HIGHOCC	0.12	3
MISSOCC	0.05	2
NONW × REM	0.10	1
NONW × ENR	0.19	3
NONW × ALG	−0.18	−1
PREARITH × REM	−1.45	−3
PREARITH × ENR	−0.10	−1
PREARITH × ALG	−0.54	−2
NONW × PREARITH	−0.19	−1

class type, being female, father being in a high occupational category, and finding mathematics useful to future needs also had strong effects. The OTL variables had very small effects overall, perhaps due to the fact that each item's OTL variable has rather little power in predicting this general trait. Given the results of the previous sections, this modeling approach can be extended to include a multidimensional model for the set of both pretest and posttest items, predicting posttest factors from pretest factors, using instructional and student background variables as covariates, and allowing for differential item functioning in terms of exposure sensitivity.

ANALYZING CHANGE

The structural modeling discussed in the previous section is also suitable for modeling of change from pretest to posttest. Earlier in the chapter we pointed out that in terms of change, the SIMS data again exemplified complex population heterogeneity. For each item, a student may belong to one of three OTL groups, corresponding to the two types of no new learning and learning during the year. To again reach the goal of instructionally sensitive psychometrics for this new situation, we should explicitly model this heterogeneity. However, to properly model such complex heterogeneity is a very challenging task.

A basic assumption is that change is different for groups of students of different class types and OTL patterns. In a structural model where posttest factors are regressed on pretest factors, the slopes may be viewed as varying across such student groups, where student groups for whom a large degree of learning during the year (as measured by the set of OTL variables) has taken place, are assumed to have steeper slopes than the other students. Such an approach is rare in the field of psychometrics.

REFERENCES

Bock, R. D. (1972). Review of The dependability of behavioral measurements. *Science, 178*, 1275–1275A.

Bock, R. D., & Mislevy R. J. (1986). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*, 431–444.

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.

Gustafsson, J. E. (1988a). Hierarchical models of individual differences in

cognitive abilities. In R. J. Sternberg, (Ed.), *Advances in the psychology of human intelligence*, (Vol. 4). Hillsdale, NJ: Lawrence Erlbaum.

Gustafsson, J. E. (1988b). Broad and narrow abilities in research on learning and instruction. *Learning and individual differences: Abilities, motivation, methodology*. Symposium conducted at the Minnesota Symposium, Minneapolis.

Kao, C. F. (1990). *An investigation of instructional sensitivity in mathematics achievement test items for U.S. 8th grade students*. Doctoral dissertation, University of California at Los Angeles, Los Angeles, CA.

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

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

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

Mehrens, W. A., & Phillips, S. E. (1987). Sensitivity of item difficulties to curricular validity. *Journal of Educational Measurement, 24*, 357–370.

Miller, M. D., & Linn, R. L. (1988). Invariance of item characteristic functions with variations in instructional coverage. *Journal of Educational Measurement, 25*, 205–219.

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

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

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*. Mooresville, IN: Scientific Software, Inc.

Muthén, B. (1988). Some uses of structural equation modeling in validity studies: Extending IRT to external variables. In H. Wainer & H. I. Braun (Eds.), *Test Validity*. Hillsdale, NJ: Lawrence Erlbaum.

Muthén, B. (1989a). Latent variable modeling in heterogeneous populations. *Psychometrika, 54*, 557–585.

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

Muthén, B., Kao, C. F., & Burnstein, L. (1991). Instructional sensitivity in mathematics achievement test items: Application of a new IRT-based detection technique. *Journal of Educational Measurement, 28*, 1–22.

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

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika, 22*, 53–61.