Using Restricted Latent Class Models to Map the Skill Structure of Achievement Items
Author(s): Edward H. Haertel
Source: *Journal of Educational Measurement*, Vol. 26, No. 4 (Winter, 1989), pp. 301-321
Published by: National Council on Measurement in Education
Stable URL: http://www.jstor.org/stable/1434756
Accessed: 30/08/2010 06:34

Journal of Educational Measurement
Winter 1989, Vol. 26, No. 4, pp. 301–321

Using Restricted Latent Class Models to Map the Skill Structure of Achievement Items

Edward H. Haertel
Stanford University

This paper presents a new method for using certain restricted latent class models, referred to as binary skills models, *to determine the skills required by a set of test items. The method is applied to reading achievement data from a nationally representative sample of fourth-grade students and offers useful perspectives on test structure and examinee ability, distinct from those provided by other methods of analysis. Models fitted to small, overlapping sets of items are integrated into a common skill map, and the nature of each skill is then inferred from the characteristics of the items for which it is required. The reading comprehension items examined conform closely to a unidimensional scale with six discrete skill levels that range from an inability to comprehend or match isolated words in a reading passage to the abilities required to integrate passage content with general knowledge and to recognize the main ideas of the most difficult passages on the test.*

The internal structure of tests is a perennial concern of both classical test theory and item response theory (IRT). Factor analysis of test items continues to be an active area of research (Bartholomew, 1980; Christoffersson, 1975; Mislevy, 1986; Muthén, 1978), and new methods are appearing for testing dimensionality in the framework of latent trait models (Gustafsson, 1980; Molenaar, 1983). Most of this work is predicated on the assumption that examinee ability can be represented by one or more continuous latent variables. But there has also been increased attention to the use of discrete latent structure models, or *latent class models,* for item response data (Bergan, 1983; Haertel, 1984a, 1984b; Macready & Dayton, 1980; Traub & Lam, 1985). Under these models, examinees are assumed not to possess variable amounts of continuously distributed abilities, but rather to conform to exactly one of some small number of discrete latent classes.

This paper introduces a family of latent class models referred to as *binary skills models,* in which each latent class is identified with a distinct pattern of dichotomous skills. A procedure is demonstrated for creating a map of binary

Preparation of this paper was supported in part by a grant from the Spencer Foundation. I am grateful to David E. Wiley for his helpful discussions of both the methods developed in this paper and the logical foundations of the binary skills model. I thank Clifford C. Clogg for information about the application of latent class models to large numbers of items. I also thank Wendy M. Yen and anonymous reviewers for their helpful suggestions.

Copies of this paper may be requested from Edward H. Haertel, School of Education, Stanford University, Stanford, California 94305-3096.

skills that could account for patterns of performance across the items of an achievement test. Binary skills might arise from several sources, including both the organization of the school curriculum and the information processing requirements of different test items. Possible reading comprehension skills might include facility with particular forms of anaphoric reference, possession of schemata for specific genres of prose passages, or possession of sufficient specific background information about the topic of a particular reading passage. Whether any of these possible skills exists is an empirical question, and it is clear that none of them is truly dichotomous. Even though degrees of proficiency might be observed, however, research has shown that for typical school populations, abilities in reading, algebra, and other areas can be well approximated by skill dichotomies (Haertel, 1984a, 1984b; Harris & Pearlman, 1978; Macready & Dayton, 1977).

Method

Subjects and Data Source

The methods developed in this paper are illustrated using 37 items from the reading comprehension subtest of the Metropolitan Achievement Tests (MAT), Form F, elementary level (Durost, Bixler, Wrightstone, Prescott, & Balow, 1970). This 45-item subtest consists of eight brief reading passages, each followed by several four-choice questions. It is recommended for use with fourth- and fifth-grade children. For this study, item response data were obtained for the nationally representative sample of nearly 63,000 fourth-grade students in nearly 1,000 schools tested in 1972 as part of the Anchor Test Study Norming Sample (Loret, Seder, Bianchini, & Vale, 1974).

Each item of the MAT reading comprehension test is classified by the test's authors as measuring literal comprehension, inference, word meaning in context, or passage main idea. Earlier analyses of these data showed that about half of the items, primarily the more difficult, so-called inference items, conform to a two-latent-class model, implying that these items could be referenced to a single latent skill dichotomy. The structure of the remaining items, however, primarily the easier, literal comprehension items, resisted interpretation (Haertel, 1984b). This paper offers further insight into the variety of skills these items measure.

In analyzing these data, each examinee's responses were weighted according to the sampling weights included on the Anchor Test Study public-use data tape. This weighting had only a small effect, but was necessary to obtain unbiased estimates of response pattern proportions in the population defined for the Anchor Test Study. Unweighted sample proportions would not represent any natural population. A design effect adjustment (Kish, 1967) was incorporated as a rough correction for the effects of stratified cluster sampling on precision. This adjustment affected fit statistics and standard errors, but had no effect on parameter estimates.

Binary Skills Models

The latent class models used in this paper characterize proficiency in terms of unobservable binary skills, defined independently of any particular items. Each

examinee commands a subset of these skills, and specification of that subset completely characterizes his or her ability. It is typically assumed that some patterns of skill presence and absence cannot occur. With two skills, for example, it might be assumed that no one possesses the second skill who does not possess the first. That would leave as permissible patterns neither skill, only the first, or both skills. There is one latent class for each permissible skill pattern. These always include a class for possession of none of the skills (the null class) and a class for possession of all of them (the full class). Thus, letting z be the number of dichotomous skills and k the number of latent classes, $2 \leq k \leq 2^z$. The population distribution of examinee abilities is completely summarized by the vector $\lambda = (\lambda_1, \ldots, \lambda_k)'$, where λ_j is the proportion of examinees commanding just that subset of skills identified with class j. Note that $\Sigma \lambda_j = 1$.

Each item requires some specific subset of the binary skills for its solution. Together with examinee skill possession, these item skill requirements completely determine the latent response of each examinee to each item, which is 1 if an examinee possesses all of an item's requisite skills and 0 otherwise. Because examinees in the same latent class possess the same skills, they must have the same latent response to any given item. Let v_{ji} denote the latent response to item i for class j. Define the latent response pattern $\mathbf{v}_j = (v_1, \ldots, v_n)'$ for latent class j as the vector of latent responses to all items for examinees conforming to that class. Latent response patterns for the null class and for the full class are vectors of all zeros and of all ones, respectively. It will be convenient to define a latent difficulty parameter $\delta_i = \Sigma v_{ji}\lambda_j$ for each item i, equal to the proportion of examinees whose latent response to that item is one.

The relation between a latent response and the corresponding manifest response is probabilistic and is governed by two classification parameters unique to each item: (a) the conditional probability of a correct manifest response to item i given a latent response of 0 is π_{1i}, referred to as item i's *false positive probability*, and, (b) the conditional probability of a correct manifest response given a latent response of 1 is π_{2i}, referred to as item i's *true positive probability*. The (conditional) probability of a false negative response is $1 - \pi_{2i}$. It follows from these definitions that an item's manifest p-value, $p_i = \pi_{1i}(1 - \delta_i) + \pi_{2i}\delta_i$. By writing $p_i = \delta_i + \pi_{1i}(1 - \delta_i) - (1 - \pi_{2i})\delta_i$, it can be seen that the discrepancy between p_i and δ_i represents a net bias due to false positives, which occur at a rate of π_{1i} among examinees whose latent response is zero, and false negatives, which occur at a rate of $(1 - \pi_{2i})$ among examinees whose latent response is one.

For any latent class j, let $\pi_{ji} = \pi_{1i}(1 - v_{ji}) + \pi_{2i}v_{ji}$ denote the conditional probability of a manifest correct response to item i. Note that this probability is equal to either π_{1i} or π_{2i}, according to the value of v_{ji}. Let $\mathbf{u} = (u_i \ldots u_n)'$ denote an arbitrary pattern of manifest responses to the n items, where $u_i = 0$ if item i is answered incorrectly and $u_i = 1$ if it is answered correctly. Manifest responses are assumed to be conditionally independent given latent class, and so the conditional probability of $\mathbf{u}$ given possession of the skill subset for latent class j is

$$P(\mathbf{u}|\lambda_j) = \prod_{i=1}^{n} (1 - \pi_{ji})^{(1-u_i)}\pi_{ji}^{u_i}. \qquad (1)$$

The unconditional probability of this response pattern is

$$P(\mathbf{u}) = \sum_{j=1}^{k} \lambda_j \prod_{i=1}^{n} (1 - \pi_{ji})^{(1-u_i)} \pi_{ji}^{u_i}. \quad (2)$$

Formally, the models described by equations 1 and 2 are restricted latent class models with exactly two distinct values of classification parameters for each item.

Relation of Binary Skills models to earlier measurement models. These are the first latent class models to employ discrete latent response variables, but the idea of unobservable (latent) responses intervening between some more fundamental latent structure and manifest responses can be traced to the discriminal process variables of psychophysical scaling (Bock & Jones, 1968; Thurstone, 1927). A related concept is central to modern methods of factor analysis for dichotomized variables (Mislevy, 1986). In these Thurstonian models, continuous, stochastic latent response variables mediate between continuous latent processes and discontinuous observable responses. They are related to manifest responses via threshold parameters.

In the binary skills models of this paper, both underlying skills and manifest responses are discrete. The intervening latent response represents the true state of the examinee, either able or unable to solve a particular item. The intent of the measurement is to determine this true state, but random, ancillary influences may result in manifest responses that do not match the corresponding latent responses. The conditional probabilities of such random influences changing responses are π_{1i} (incorrect to correct) and $1 - \pi_{2i}$ (correct to incorrect). A similar construction is suggested in the context of a simpler latent class model by Harris and Pearlman (1978), who interpret latent class membership as determining an examinee's "Platonic true score."

Binary skills models are distinguished from earlier latent class models for item response data in both the form of their parameter restrictions and the interpretive framework from which these restrictions are derived. Formally, they include as a subset Lazarsfeld's latent distance models (Lazarsfeld, 1950a, p. 410, 1950b, p. 441; Lazarsfeld & Henry, 1968, chap. 5), but latent distance models treat classes as representing linearly ordered categories rather than patterns of binary skills. A binary skills model is formally equivalent to a latent distance model if and only if its latent response patterns form a (latent) Guttman scale.

Model estimation. Applications of latent class models have been limited by problems of estimation. Although efficient algorithms are now known for maximum likelihood estimation of model parameters, available computer programs still require a tabulation of frequencies for all possible response patterns, sharply limiting the number of variables, or items, that can be included in a single analysis. Maximum likelihood estimation for larger numbers of items is also complicated by the fact that likelihood functions for models of even moderate complexity possess multiple local maxima (Hartigan, 1975). Models with up to seven latent classes were fitted to 38 binary variables by Aitkin, Anderson, and Hinde (1981), but they chose not to interpret models with more than three classes due to ambiguities arising from the multiple solutions they obtained with different starting values.

For this study, maximum likelihood estimates of the λ_j and the π_{ji} were obtained using Goodman's (1974, 1975) iterative proportional scaling algorithm, as implemented in the maximum likelihood latent structure analysis (MLLSA) computer program (Clogg, 1977). This has been shown to be an EM algorithm (Dempster, Laird, & Rubin, 1977; Everitt, 1984; Goodman, 1979), assuring its asymptotic properties. The MLLSA program was modified by the present author to include calculation of approximations to the asymptotic standard errors of parameter estimates, based on standard theory for maximum likelihood estimation (Rao, 1973).

Testing goodness of fit. The likelihood ratio chi square computed for each model is $G^2 = -2\Sigma\, f_m \ln (f_m/e_m)$, is computed for each model, where f_m and e_m are the observed and expected (predicted) frequencies for response pattern m, and the summation is over all of the 2^n correct/incorrect response patterns to the n items for which $f_m > 0$. For a binary skills model with k latent classes, this statistic is asymptotically distributed as chi square on $2^n - 2n - k$ degrees of freedom (Rao, 1973). A nonsignificant chi square only indicates that the data are consistent with the model and cannot confirm that the model is correct.

When an additional latent class is introduced, the reduction in the chi square is referred to as a difference chi square and may be referenced to the chi square distribution on one degree of freedom to determine whether the improvement in fit associated with the additional class is statistically significant. It is generally recognized (Aitkin, Anderson, & Hinde, 1981, p. 424; Everitt & Hand, 1981) that the asymptotic distribution of this statistic is not necessarily chi square due to violation of a regularity condition. The less inclusive model is obtained from the more inclusive by constraining the proportion in one class to zero, which is at a boundary, not an interior point, of the parameter space. Both this caveat and the approximate nature of the design effect adjustment militate against rigid interpretation of probability levels for difference chi squares.

Model identification. A conceptually distinct latent class exists for each permissible skill pattern, but it may be impossible to distinguish all these classes empirically using a given set of items. Depending upon the items' skill requirements, latent response patterns for two or more classes may be identical. If classes j and j′ have the same latent response pattern, then λ_j and $\lambda_{j'}$ are not identified. A fully identified model can be obtained, however, by treating classes j and j′ as a single, pooled class and estimating $(\lambda_j + \lambda_{j'})$. The remaining model parameters are unaffected by this collapsing.

It can also happen that some parameters are not identified even though all of the latent response patterns are distinct (Haertel, 1984a, p. 335). Suppose, for example, that the skills required by item i are a subset of those required by each of the other items taken one at a time. It would follow that the latent response to item i is one for all classes except the null class. (Otherwise, there would be a class for examinees unable to solve item i but able to solve some other item, a contradiction.) Suppose there is a class j for examinees able to solve item i but no other item. (Note that $\mathbf{v}_j$ includes a 1 for item i and a zero for every other item.) Let the null class be class 1. There is then a dependency among λ_1, λ_j, and π_{1i}. The sum $\lambda_1 + \lambda_j$ can be determined, but changes in the relative sizes of classes 1 and j

may be compensated by adjustments to π_{1i}. An interpretation of this dependency is that there is no way to tell whether only a few examinees are unable to solve item i, and, for those few, false positives are quite unlikely or whether a larger fraction of examinees are unable to solve item i, but each of them is more likely to give a false positive response. An analogous problem can arise if there is a class j′ for which the latent response vector contains only a single zero. Denoting the full class by k, a dependency then exists among $\lambda_{j'}$, λ_k, and π_{2i}, although the sum $\lambda_{j'} + \lambda_k$ can be estimated. These situations in which a model is not fully identified are analogous to the problem in factor analysis of identifying a factor measured by only one variable.

For purposes of estimation, a fully identified model can again be obtained by pooling classes 1 and j, or j′ and k, but, unlike the case of pooling latent classes with identical latent response patterns, this pooling results in an inflated estimate of π_{1i} or a depressed estimate of π_{2i}. These identification problems, as well as extensions to more complex identification problems, are further discussed and illustrated in the Appendix.

Procedure for Mapping the Skill Requirements of a Test

A given set of items can be used to distinguish at most as many latent classes as yield distinct latent response patterns for that set of items, and, as explained above and illustrated in the Appendix, even classes with distinct latent response patterns may not be distinguishable. As more items are considered, more latent classes may be identified, and so the binary skills model for a large set of items can easily include more latent classes than are distinguishable using some particular subset of those items.

If the binary skills model for an entire set of items were known, it would be straightforward to derive the model for any subset. This would entail first pooling classes with common latent response patterns across the subset of items, and then further collapsing latent response patterns with only a single zero or only a single one into the null or full classes, respectively. The latter step would require adjustments to the values of false positive or true positive probabilities for one or more items.

In fact, the problem is to proceed in the opposite direction. The binary skills model for a large set of items must be inferred from the separate analyses of smaller item sets. In the procedure of this paper, this is accomplished by (a) fitting models to overlapping six-item sets, then (b) revising these models by partitioning some latent classes into two or more classes and if necessary adjusting the π_{ji}, and finally (c) determining which of these revised latent classes, detected using different six-item sets, in fact represent the same class in the binary skills model for the full set of items.

The first of these steps involves estimation of fully identified, restricted latent class models. For each six-item set, several models are fitted and one final model is chosen based on goodness of fit and on the estimated size of additional latent classes in alternative models. These final models yield distinct estimates of π_{1i}, π_{2i}, and δ_i from each six-item set in which item i appears. In the second step, discrepancies among these distinct estimates for each item are resolved by the

introduction of additional, nonidentified latent classes into the initial models. The third step, matching classes across sets, involves (a) summarizing information about each pair of items appearing together in one or more six-item sets, (b) using this summary to find clusters of items all requiring the same skills, (c) relating these clusters to one another to form a partial map, and finally, (d) completing this map by incorporating those items not appearing in a cluster. The application of these three steps to 37 items from the MAT is described in the following section.

Results

Selection and Analysis of Separate Item Sets

Analysis of the MAT began with the selection of 12 six-item sets, as described by Haertel, Korpi, and Capell (1982). In order to increase the probability of detecting any distinct skills that subsets of the items might require, each of these sets included items with contrasting features, for example, easy versus difficult items, or items classified as requiring literal comprehension versus inference. Of the 45 items on the MAT, 37 were included in one or more of these twelve sets. An additional six sets were later added, using these same 37 items.

Several binary skills models were fitted to each of the 18-item sets, beginning with the two-latent-class model and proceeding to more complex models by incorporating additional classes. Additional classes were chosen based on an analysis of residuals from previous models as described by Haertel (1984a, p. 337), and analyses of each six-item set were continued until the improvement in fit associated with further additional latent classes was not statistically significant. No latent class was retained unless its inclusion resulted in a statistically significant difference chi square.

The statistical significance of a new class j could be judged in two ways, using either a difference chi square comparing models with and without class j, or by determining whether a confidence interval about λ_j, constructed using its asymptotic standard error, including zero. Despite the lack of rigorous justification for the difference chi square, these two tests of significance were found to be highly consistent. Denoting the asymptotic standard error of λ_j by σ_λ, it was generally found across a variety of models and item sets that $(\lambda_j/\sigma_\lambda)^2$ closely approximated the corresponding difference chi square.

Due to the enormous sample size in this study, even very small latent classes were often statistically significant by these criteria. The final models chosen before proceeding to the second step were generally much simpler than the most complex models tried for each item set. In most cases, the final model chosen was one that had the fewest latent classes of any model yielding a nonsignificant chi square. One additional latent class beyond that minimum number was included if (a) the proportion of examinees in that additional class was greater than .1, or (b) the proportion in that additional class was at least .05 and its inclusion resulted in a difference chi square of at least 20 on one degree of freedom. An additional latent class satisfying one of these two conditions was included in the final models for six of the 18-item sets. These exceptions to the general rule of accepting the most parsimonious model were made because (a) excluding latent classes repre-

senting 10 percent or more of examinees could significantly distort the skill map to be constructed, and (b) very large reductions in the chi square provide strong evidence for the existence of additional latent classes. It should be emphasized that the final skill map does not depend critically upon the rules adopted at this stage. In analyses not reported here, alternative maps based upon somewhat different sets of final models were compared, and were found to differ only slightly. Reliance in this decision rule upon the numerical values of difference chi squares is an imperfect expedient. The magnitude of these statistics is a function of the sample size and the design effect adjustment, and they are properly interpreted only as significant versus nonsignificant. In another study with a different sample size or sample design, a value other than 20 might be more appropriate.

Model Revision

Each of the 18 final models yielded an estimate of π_{1i}, of π_{2i}, and of δ_i for each of its six items. All of these estimates from the 18 models were next sorted by item and compared. Under the assumptions of the binary skills model, each of these quantities for a given item should be invariant across item sets, except for sampling fluctuations and for the effects of pooling latent classes as described above. Thus, a finding that estimates of one or more of these quantities obtained using one item set differed sharply from those obtained using other item sets suggested that the model for the discrepant set should be revised. Such a revision is illustrated in Table 1 for a single item set, and is further discussed in the Appendix.

Multiple estimates were available for nearly all of the 37 items. Small variations among these estimates would be expected, due to sampling, but larger discrepancies were also found, indicating pooled classes. The problem was to distinguish which discrepancies were large enough to justify model revisions. If the multiple estimates of each parameter were statistically independent, their asymptotic standard errors would show how much variation should be expected due to sampling. In these analyses, however, errors in estimates were highly intercorrelated, and so their separate standard errors could not be used directly to judge the magnitude of discrepancies among estimates. Not only were all estimates obtained using the same examinee sample, but the degree of correlation between errors depended in complex ways upon the number of items two six-item sets had in common. Analysis of independent subsamples of examinees might have been attempted, but such an approach would have had to take into account the stratified cluster sampling design of the Anchor Test Study.

Rather than using asymptotic standard errors to detect pooling of latent classes, standard deviations were calculated for each item of all estimates of π_{1i}, of π_{2i}, and of δ_i. Unusually high standard deviations indicated discrepancies in need of examination. A stem-and-leaf of these within-item standard deviations is presented in Table 2.

Each of the three initial stem-and-leafs in Table 2 shows several clear outliers. The models giving rise to these discrepant estimates were examined, and where possible the discrepancies were eliminated by introducing additional latent

Table 1
Set F Model Revision to Resolve
Discrepant Estimates for Item 2

Original Latent Response Patterns						
Item: 02	15	23	37	38	43	Proportion (λ_j)
0	0	0	0	0	0	.356
0	0	0	1	0	1	.145
1	1	1	1	1	1	.499
χ^2 (49) = 36.14						
Revised Latent Response Patterns						
Item: 02	15	23	37	38	43	Proportion (λ_j)
0	0	0	0	0	0	.168
1	0	0	0	0	0	.188
0	0	0	1	0	1	.145
1	1	1	1	1	1	.499
χ^2 (49) = 36.14 (unchanged)						

Note. For the original model, estimated false positive and true positive probabilities for item 2 were .553 and .718, respectively. For the revised model, corresponding values were .454 and .718, as explained in the Appendix.

classes, following the computational procedure illustrated in the Appendix. Additional classes were introduced for one item at a time, beginning with the items producing the largest standard deviations shown in Table 2 and proceeding through items showing successively smaller standard deviations. The process ended when items were encountered for which (a) an appropriate model revision could not be determined, (b) the additional class created by a model revision would represent only two or three percent of the examinees, or (c) the class indicated by inspection of the π_{1i}, π_{2i}, and δ_i would have been identifiable in the initial analyses of the six-item set involved. Following this procedure, successive model revisions create smaller and smaller latent classes, until each revision has only a trivial effect. Thus, the precise stopping point for model revisions is not at all critical in determining the form of the final skill map. Introduction of additional classes had absolutely no effect upon fit statistics, degrees of freedom, or predicted proportions for each possible response pattern.

This procedure was carried out first using the standard deviations of the δ_i. Nineteen separate revisions indicated by large values of σ_δ also eliminated most of the outliers shown in the stem-and-leafs of within-item standard deviations for the π_{1i} and the π_{2i}, but the remaining outliers in these distributions were examined following the same procedure, and four more revisions were made. Taken together, these revisions affected 13 of the 18 six-item sets. Table 2 also shows stem-and-leafs of standard deviations of estimates following all model revisions. Although some possible outliers remain, the revised-model distributions are considerably tighter.

Table 2
Stem-and-Leafs of Within-Item Standard Deviations
Before and After Model Revisions

$\sigma_{P(\text{able to solve})}$		$\sigma_{\text{false positive}}$		$\sigma_{\text{true positive}}$	
Before	After	Before	After	Before	After
.15 | 0					
.14 | 5					
.13 |					
.12 | 9					
.11 |		1			
.10 | 6					
.09 | 138		1			
.08 | 234	3	9	9		
.07 | 02	02	345		8	
.06 | 588	8	017	0		
.05 | 245	1	3		02	2
.04 | 13	3	019	04	068	16
.03 | 0255779	022455779	355	333558	11234778	11137
.02 | 179	11237778999	357899	34778999	13469	113466889
.01 |	77	11123367	11123367779	11133569	111335569
.00 | 3467	346778	1489	14789	1233788	12337888

Note. The column headings $\sigma_{P(\text{able to solve})}$, $\sigma_{\text{false positive}}$, and $\sigma_{\text{true positive}}$ refer to the within-item standard deviations of estimates of ∂_i, π_{1i}, and π_{2i}, respectively.

Constructing a Skill Map for the MAT Items

After revising the initial models, they were integrated into a single skill map as follows. First, information was summarized about each pair of items occurring together in at least one set. Second, clusters were identified of items that appeared to require identical skills. Third, the skill requirements of these item clusters were related to one another to create a partial skill map. Finally, this skill map was elaborated to accommodate items not included in the clusters.

Summarizing information about item pairs. The six items of each set taken two at a time yielded 15 different item pairs, so that across the 18 item sets examined, there were 270 item pairings. For each of these, the latent response patterns and estimated latent class sizes were used to calculate estimated fractions of examinees able to solve neither item, only the first, only the second, or both items. The fraction able to solve neither item was always at least as large as the null class for the item set, and the fraction able to solve both was at least as large as the full class. If the fractions able to solve one item and not the other were both zero, then the two items were judged provisionally to require identical skills.

These paired relationships are depicted in Figure 1. Item numbers are listed across the top and down the left margins of the figure. In the body of the figure, "S" indicates a finding that two items require the same skill, "<" indicates that the row item requires fewer skills than the column item, ">" indicates the converse, and "*" indicates that each item requires skills the other does not. A blank indicates that the two items never occurred together in a six-item set.

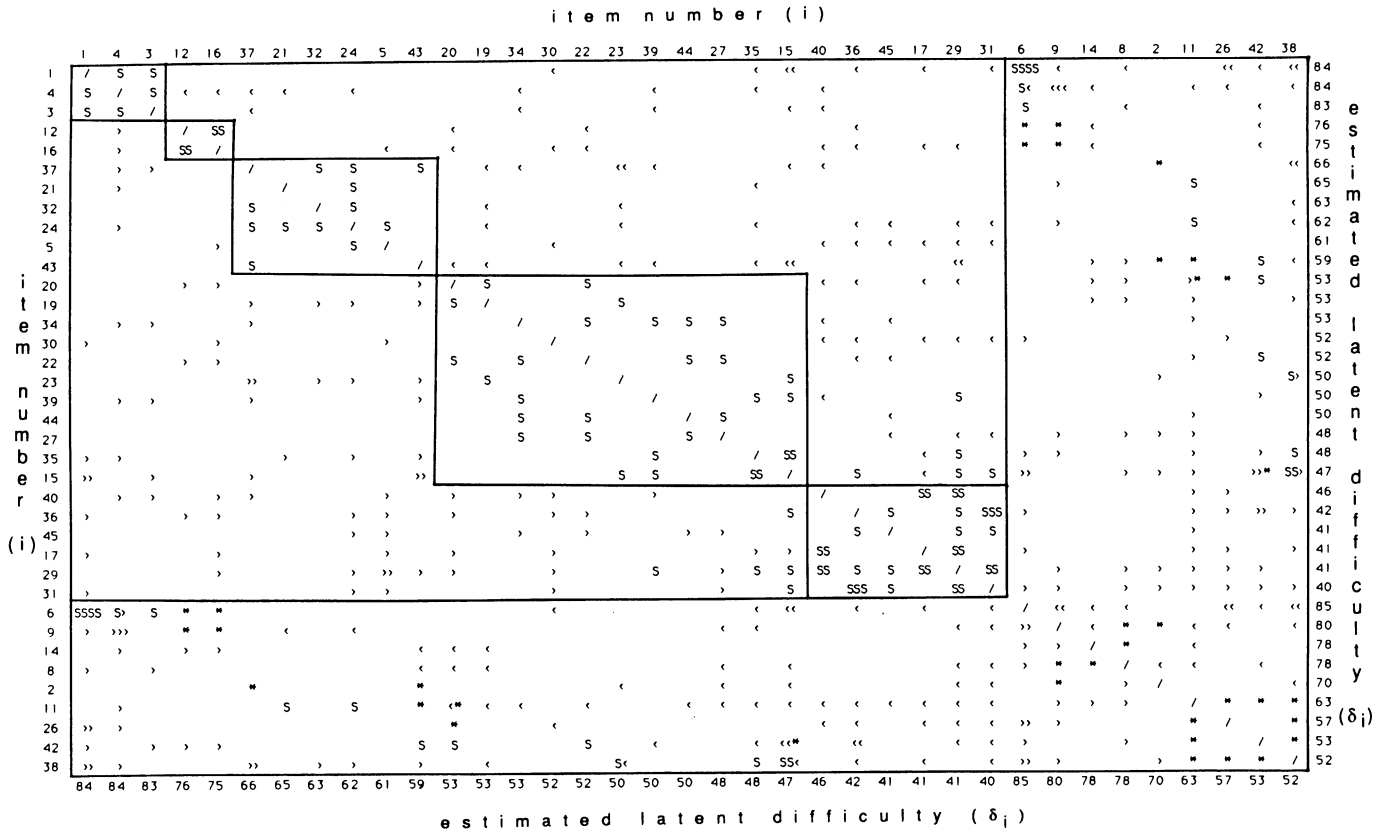

FIGURE 1. *Paired relationships among MAT items determined from multiple item sets*

The 270 item pairs represented 238 distinct two-item combinations. Of the combinations occurring in more than one set, less than 20 percent showed any inconsistencies across item sets. In these few cases, it was assumed that the discrepancy reflected omission of an actual latent class, rather than inclusion of a spurious class.

Identifying item clusters. To locate clusters of items requiring the same skills, the rows and columns of Figure 1 were sorted according to values of δ_i for the 37 items, and by inspection the largest possible clusters were located within which no pairwise relations other than "S" occurred. There were five such clusters, each containing from two to eleven items. In all, 28 of the 37 items analyzed were included in a cluster. These five item clusters are indicated in Figure 1 by rectangular boxes. The nine nonconforming items have been placed at the bottom and right of the figure.

The $\hat{\delta}_i$ are shown along the bottom and right-hand margins of Figure 1, and can be seen to be fairly homogeneous within clusters. Each of these values was obtained by averaging estimates of δ_i from all sets in which a given item was included. As would be expected, these values are less variable within clusters than are the manifest item p-values. For example, the $\hat{\delta}_i$ in one cluster range from .59 to .66, and the p-values of these same items range from .49 to .81. In another cluster, the $\hat{\delta}_i$ range from .47 to .53, and difficulties from .41 to .65. The correlation across the 37 items between δ_i and the manifest item difficult p_i is .88, but for some items the difference between these two statistics, representing net bias due to false positives and false negatives, is substantial.

Constructing a skill map for item clusters. It was straightforward to relate the five clusters to one another. As shown by Figure 1, they can be seen to form a scale, such that examinees able to solve items within a given cluster can also solve those in the preceding clusters. The $\hat{\delta}_i$ within clusters were averaged to obtain cluster-level estimates, and successive differences between these estimates were taken to find the proportion of examinees possessing the sets of skills defined by the items at each level. These are the *major classes* shown in Table 3. A sixth major class is shown for examinees unable to solve even those items in the lowest cluster of the scale.

Incorporating nonconforming items. To complete the skill map, the nine items not included in any of the clusters were examined. Rather than attempting to characterize the interrelationships among these items, each was related, independent of the others, to the scale defined by the six major classes. The relationship of each nonconforming item to as many of the successive major classes as possible was determined from its pairwise relationships to items in each cluster, tabulated earlier. Occasional inconsistencies were assumed to reflect omitted classes rather than spurious included classes. Finally, the column of Table 3 for each nonconforming item was calculated, giving the percentage of examinees in each major class able to solve that item. When there were no pairings of a nonconforming item with any items in some major class, it was sometimes still possible to infer the proportion of that major class able to solve the item. It might be, for example, that the overall proportion able to solve the item

Table 3
Percents of Examinees Possessing Skills Required by Different Item Sets

Major Class	Items	Percent of Sample	Percent of Major Class Able to Solve Each Nonconforming Item							
			6	9	14	8	2	11	26	42	38
Null	None	16	6	0	0	0	*	0	0	0	0
Matching	1,3,4	09	100	65	33	33	*	*	*	0	0
John & Juan	12,16	12	100	93	100	100	*	*	*	0	0
Sentence Comprehension	5,21,24,32,37,43	12	100	100	100	100	0	75	*	36	8
Inference	15,19,20,22,23,27, 30,34,35,39,44	09	100	100	100	100	100	87	58	74	100
Sophisticated Inference	17,29,31,36,40,45	42	100	100	100	100	100	100	100	100	100

Note. Examinees in each major class can also solve all items listed for the preceding classes. Asterisks indicate values that cannot be determined from the analyses completed.

was accounted for by examinees in other major classes who were able to solve it, so that the proportions for the remaining major classes had to be near zero. Cases in which such inferences could not be made are indicated by asterisks in Table 3.

The presence of the nonconforming items indicates that some of the major classes in fact represent a pooling of several smaller classes, distinguished by the presence or absence of additional skills that these nine items require. The Null major class, for example, is comprised of at least two classes, for examinees able to solve none of the items and examinees able to solve only item 6. It may also include separate classes for examinees able to solve only item 2, and for examinees able to solve both item 2 and item 6 but no others. The inference major class might represent up to eight separate classes. All of the examinees in this major class are able to solve conforming items through the inference level as well as six of the nonconforming items. They may or may not be able to solve each of items 11, 26, and 42. Since all examinees in the Sophisticated Inference class can solve all of the nonconforming items, this major class does represent only a single latent class.

It can be seen from Table 3 that the nonconforming items represent only minor departures from the linear scale defined by the major classes. Item 42, for example, is soluble by virtually the same examinees as can solve the items in the Inference cluster. The only examinees able to solve the Inference items and

unable to solve item 42 are 100 − 74 = 26 percent of those examinees at the Inference level. This level represents 9 percent of all the examinees, so 26 percent of them are about 2.3 percent of the total population. The only examinees able to solve item 42 and unable to solve the inference items are 36 percent of those at the sentence comprehension level, or about 4.3 percent of the total population. Thus, classification of item 42 in the Inference major class would misrepresent its solubility for only about 6.6 percent of the total population.

Interpretation of Skill Structure

The item clusters shown in Table 3 were derived statistically, without any consideration of the nature or content of the items themselves. It remained to examine the items in each cluster, looking for features they might share in common in order to clarify their interpretation. This was somewhat like interpreting the findings of an exploratory factor analysis.

The first cluster includes items 1, 3, and 4, all of which are based on the first of the eight passages on the MAT. They share the property that their correct response alternatives are words appearing in the passage, whereas none of their distractors appears in the passage. Other items based on that passage do not share this property. Since items 1, 3, and 4 can be solved by matching response alternatives to words in the passage, the major class for this cluster was labeled Matching. Of the first nine items on the test, 1, 3, and 4 are also the only ones for which choice "a" is correct. Thus, examinees marking the first choice for all items would get these three correct and surrounding items incorrect. Nonconforming item 6, which could be solved by nearly the same examinees as items 1, 3, and 4, is based on the second passage. It too could be solved by a matching strategy.

The next item cluster, labeled John and Juan, includes two items based on a paragraph about John helping Juan learn English. Item 12 tests recognition that the two were probably friends, and item 16 that Juan was probably grateful to John. The similar content of these items may account for their appearance as a cluster. Surrounding items tested comprehension of unrelated details.

These two items may also form a cluster because either can be answered by most examinees without having read the passage. In a study of passage dependency, Tuinman (1972, 1973) administered items from several reading comprehension tests, including the MAT, without their associated passages. In his sample of fourth graders, over 70 percent could answer each of items 12 and 16 correctly under this no-passage condition. Corresponding percentages for the other items based on the same passage ranged from 25 to 53 percent.

The third class is labeled Sentence Comprehension because its items each appear to require comprehension of only a single sentence or a pair adjacent of sentences. For items 5, 21, and 37, these are sentences in the passage. For items 24, 32, and 43, the sentence to be comprehended is the question stem itself. These latter items test propositions, that diseases are caused by germs, for example, which appear likely to be common knowledge among fourth graders. Tuinman (1972, 1973) found that 72 to 83 percent of fourth graders could answer items 24, 32, and 43 correctly under his no-passage condition, supporting the hypothesis that these items test general knowledge.

Items in the Inference cluster all require the integration of some background knowledge with the information presented in the passage. They called for interpretation of word meaning in context, inferences about when an event was likely to have occurred, or comprehension of propositions conveyed by subordinate clauses or single words. Item 35, for example, requires the examinee to infer from an incidental reference to "a seated cat," that the cat in a museum is sitting rather than black, gold, or in a coffin. The inference items had difficulties ranging from .14 to .57 under Tuinman's no-passage condition.

The Sophisticated Inference cluster included items of two kinds. Three of them, numbers 17, 29, and 40, each asked for the best name, main subject, or main idea of a passage, and included distractors likely to mislead students unable to distinguish major propositions from supporting details. The remaining three items, 31, 36, and 45, each required for their solution that three of the four alternatives be logically excluded, leaving the remaining choice as the only possible answer. Exclusion of the three distractors required precise interpretation the words "*not*" in item 31, "must be," "all" and "entirely" in item 36, and "every" in item 45.

Inspection of the nonconforming items yielded numerous hypotheses about their distinctive features, but additional items would have to be written and additional data collected to confirm or disconfirm any of these speculations. Items 2, 6, 8, 9, and 14 each appear to be soluble by idiosyncratic methods, whereas items 11, 26, 38, and 42 may present some unusual complexities that could thwart solution attempts by otherwise knowledgeable examinees.

Discussion

The items of the MAT reading comprehension subtest clearly do not all measure the same skills. They can be located at a series of skill levels that represent at least part of what is commonly regarded as reading comprehension, but some of the easier items can be solved in ways that do not imply any real understanding, and a few of the most difficult items appear to test elements of formal logic in addition to reading comprehension skills. At the highest skill level, about 42% of the examinees could solve all of the items analyzed. At the lowest three levels, 37% could solve no more than the five or so items yielding to the matching strategy or asking about John and Juan. The remaining 21% of the examinees fell at one of two intermediate levels. Inspection of the items after the skill map was constructed revealed striking similarities among the items at each level, and confirmed that items at different skill levels are difficult in different ways (Drum, Calfee, & Cook, 1980).

The linear skill structure established in these analyses is consistent with a plausible developmental sequence. Bearing in mind that cross-sectional data cannot establish the course of development, it nonetheless appears likely from these data that children progress from comprehension or matching of isolated words on these multiple-choice tests through the literal comprehension of sentences, then to inferences requiring the relation of passage content to general knowledge, and finally to comprehension of the passage as a whole.

Robustness of Procedures

In carrying out these analyses, some arbitrary decisions were required. It is appropriate to ask to what extent the final skill map or its interpretation would have differed if other item sets had been chosen for initial analyses, if the initial models for these sets had been chosen by slightly different criteria, if more or fewer model revisions had been made to resolve discrepancies, or if a much smaller sample had been used.

Perhaps the best way to resolve these questions would be through cross-validation. The examinee sample could be subdivided (randomly assigning primary sampling units within strata to different subsamples), and all of the steps from item set selection through skill map interpretation could be replicated independently. Such a complete, independent analysis would certainly increase confidence in the generalizability of the results presented, but even without full cross-validation, some assurances can be offered concerning the robustness of the findings.

While developing the skill mapping procedure, the effects of alternative selections of models in the first step of the procedure were investigated and found to be small. Likewise, precise rules appear unimportant in deciding how many model revisions should be made to resolve discrepancies in parameter estimates. The possible effects of choosing different item sets initially, perhaps choosing more or fewer sets, are more difficult to predict, but the inclusion of each item in an average of three different item sets appears to provide a reasonable basis for skill map construction.

Relation to Previous Analyses

In an earlier study applying latent class models to the same MAT data, Haertel (1984b) reported that 24 of these 37 items could be referenced to a single skill dichotomy. The skills required by the remaining items were not discussed. The earlier analyses and those of the present study used different latent class models to achieve different purposes. Given these differences, their findings are quite compatible.

One purpose of the earlier paper was to establish the applicability of two-latent-class models to a type of achievement commonly modeled as continuous. To that end, the largest possible set of items was located such that this simple model, applied to any six-item subset, would yield a nonsignificant chi square. This paper presents a systematic procedure for mapping the skills required by an entire collection of items examined, using a family of binary skills models of which the two-latent-class model is only the most elementary example.

Different latent class models may provide satisfactory fits to the same item response data, just as factor models with different numbers of factors may fit the same data. If a particular factor model, or a particular latent class model, is a true representation of the processes giving rise to the observations, then in principle no model with a different number of factors or classes can also be correct. In practice, however, just as several highly correlated factors can be modeled successfully as a single factor, so several latent classes representing very similar skill profiles can be modeled successfully as a single, larger class.

The 24 items found earlier to measure a common skill included the five most difficult of the nonconforming items shown in Table 3 and, with just four exceptions, all of the items at the Inference and Sophisticated Inference levels. These 24 items were characterized in the earlier study as difficult inference items, and it was estimated that 46% of the examinees could solve them and 54% could not. Table 3 shows that 42% of the examinees can solve all of these items at the Inference and Sophisticated Inference levels, 9% can solve most of them, and 49% can solve none. It also shows that the five most difficult of the nonconforming items can be solved by virtually the same examinees as the Inference items, as illustrated above for item 42. The four exceptions include three items from the Sentence Comprehension cluster that were included in the earlier set of 24, and one item from the Inference cluster that was omitted. In retrospect, it appears that all four of these items would have been classified differently in the earlier study if more item sets had been analyzed.

Summary and Conclusion

Further experience with binary skills models for different tests in different content areas would be very helpful in evaluating their strengths and weaknesses. In future work, it would seem better to select initial item sets following a systematic design, so that every pair of items occurred together in some minimum number of sets. Based on present experience, that minimum number should be at least three. Six-item sets appear to be a good size to work with. Estimation for six-item sets can easily be carried out on a personal computer, but such sets are large enough to provide ample degrees of freedom for testing goodness of fit.

Binary skills models may take their place along with the models of factor analysis, multidimensional scaling, and related procedures as tools for examining the internal structure of tests and for characterizing the individual differences they measure. Neither continuous nor discrete latent structure models can necessarily offer a complete picture of examinee abilities, but binary skills models applied along with continuous models may offer useful new perspectives on the determinants of test performance and the characteristics of test items.

APPENDIX

Model Revisions to Resolve Discrepancies

The resolution of discrepant estimates is illustrated for Item 2, which occurs in six-item sets F and R. The proportion able to solve item 2 is estimated in set F as .499 and in set R as .711, and the standard deviation of these two values is .150. As shown in Table 2, this standard deviation is clearly an outlier. There is also a large discrepancy between the estimates of the false positive probability for item 2: .553 according to set F versus .478 according to set R. The true positive estimates, .718 and .703, closely agree.

The fact that the false positive estimates are discrepant and the true positive estimates are close implies that in one of the sets the estimated proportion able to solve is depressed and the false positive rate is inflated. Discrepant true positive estimates would have implied instead that an estimate of the proportion able to solve was inflated and the corresponding true positive estimate was depressed. Inspection of the final models for sets F and R confirms this assessment. Set R includes three latent classes for which examinees can solve item 2, implying that π_{22} must be identified. Thus, the discrepancy cannot be

resolved by modifying set R. It can easily be resolved, however, by introducing an additional class into the model for set F, for examinees able to solve item 2 but none of the other five items. In other words, the original null class obtained for set F represents two pooled classes, with latent response patterns 000000 and 100000 (see Table 1).

The next step is to determine the proportion of examinees conforming to this additional latent class. Led $\tilde{\lambda}_2$ represent this proportion. Let $\delta_2 = .499$ represent the original proportion of examinees found able to solve item 2 using set F, and let $\tilde{\delta}_2$ represent the estimated proportion able to solve item 2 after the null class is divided into two classes. Note that $\tilde{\delta}_2 = \delta_2 + \tilde{\lambda}_2$. It has been explained that model revision will reduce the set F estimate of π_{12}. Denoting the original set F estimates of item 2's false positive and true positive probabilities by $\pi_{12} = .553$ and $\pi_{22} = .718$ and the revised estimate of the false positive probability by $\tilde{\pi}_{12}$, it is readily shown that

$$\tilde{\pi}_{12} = \pi_{12} - (\pi_{22} - \pi_{12})\,\tilde{\lambda}_2/(1 - \delta_2 - \tilde{\lambda}_2) = \pi_{12} - (\pi_{22} - \pi_{12})\,\tilde{\lambda}_2/(1 - \tilde{\delta}_2). \quad (1)$$

Thus, the value chosen for $\tilde{\lambda}_2$ determines the value of $\tilde{\pi}_{12}$. Ideally, a value of $\tilde{\lambda}_2$ could be found that would bring estimates of both the proportion able to solve and the false positive probability into agreement with the estimates obtained using all other item sets. This objective can be approximated as follows. Let target values for $\tilde{\delta}_2$ and for $\tilde{\pi}_{12}$ be denoted $\bar{\delta}_2$ and $\bar{\pi}_{12}$, and choose a value of $\tilde{\lambda}_2$ such that

$$\tilde{\delta}_2 - \bar{\delta}_2 = \tilde{\pi}_{12} - \bar{\pi}_{12}. \quad (2)$$

The target values $\bar{\delta}_2$ and $\bar{\pi}_{12}$ are the means of the corresponding estimates obtained using all other item sets that include item 2. There is only one other set, and so the target values are simply the estimates for set R, $\bar{\delta}_2 = .711$ and $\bar{\pi}_{12} = .478$. A direct solution for the value of $\tilde{\lambda}_2$ satisfying Equation 2 is cumbersome, but the required numerical value is readily obtained using Newton's method. The iterative equation required is

$$\tilde{\lambda}_2^{(t+1)} = \tilde{\lambda}_2^{(t)} - (\bar{\pi}_{12} - \tilde{\pi}_{12}^{(t)} - \bar{\delta}_2 + \delta_2 + \tilde{\lambda}_2^{(t)})/[1 + (\pi_{22} - \pi_{12})(1 - \delta_2)/(1 - \delta_2 - \tilde{\lambda}_2^{(t)})^2],$$

where $\tilde{\pi}_{12}^{(t)}$ is calculated from $\tilde{\lambda}_2^{(t)}$ according to Equation 1. Beginning with the value $\tilde{\lambda}_2^{(0)} = 0$, the desired solution is obtained in two or three iterations. In this example, the solution is $\tilde{\lambda}_2 = .18794$, $\tilde{\pi}_{12} = .45394$, which makes each side of Equation 2 equal to $-.02406$.

This illustration shows the solution to the identification problem in which a class is introduced for examinees able to solve only one item. There is a corresponding solution for the case in which some examinees can solve all except one item, which affects estimates of that item's true positive probability and of the proportion conforming to the full latent class.

Occasionally, additional classes for two or three different items must be introduced into the model for the same item set. This requires that classes also be introduced for examinees able to solve all possible combinations of these respective items. It can be shown that if such a model is to be mathematically equivalent to the original model, then the distinct binary skills associated with each of these items must be uncorrelated. This fact can be used to solve for the proportions of examinees able to solve each combination of the items. Suppose, for example, that classes for examinees able to solve only item x and only item y are both indicated for some item set S. This implies that the null class for set S must be divided into four classes: a smaller null class, and additional classes for examinees able to solve only item x, only y, and both x and y but no other items. Suppose that separate analyses like the one illustrated above, each beginning with the original set S model, yield estimates $\tilde{\lambda}_x$ and $\tilde{\lambda}_y$ for the respective proportions able to solve item x only and item y only.

Then the final proportion able to solve only item x would be $\tilde{\lambda}_x - \tilde{\lambda}_x\tilde{\lambda}_y$, the proportion able to solve only item y would be $\tilde{\lambda}_y - \tilde{\lambda}_x\tilde{\lambda}_y$, and the proportion able to solve items x and y but no others would be $\tilde{\lambda}_x\tilde{\lambda}_y$. The null latent class proportion would be reduced by $\tilde{\lambda}_x + \tilde{\lambda}_y - \tilde{\lambda}_x\tilde{\lambda}_y$. Note that if a fourfold table were constructed crossing possession versus nonpossession of the item x binary skill with possession versus nonpossession of the item y binary skill, the covariance in that fourfold table would be zero.

The extension to more than two items is immediate. With three items, x, y, and z, the final proportions able to solve only x; only x and y; and only x, y, and z would be $\tilde{\lambda}_x - \tilde{\lambda}_x\tilde{\lambda}_y - \tilde{\lambda}_x\tilde{\lambda}_z + \tilde{\lambda}_x\tilde{\lambda}_y\tilde{\lambda}_z$, $\tilde{\lambda}_x\tilde{\lambda}_y - \tilde{\lambda}_x\tilde{\lambda}_y\tilde{\lambda}_z$, and $\tilde{\lambda}_x\tilde{\lambda}_y\tilde{\lambda}_z$, respectively. Interactions of this kind among separate revisions to the same item set occurred for six of the 18-item sets. Note that there is no such interaction if one revision affects a false positive probability and the null class and another affects a true positive probability and the full class.

References

Aitkin, M., Anderson, D., & Hinde, J. (1981). Statistical modeling of data on teaching styles (with discussion). *Journal of the Royal Statistical Society-Series A, 144,* 419–461.

Bartholomew, D. J. (1980). Factor analysis for categorical data. *Journal of the Royal Statistical Society-Series B, 42,* 293–312.

Bergan, J. R. (1983). Latent-class models in educational research. In E. W. Gordon (Ed.), *Review of Research in Education, 2,* (pp. 305–360). Washington, DC: American Educational Research Association.

Bock, R. D., & Jones, L. V. (1968). *The measurement and prediction of judgment and choice.* San Francisco: Holden-Day.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika, 40,* 5–32.

Clogg, C. C. (1977). *Unrestricted and restricted maximum likelihood latent structure analysis: A manual for users* (Working Paper No. 1977–09). University Park, PA: Pennsylvania State University, Population Issues Research Office.

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, Series B, 39,* 1–22.

Drum, P. A., Calfee, R. C., & Cook, L. K. (1980). The effects of surface structure variables on performance on reading comprehension tests. *Reading Research Quarterly, 16,* 486–513.

Durost, W. N., Bixler, H. H., Wrightstone, J. W., Prescott, G. A., & Balow, I. H. (1970). *Metropolitan Achievement Tests, Form F* (Elementary level). New York: Harcourt, Brace, Jovanovich.

Everitt, B. S. (1984). A note on parameter estimation for Lazarsfeld's latent structure model using the EM algorithm. *Multivariate Behavioral Research, 19,* 79–89.

Everitt, B. S., & Hand, D. J. (1981). *Finite mixture distributions.* London: Chapman and Hall.

Goodman, L. A. (1974). The analysis of qualitative variables when some of the variables are unobservable. Part I—a modified latent structure approach. *American Journal of Sociology, 79,* 1179–1259.

Goodman, L. A. (1975). A new model for scaling response patterns: An application of the quasi-independence concept. *Journal of the American Statistical Association, 70,* 755–768.

Goodman, L. A. (1979). On the estimation of parameters in latent structure analysis. *Psychometrika, 44,* 123–128.

Gustafsson, J- E. (1980). A solution of the conditional estimation problem for long tests in the Rasch model for dichotomous items. *Educational and Psychological Measurement, 40,* 377–383.

Haertel, E. H. (1984a). An application of latent class models to assessment data. *Applied Psychological Measurement, 8,* 333–346.

Haertel, E. H. (1984b). Detection of a skill dichotomy using standardized achievement test items. *Journal of Educational Measurement, 21,* 59–72.

Haertel, E. H., Korpi, M., & Capell, F. J. (1982, April). *Detection of distinct skills required by reading comprehension test items.* Paper presented at the meeting of the American Educational Research Association, New York.

Harris, C. W., & Pearlman, A. P. (1978). An index for a domain of completion or short answer items. *Journal of Educational Statistics, 3,* 285–303.

Hartigan, J. A. (1975). *Clustering algorithms.* New York: John Wiley & Sons.

Kish, L. (1967). *Survey sampling.* New York: John Wiley & Sons.

Lazarsfeld, P. F. (1950a). The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen, *Studies in Social Psychology in World War II: Vol. 4. Measurement and prediction* (pp. 362–412). Princeton, NJ: Princeton University Press.

Lazarsfeld, P. F. (1950b). Some latent structures. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen, *Studies in Social Psychology in World War II: Vol. 4. Measurement and prediction* (pp. 413–472). Princeton, NJ: Princeton University Press.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent Structure Analysis.* Boston: Houghton Mifflin.

Loret, P. G., Seder, A., Bianchini, J. C., & Vale, C. A. (1974). *Anchor test study—The equating and norming of selected reading achievement tests (grades 4, 5, and 6).* Washington, DC: U.S. Department of Health, Education and Welfare, Office of Education.

Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics, 2,* 99–120.

Macready, G. B., & Dayton, C. M. (1980). The nature and use of state mastery models. *Applied Psychological Measurement, 4,* 493–516.

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

Molenaar, I. W. (1983). Some improved diagnostics for failure of the Rasch model. *Psychometrika, 48,* 49–72.

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

Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York: John Wiley & Sons.

Thurstone, L. L. (1927). Psychophysical analysis. *American Journal of Psychology, 38,* 368–389. (Reprinted in L. L. Thurstone, *The measurement of values,* pp. 19–38. Chicago: University of Chicago Press, 1959).

Traub, R. E., & Lam, Y. R. (1985). Latent structure and item sampling models for testing. In M. R. Rosenzweig & L. W. Porter (Eds.), *Annual review of psychology* (Vol. 36, pp. 19–48). Palo Alto, CA: Annual Reviews, Inc.

Tuinman, J. J. (1972). *Inspection of reading comprehension passages as a function of passage dependency of test items.* Bloomington, IN: Indiana University, Institute for Child Study.

Tuinman, J. J. (1973). Determining the passage dependency of comprehension questions in 5 major tests. *Reading Research Quarterly, 9,* 206–223.

Author

EDWARD H. HAERTEL, Associate Professor of Education, Stanford University School of Education, Stanford, CA 94305-3096. *Degrees:* BA (honors), Mathematics, University of Wisconsin–Madison, 1971; PhD, Education, University of Chicago, 1980. *Specializations:* statistical models for achievement tests, testing and educational policy; large-scale educational assessments, criterion-referenced measurement.