1.027 and 1.050, respectively. Since X_2 and Y_2 correspond to parallel sections this suggests a consistent though slight departure from the constant practice effects model for the type of section X_2 and Y_2 represent.

In Table I we also give estimated values of K for each of the eight sections for which it can be estimated. The values range from $-.02$ to .21, which is typical of the data for this test.

Acknowledgments

The version of SPE described in this paper was developed by a team of Educational Testing Service staff that included R. Durso, J. Faggen, L. Hecht, L. Leary, M. McPeek, E. Stewart, L. Wightman, and the authors. Professor D. Rubin suggested several crucial ideas.

References

Angoff, W. H. (1984). *Scales, norms, and equivalent scores.* Princeton, NJ: Educational Testing Service. Reprinted from R. L. Thorndike (Ed.). (1971). *Educational measurement* (2nd ed.). Washington, DC: American Council on Education.

Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Eds.), *Test equating*. New York: Academic Press.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society (B), 39,* 1–38.

Holland, P. W., & Rubin, D. B. (Eds.). (1982). *Test equating.* New York: Academic Press.

Holland, P. W., & Thayer, D. T. (1981). *Section pre-equating the graduate record examination* (Program Statistics Research Technical Report No. 81-13). Princeton, NJ: Educational Testing Service.

Holland, P. W., & Wightman, L. E. (1982). Section pre-equating. A preliminary investigation. In P. W. Holland & D. B. Rubin (Eds.), *Test equating.* New York: Academic Press.

Koutsopoulos, C. J. (1961). *A linear practice effect solution for the counterbalanced case of equating* (Research Bulletin No. 61-19). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1950). *Notes on comparable scales for test scores* (Research Bulletin No. 50-48). Princeton, NJ: Educational Testing Service.

Rubin, D. B. (1974). Characterizing the estimation of parameters in incomplete-data problems. *Journal of the American Statistical Association, 69,* 467–474.

Rubin, D. B., & Szatrowski, T. H. (1982). Finding MLE of patterned covariance matrices by the EM algorithm. *Biometrika, 69*(3), 657–660.

Authors

PAUL W. HOLLAND, Director, Research Statistics Group, Educational Testing Service, Rosedale Road, Princeton, NJ 08541. *Specialization*: Statistics.

DOROTHY T. THAYER, Advanced Research Systems Specialist, Educational Testing Service, Rosedale Road, Princeton, NJ 08451. *Specialization*: Computational and statistical methodology.

Journal of Educational Statistics
Summer 1985, Volume 10, Number 2, pp. 121–132

A METHOD FOR STUDYING THE HOMOGENEITY OF TEST ITEMS WITH RESPECT TO OTHER RELEVANT VARIABLES

BENGT MUTHÉN
University of California, Los Angeles

KEY WORDS. *Multivariate probit, structural regression, latent variable, unidimensionality.*

ABSTRACT. Drawing on recently developed methodology for structural equation modeling with categorical data, this article proposes a new approach for investigating the behavior of a set of dichotomously scored test items in relation to a set of other relevant (observed) variables. This is achieved by considering a linear structural model relating the latent ability variable to a set of observed scores. The approach gives information on hypothesized unidimensionality and homogeneity of items with respect to these other variables. Some examples are given where a set of biology anchor items is related to a set of other related test scores obtained from the examinees.

In this article we propose an extension of Item Response Theory (IRT) modeling, which may be useful in gaining further knowledge of a set of dichotomously scored test items. Building on new statistical methodology presented in Muthén (1984), the latent ability variable involved can be linearly related to a set of other relevant variables, such as other test scores thought to be associated with the items in question, or demographic variables. This approach allows for a thorough investigation of the behavior of the test items, gaining important information on their homogeneity with respect to other relevant variables. This methodology would seem to be useful for scrutinizing a new set of test items, yielding information beyond that of the IRT approach.

In the next section, the methodology is outlined; in subsequent sections, data sets containing responses to some science items are analyzed, and further generalizations are considered.

Modeling

Consider the two-parameter normal ogive model of IRT for a set of p dichotomous test items. Using standard notation (see, e.g., Lord, 1980), we assume for the i^{th} variable,

$$Pr\,(y_i = 1 \mid \theta) = \Phi\,[a_i(\theta - b_i)]\,, \qquad (1)$$

This research was supported by grant no. SES-8312583 from the National Science Foundation.

#13

where $\Phi(.)$ is the standard normal distribution function. Also, conditional independence is assumed to hold. It is well known that the model may be reformulated as follows. Assume a latent response variable y_i^* underlying each y_i, such that

$$y_i = \begin{cases} 1, \text{ if } y_i^* > \tau_i, \\ 0, \text{ otherwise} \end{cases} \quad (2)$$

where τ_i is a threshold parameter for y_i^*, and assume that

$$y_i^* = \lambda_i \eta + \epsilon_i, \quad (3)$$

where λ_i is a slope ("loading") parameter, η replaces the latent ability variable θ of Equation 1, and ϵ_i is a residual variable with mean zero. The assumption that the ϵ_i's are multivariate normal and independent among themselves and with η gives the standard normal ogive model (see, e.g., Lord, 1980; Muthén & Christoffersson, 1981).

In IRT modeling, a crucial specification is that of unidimensionality, that is, the single variable should completely account for the interrelationships among the y's. In our model reformulation this is reflected by the ϵ's being uncorrelated. The appropriateness of this specification can be tested using the y-response patterns. It seems, however, that a valuable further test of the single η concept has not been considered in item analysis, namely by relating the y-responses to a set of other (observed) relevant variables.

Consider the following "structural equation" specification, adding to Equations 2 and 3:

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

where $\boldsymbol{\gamma}$ is a $q \times 1$ vector of "structural" regression coefficients, $\mathbf{x}$ is a $q \times 1$ vector of observed random variables, and ζ is a residual with mean zero, which is assumed to be normal and independent of $\mathbf{x}$. The x variables may, for instance, contain test scores on related subject matters and measures of general abilities needed for solving the tasks, such as verbal and quantitative skills. We may note that Equation 4 does not necessarily imply that $\mathbf{x}$ causes η, but merely that we are considering a linear relationship between the x's and η.

The specification of Equations 2, 3, and 4 implies that not only are the interrelations among the y's fully explained by η, but also that the relations between the y's and the x's are fully accounted for by the "intervening" variable η. This means that the y items are homogenous with respect to their relationship with the x's, so that no y is directly related to any x, but only indirectly via η. Hence, if we write

$$y_i^* = \lambda_i \eta + \boldsymbol{\beta}'_i \mathbf{x} + \epsilon_i, \quad (5)$$

it must hold that $\boldsymbol{\beta}_i = \mathbf{0}$ for $i = 1, 2, \ldots, p$. Combining Equations 3 and 4, it is seen that our model implies

$$y_i^* = \lambda_i \boldsymbol{\gamma}'\mathbf{x} + \lambda_i \zeta + \epsilon_i. \quad (6)$$

Consider in this connection the regressions of each y_i^* on x,

$$y_i^* = \boldsymbol{\pi}'_i \mathbf{x} + \delta_i. \quad (7)$$

Together with Equation 2, this defines a set of probit regression of y_i's on $\mathbf{x}$, where usually the variance of the residual δ_i is fixed to one. Using econometric terms, the $p \times q$ regression coefficients (the π-parameters) are "reduced-form" (unrestricted) regression coefficients, not taking into account the specific structure imposed by the single η concept. In contrast, Equation 6 expresses these $p \times q$ π's in terms of only $p-1+q$ regression coefficients by the terms $\lambda_i \boldsymbol{\gamma}'$ ($p-1$ λ's and q γ's). One λ-coefficient needs to be fixed (say to one) to determine the metric of η (this is similar to the case of confirmatory factor analysis modeling; see, e.g., Muthén, 1978). Hence, Equation 6 may impose a considerable amount of restrictions on the reduced-form π's of Equation 7, and yields a much more parsimonious model. For example, with $p=q=5$, 16 restrictions are imposed. We may note that these restrictions are such that the reduced-form (probit) regression coefficients differ only by a proportionality factor λ_i across the y-variables.

In addition to the restrictions that our model imposes on the reduced-form probit regression coefficients, further restrictions are involved. These can be viewed as restrictions on the residual correlations of the multivariate probit model's reduced-form. Comparing Equations 6 and 7, we find that covariances among reduced-form δ's can be written as

$$V(\delta_i, \delta_j) = \lambda_i \lambda_j V(\zeta); \; i \neq j. \quad (8)$$

As in the multivariate probit model, we will standardize the variances of the residuals of our structural model in Equation 6 to one,

$$\lambda_i^2 V(\zeta) + V(\epsilon_i) = 1, \quad (9)$$

so that Equation 8 represents correlations. We note that Equation 8 has the structure of an ordinary (observed continuous variables) single-factor model. Whereas the correlations among the y^*'s (and thereby the y's) are all zero conditional on η, Equation 8 gives the correlations among the y^*'s conditional on the x vector. There are $p(p-1)/2$ reduced-form correlations in Equation 8, expressed in terms of only $p-1$ λ's and a single $V(\zeta)$ parameter. Hence, we find that our structural model imposes two kinds of restrictions; across y's and x's restricting reduced-form probit regression coefficients, and among the y's conditional on the x's restricting reduced-form probit residual correlations. The total number of restrictions imposed by our structural model is therefore $p \times q + p(p-1)/2 - (p-1+1+q)$. To continue the example with $p=q=5$, we have a total of 25 restrictions.

The modeling of Equations 2, 3, and 4 would seem to be of great usefulness in test item analysis. It extends the IRT framework to also consider relations with other relevant variables. This should be a valuable tool in studying item homogeneity. For instance, it may be found that certain science test items are directly related to, say, vocabulary scores over and beyond their relation to science ability. Such information may be useful in item bias studies, where the source of item bias may be located to aberrant relationships with some other variable. The relationship to demographic variables may also be considered.

Furthermore, our approach gives valuable information regarding validity and reliability in the items' measurement of η. Say that the y's consist of a newly developed set of items administered to a sample that also took a conventional test x. Validity would then be reflected by the estimated structural relation between y and x. At the same time, we may consider item reliability in the measurement of η as the amount of variation in y^* accounted for by η.

The new type of modeling outlined above may be viewed in the general framework of latent variable structural equation modeling with multiple categorical indicators, as recently developed by Muthén (1984), extending work in Muthén (1979). Following Muthén, estimation of such models may be carried out by a three-stage limited information procedure, which also gives large-sample standard errors of estimates and a large-sample chi-square measure of fit to the restrictions imposed; see Muthén (1984) for technical details and Muthén (1983) for an overview of related modeling and estimation. A computer program LISCOMP (Analysis of linear structural equations with a comprehensive measurement model) has been developed by the author for general users. It was used in the following applications.

Applications

We will consider a set of dichotomously scored biology items from the IEA Study (Comber & Keeves, 1973). Two samples will be used, females in England for grade 9 ($N=712$) and grade 12 ($N=1075$). We will be particularly concerned with four anchor items that were administered in both grade 9 and grade 12. For grade 12, we will consider an additional biology item taken from the same test. The item wording is given in Table I together with item means (p-values).

In grade 9 the biology test consisted of 19 items, whereas in grade 12 the test consisted of 16 items. For the set of x variables to be considered, we have chosen the following related test scores. Lacking a separate biology test, a biology score was produced by deleting the four anchor items in each grade. Hence the biology score consists of 15 and 12 summed items for the two grades, respectively. In addition, the following scores were used—chemistry, physics, practice, reading comprehension, and word knowledge. The chemistry test consisted of 19 and 16 items, respectively. Physics consisted of 20 and

16 items. Practice consisted of 20 and 16 items and was intended to measure practical work in the laboratory or the field. Reading comprehension consisted of 52 and 54 items, respectively (tests C and D combined). Word knowledge consisted of 40 items in both grades (not all the same) and was intended to measure general ability, where items consisted of a pair of words that had either nearly the same or nearly opposite meaning. KR-20 test reliability coefficients are available only for males and females together. They were for chemistry, .70/.75; physics, .72/.75; practice, .68/.64; reading comprehension, .87/.72; word knowledge, .83/.77; and biology, .60/.52 (all 19/16 items). Means, standard deviations, and correlations for these six x variables are given in Table II.

Consider first the analysis of the grade 9 data. In relation to the model of Equations 2, 3, and 4, we have $p=4$ and $q=6$, where the four y's (the anchor items) will be denoted $A1$, $A2$, $A3$, $A4$. We hypothesize that these anchor items measure a single ability variable, with all items being only indirectly related to the six x-variables via the ability variable. Throughout our analysis, we will (arbitrarily) choose to determine the scale of η by fixing the loading of $A3$ to one. The total number of restrictions imposed by the model is 20. The large-sample chi-square test with 20 degrees of freedom obtained the value 14.99, hence indicating an excellent fit. The concept of a single ability variable cannot be rejected even when relating the items to a set of other variables. At grade 9, the four anchor items are homogeneous with respect to these other variables.

The estimated parameters are given in Table III. First, consider the measurement parameters, the loadings of each of the four anchor items. We note that $A4$ has a considerably weaker relationship with the ability than do the other measures, whereas the others have largely identical relations. This is also reflected in the estimated item reliabilities. These are calculated by using the sample covariance matrix of $\mathbf{x}$ and the estimated parameters to form the proportion of variation in each y^* accounted for by η variation.

Turning to the structural parameters, most other relevant variables have coefficients that are significant at the 5% level (the ratios Estimate/Standard error have approximate large-sample standard normal distributions). The structural regression R^2 is defined as the variation in η accounted for by x variation. The estimated R^2 is obtained from the sample covariance matrix for x and the estimated structural regression coefficients and residual variance. We note that the summed score of the remaining 15 biology items (biology (15)) has a highly significant regression coefficient. We may conclude that there seems to be reasonable validity in the anchor items' measurement of the biology ability. We may also note that the Reading Comprehension score has an insignificant relationship to the ability in question, whereas the highly correlated (.678) word knowledge score seems to be almost as important as the

TABLE I

Item Wording and Means for a Sub-Set of Biology Test Items[a]

Item	Wording	Mean Grade 9	Mean Grade 12
A1	All of the following are aspects of the reproductive process. Which one of them must occur before we can be certain that fertilization has taken place? A. A male organism must find a mate B. Reproductive organs must be produced *C. The nucleus of a male gamete must fuse with that of a female gamete D. A spermatozoon must reach an egg cell E. A female gamete must provide a store of food for embryo	.392	.828
A2	The energy for photosynthesis is generally obtained from A. chlorophyll B. chloroplasts *C. sunlight D. carbohydrates E. carbon dioxide	.677	.931
A3	The diagram below shows an example of interdependence among aquatic organisms. During the day the organisms either use up or give off (a) or (b), as shown by the arrows. *A. (a) is oxygen and (b) is carbon dioxide B. (a) is oxygen and (b) is carbohydrate C. (a) is nitrogen and (b) is carbon dioxide D. (a) is carbon dioxide and (b) is oxygen E. (a) is carbon dioxide and (b) is carbohydrate	.565	.853
A4	Why is it that your body temperature does not fall even though you lose heat continually? A. The blood distributes heat round the body *B. Respiration results in the liberation of heat C. Heat is continually being absorbed from the sun D. Hot meals are eaten regularly E. Warm clothes are good insulators	.150	.384
I1	A student wrote the following note on a laboratory project: "Using a cork-borer, I obtained several cylinders from a large potato. The cylinders were 7 cm long and 0.5 cm in diameter. I kept them on a dry plate and measured them again on the following day. I found that all of them had become shorter and thinner." The student then put the cylinders into a beaker of tap water and wrote: "If I measure them tomorrow I shall find that they have all returned to their original size." In writing this last sentence, the student was A. making a statement of fact B. making an observation C. drawing a tentative conclusion D. describing an experimental procedure *E. making a hypothesis	—	.284

[a] Source: Comber, L. C., & Keeves, J. P. (1973).

biology (15) score. Note, however, that the word knowledge score is composed of a larger set of items than biology (15) with a more reliable total score, presumably yielding less of an attenuation in regression slope than the more unreliable biology (15) score.

It is now of interest to study the behavior of the anchor items for a group of individuals at three grades later. Testing the model in grade 12 resulted in a 20 degrees of freedom chi-square of 26.41, with a probability level of 0.16. As for grade 9, there is no evidence of overall misfit to the single ability structural model. However, in this case, a significant improvement in the model is obtained by allowing a direct relationship ("path") between $A4$ and Word Knowledge. This modified model is of the type depicted in Equation 5, with a single non-zero β-coefficient. An interpretation will be attempted shortly. The chi-square value was 21.85 with 19 degrees of freedom. The chi-square difference of 4.56 is with one degree of freedom significant at the 5% level, so that the inclusion of this path seems warranted.

The estimated model is given in Table IV. The estimated loadings are not directly comparable to those for grade 9 because the ability variable does not have the same variance across grades. A more appropriate comparison is obtained by the estimated item reliabilities, where we note a drop in reliability

TABLE II
Descriptive Statistics for Other Relevant Variables

	Grade 9 (N = 712)		Grade 12 (N = 1075)	
Variable	Mean	Std. dev.	Mean	Std. dev.
Chemistry	6.909	3.600	5.207	2.906
Physics	9.899	3.560	6.335	2.824
Practice	6.618	3.181	5.186	2.286
Reading Comprehension	26.376	6.573	30.992	4.940
Word Knowledge (WK)	33.275	9.130	38.620	6.664
Biology (15/12)	6.746	2.670	6.367	2.068

Correlation Matrix[a]

	Chemistry	Physics	Practice	Reading	WK	Biology
Chemistry	1.000	.580	.549	.082	.262	.403
Physics	.648	1.000	.577	.157	.353	.428
Practice	.645	.654	1.000	.174	.338	.446
WK	.653	.643	.607	.678	1.000	.381
Biology	.562	.574	.562	.487	.631	1.000

[a] Grade 9 correlations are given below the diagonal and Grade 12 correlations above.

TABLE III
Estimated Model for Grade 9: Anchor Items

	Estimate	Standard error	Estimate/ Std. error	
		Loadings		Reliability
A1	.909	.138	6.59	.22
A2	.992	.154	6.44	.25
A3	1.000[a]	—	—	.26
A4	.574	.123	4.67	.10
	Structural regression coefficients			
Chemistry	.037	.013	2.85	
Physics	.024	.013	1.85	
Practice	.046	.015	2.97	
Reading Comp.	−.006	.006	−1.00	
Word Knowledge	.017	.005	3.40	
Biology (15)	.056	.016	3.50	
	Structural regression residual variance			
	.045	.037	1.216	
	Structural regression $R^2 = 0.86$			

[a] Fixed parameter.

in $A2$ and $A3$. In Table IV, the $A4$ reliability is not considered, because it is not only related to the ability variable but also directly to word knowledge. The pattern of ratios Estimate/Standard errors is largely unaltered as compared to grade 9. However, there is a slight increase for most variables in grade 12, and the R^2 is also larger.

Turning to the direct path between $A4$ and WK, we find a negative estimate. We may interpret this in the following way. The model states that $A4$ is predicted by the ability variable, but that the probability correct on $A4$ is overstated as explained by this ability alone. The (reduced-form) relationship between $A4$ and word knowledge is not as strong as predicted by the model postulating an indirect relationship via the ability variable. A negative direct path adjusts for this. This means that relative to the other anchor items, $A4$ does not have as strong a component of word knowledge. The anchor items are not homogeneous with respect to this variable in grade 12. No such phenomenon could be observed in grade 9.

As a final example of our methodology, let us consider another item in the biology test ($I1$) in connection with the four anchor items for grade 12. For the anchor items the model is the same as above (reported in Table IV). Let us test the additional hypothesis that $I1$ is only indirectly related to the x's via the ability variable, analyzing the five dichotomous items together. This model

obtained a chi-square value of 61.45 with 28 degrees of freedom, which represents a severe misfit. Inspection of the source of this misfit revealed that additional direct paths were required between $I1$ and both reading comprehension and biology (12); this brought down the chi-square to 24.73, losing only two degrees of freedom. Both of these extra paths obtained positive values. The positive direct path to reading comprehension seems plausible in light of the wording of these five items; the items are not homogenous in this sense. To some extent the biology relation is to be expected because $I1$ was included in biology (12), although it would seem that this would not have as strong an effect. Furthermore, $I1$ does not obtain a significant loading as measuring the ability variable (see Table V), which could mean that $I1$ does not tap the same ability as that measured by the four anchor items.

It should be noted that because both $I1$ and $A4$ have direct paths to x-variables that are correlated, conditional independence between these two items, given the ability variable, cannot hold. By our modeling, this violation of assumption for IRT modeling is given an interpretation.

Apart from $I1$, the estimates remain largely the same as in Table IV (see Table V).

TABLE IV
Estimated Model for Grade 12: Anchor Items

	Estimate	Standard error	Estimate/ Std. error	
		Loadings		Reliability
A1	.798	.153	5.22	.10
A2	.846	.187	4.52	.13
A3	1.000[a]	—	—	.17
A4	1.083	.199	5.44	—
	Structural regression coefficients			
Chemistry	.038	.013	2.92	
Physics	.027	.013	2.08	
Practice	.057	.016	3.56	
Reading Comp.	.000	.006	.00	
Word Knowledge	.018	.006	3.00	
Biology (15)	.077	.017	4.53	
	Structural regression residual variance			
	.010	.032	.31	
	Structural regression $R^2 = 0.95$			
	Direct path A4—Word Knowledge			
	−.022	.011	2.00	

[a] Fixed parameter.

TABLE V
Estimated Model for Grade 12: I1 and Anchor Items

	Estimate	Standard error	Estimate/ Std. error	
		Loadings		Reliability
I1	.179	.156	1.15	—
A1	.802	.153	5.24	.11
A2	.855	.189	4.52	.13
A3	1.000[a]	—	—	.16
A4	1.089	.200	5.45	—
	Structural regression coefficients			
Chemistry	.039	.012	3.25	
Physics	.027	.013	2.08	
Practice	.055	.016	3.44	
Reading Comp.	.000	.006	.00	
Word Knowledge	.018	.006	3.00	
Biology (12)	.076	.016	4.75	
	Structural regression residual variance			
	.007	.031	.23	
	Structural regression $R^2 = 0.96$			
	Direct paths			
I1—Reading	.031	.009	3.44	
I1—Biology	.222	.033	6.73	
A4—WK	−.022	.011	2.00	

[a] Fixed parameter.

Conclusion

In this article we proposed a powerful new methodology for investigating the behavior of dichotomously scored test items as related to other relevant variables. It was shown that this approach can give new detailed insights into the hypothesized homogeneity and unidimensionality of items.

The choice of relevant variables must be made carefully. Statistically, the set of relevant variables would be one that makes the assumption of random, normal residuals δ_i in Equation 7 tenable (note, however, that the distribution of η need not be normal). The amount of measurement error in these variables should be low to avoid problems of slope attenuation. For the item responses, we note the model restriction of a zero lower asymptote for $Pr(y = 1 \mid \eta)$. Items with a high probability of guessing should be avoided in order not to unduly attenuate the probit slopes. Further extensions to include lower asymptote parameters are desirable.

It should be noted that this article exemplifies only a subset of possibilities

that are available in the framework of the methodology of Muthén (1984), extending IRT modeling. Some more complex types of modeling would also seem to be useful. We may consider multiple latent ability variables with separate sets of items measurements. The interrelations of these abilities can be explored in relation to other relevant variables. Furthermore, with categorical x variables, a powerful simultaneous analysis of multiple groups of examinees can be carried out. This makes it possible to investigate invariance hypotheses regarding both measurement and structural parameters, which would be valuable, for example, in studies of test item bias.

References

Comber, L. C., & Keeves, J. P. (1973). *Science education in nineteen countries* (International Association for the Evaluation of Educational Achievement: International Studies in Evaluation I). Stockholm: Almqvist & Wiksell.

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

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

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

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

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

Muthén, B., & Christoffersson, A. (1981). Simultaneous factor analysis of dichotomous variables in several groups. *Psychometrika*, *46*, 407–419.

Author

BENGT MUTHÉN, Associate Professor, Graduate School of Education, UCLA, Los Angeles, CA 90024. *Specializations*: Latent variable structural equation modeling with categorical and other nonnormal data.

Journal of Educational Statistics
Summer 1985, Volume 10, Number 2, pp. 133–142

MULTIPLE GROUP IRT MODELING: APPLICATIONS TO ITEM BIAS ANALYSIS

BENGT MUTHÉN
and
JAMES LEHMAN
University of California, Los Angeles

KEY WORDS. *Measurement invariance, factor analysis, random disturbances.*

ABSTRACT. This article shows the applicability of new methodology for multiple-group factor analysis of dichotomous variables. Situations are considered where the same set of test items has been administered to more than one group of examinees. The new methodology is contrasted with the IRT approach to item bias analysis. An example is given in which females and males have taken a certain biology test.

The purpose of this article is to show the applicability of IRT-related methodology developed by Muthén and Christoffersson (1981). We will be concerned with the situation in which the same set of dichotomously scored test items has been administered to more than one group of examinees. In such situations, issues of item invariance are of primary concern. IRT modeling has been proposed to study these issues in an effective way, particularly under the rubric of item bias (see, e.g., Linn, Levine, Hastings, & Wardrop, 1981; Lord, 1977, 1980). Here, biased items (not showing invariance across groups) are singled out and modified or discarded, after which the values of the latent ability variable are estimated for each individual. The alternative approach to be discussed here differs from the above in three important respects: (a) for each item, the model is in a certain sense more general than conventional IRT models; (b) a simultaneous analysis is performed of the various groups under certain testable invariance restrictions of parameters; and (c) individual values of the latent ability variable are not estimated (as is also the case in marginal maximum likelihood estimation), but rather group means and variances.

The new approach gives a very powerful multiple-group analysis. An important product of the new approach is that items identified as biased by conventional IRT analysis need not be discarded but can still be used to estimate group parameters. In the next sections the new methodology is outlined and illustrated by an achievement test example.

The research of the first author was supported by Grant No. SES-8312583 from the National Science Foundation.