An Application of Latent Class Models to Assessment Data

Edward Haertel
Stanford University

Responses of 17-year-olds to selected 1977–78 National Assessment of Educational Progress (NAEP) mathematics exercises were analyzed, using latent class models. A single model was fitted to data from five independent samples of examinees, each of which responded to a different set of six algebra or prealgebra exercises. Four categories of items were found, defining five levels of content mastery, ranging from examinees unable to solve any of the exercises (43%) through those able to solve all the exercises (19%). The methods demonstrated are broadly applicable to assessment data, including matrix-sampled data, and provide an aggregate description of examinee abilities independent of the specific characteristics of individual exercises administered.

The National Assessment of Educational Progress (NAEP) affords a rich resource for the description of the academic skills of American young people. Careful deliberation as to appropriate objectives, to the matching of item formats to objectives, and to the use of all technically sound exercises regardless of their item statistics yield item pools that tap an unparalleled array of specific competencies. Matrix sampling permits the administration of hundreds of exercises to nationally representative respondent samples without unduly burdening any individual pupils or schools.

Researchers seeking to exploit these data may be overwhelmed by the embarrassment of riches. Accustomed to conceiving items as multiple indicators of a single ability, many are ill-equipped to draw useful conclusions from hundreds of diverse items, each of interest in its own right. The most typical response has been to build scales using items that appear to be related, and restricting attention to one exercise booklet (package) at a time.

Those turning to NAEP publications of findings have found little additional guidance. In these publications, results are aggregated over booklets, but exercises are pooled into broad "report topics" categories spanning many objectives, simply because the separate objectives are so numerous. For descriptions of absolute levels of performance the reader is told, "30% of nine-year-olds know this . . .," "45% can solve this problem . . .," and so forth. The reader then is left to imagine what generalizations from these statements are appropriate. Only 7% of 17-year-olds could correctly solve the equation "$(x - 2)^2 = 9$" for x, but in another sample, 18% could "Find the solution set of $x^2 - 5x + 6 = 9$." (Exercises S0429A and S0905A from the 1977–78 mathematics assessment, March–May 1978.) Is the

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 8, No. 3, Summer 1984, pp. 333–346

0146-6216/84/030333-14$1.95

difference due to the wording of the problems? The particular numbers? The format of the equations? How is the proportion of 17-year-olds who can solve quadratic equations to be generalized? How would the proportion correct change if these items were, say, multiple-choice rather than free response?

This study investigated the use of certain restricted latent class models to address such issues, and to abstract generalizations about examinee abilities from matrix-sampled data. The goals of the study were to determine empirically (1) what and how many dichotomous skills need be assumed to account for patterns of examinee performance, (2) whether these skills formed a Guttman scale, and (3) the prevalence of each skill combination in the population. Substantive interpretations of assessment data are derived, independent of the format or other features of any particular items.

Latent Class Model

Latent class models similar to those used in this study were first investigated by Lazarsfeld and Henry (1968), and applied to item response data by Proctor (1970), Dayton and Macready (1976), Goodman (1975), Haertel (1984), and others. The basic unit of ability assumed in these models may be termed a *skill*. In this study, skills were assumed to be intermediate in scope between NAEP objectives and report topics, for example, "solving quadratic equations" or "making conversions among common units of measurement." Skills were assumed to be dichotomous, and to be determined by curricular organization as well as basic psychological processes. No attempt was made in this study to disentangle sources of skill distinctions. Skill possession cannot be observed directly, but each permissible pattern of presence and absence of skills defines a latent class. Each examinee conforms to exactly one latent class, and item responses are assumed conditionally independent given latent class. Thus, latent class membership completely determines the probabilities of correct responses to each item.

The population distribution of examinee ability is given by the vector $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k)$, which gives the proportion of examinees in each of the k latent classes. Note that $\Sigma\lambda_j = 1$. Several latent class models for item response data are reviewed by Macready and Dayton (1980). The models considered in this study are defined by a different set of parameter restrictions, and appear not to have been used before. These models permit more than two latent classes, but for each item i, correct response probabilities must assume one of just two values, ϕ_i or τ_i. These constraints permit the definition of latent response patterns for each latent class, as described below, which may or may not form a Guttman scale.

Each item is assumed to require some subset of the dichotomous skills. On the basis of the skills required by a given item i, the latent classes may be partitioned into two sets. In one set are latent classes for examinees unable to solve item i, that is, lacking one or more of the skills the item requires; in the other are latent classes for examinees possessing all the requisite skills and therefore able to solve the item. For latent classes in which all the requisite skills are present, the correct response probability is the item's true positive rate, τ_i. For classes in which at least one requisite skill is absent, the correct response probability is the item's false positive rate, ϕ_i.

The skill pattern for each latent class and the skill requirements for each item together define a *latent response pattern* for each latent class, $\mathbf{v}_j = (v_{j1}, \ldots, v_{jn})$, where v_{ji} is 1 if examinees in class j can solve item i, else it is zero. This is the pattern of manifest responses to the n items that would be observed for examinees in the jth class if no misclassifications occurred.

If the latent response patterns for all classes form a Guttman scale, then the set of dichotomous skills may be interpreted as defining a single scale with more than two discrete levels of content mastery. A scale of this kind was derived empirically for a broad domain of NAEP mathematics items. Four categories of NAEP exercises are described, defining five levels of content mastery. Estimates of the proportions of 17-year-olds at each level are constant (within sampling error) across five distinct sets of exercises

administered to independent samples of examinees. These proportions characterize the distribution of algebra skills in that population.

Because item responses are assumed to be conditionally independent given latent class, the conditional probability of any pattern of correct and incorrect responses given that the examinee is in the jth latent class is just the product of the conditional probabilities of their separate occurrences. Thus, it is possible to express the unconditional probability of any observable pattern of correct and incorrect item responses in terms of just $2n + k - 1$ parameters: ϕ_i and τ_i, $i = 1, \ldots, n$, and λ_j, $j = 1, \ldots, k - 1$. Let $\mathbf{u}_h = (u_{h1}, \ldots, u_{hn})$ be one of the 2^n observed patterns of correct and incorrect responses, where u_{hi} is 0 (incorrect) or 1 (correct). Then

$$P(\mathbf{u}_h \mid \boldsymbol{\lambda}, \boldsymbol{\phi}, \boldsymbol{\tau}) = \sum_{j=1}^{k} \lambda_j \prod_{i=1}^{n} \left[\phi_i^{(1-v_{ji})u_{hi}} (1-\phi_i)^{(1-v_{ji})(1-u_{hi})} \tau_i^{v_{ji}u_{hi}} (1-\tau_i)^{v_{ji}(1-u_{hi})} \right] . \quad (1)$$

Note that $(1-\phi_i)$ and $(1-\tau_i)$ are the true negative and false negative probabilities for item i, respectively.

Two situations are encountered in which these latent class models are not fully identified.[1] First, if there are two latent classes j and j' such that $\mathbf{v}_j = \mathbf{v}_{j'}$, then λ_j and $\lambda_{j'}$ are not identified. Their sum is identified, however. The nonidentified latent classes are pooled, and a model with $k' < k$ classes is estimated. The estimate of λ for the pooled classes is interpreted as the sum of the nonidentified parameters.

The second type of identification problem involves two latent class parameters and an item parameter, and arises whenever the lowest or highest skill level represented by any items in a set is represented by only a single item. Suppose $\mathbf{v}_1 = \mathbf{0}$, and $\mathbf{v}_2$ is the same except that $v_{2i} = 1$. It is then impossible to distinguish between an examinee unable to solve any of the items (i.e., conforming to λ_1) who gives a false positive response to item i, and an examinee able to solve only the ith item (i.e., conforming to λ_2) who gives a true positive response to that item. The parameters λ_1, λ_2, and ϕ_i are nonidentified, as can be verified by considering the following substitutions:

$$\begin{aligned} \lambda_1^* &= \lambda_1 + c \\ \lambda_2^* &= \lambda_2 - c \\ \phi_i^* &= (\lambda_1\phi_i + c\tau_i)/(\lambda_1 + c) \quad . \end{aligned} \quad (2)$$

Inspection of Equation 1 will reveal that if λ_1, λ_2, and ϕ_i are replaced with λ_1^*, λ_2^*, and ϕ_i^*, the probabilities of all manifest response patterns remain unchanged.

The procedure followed in this case is to impose the arbitrary constraint $\lambda_2 = 0$. Estimation then proceeds just as if there were $k-1$ latent classes. The estimate of λ_1 is interpreted as an estimate of $\lambda_1 + \lambda_2$, and in addition, the estimate of ϕ_i is recognized to be inflated to a degree that depends on the magnitude of λ_2. An exactly analogous effect leads to a depressed estimate of τ_i (i.e., an inflated false negative rate, $1-\tau_i$), when the highest skill level in a set of items is represented only by item i. These distorted item parameter estimates can be corrected by using additional information from analyses of other item sets to estimate the magnitude of the nonidentified parameters and then solving for the correct value of ϕ_i or τ_i.

Estimation was by the method of maximum likelihood (Rao, 1973). With samples in excess of 1,000 examinees for each booklet, the likelihood ratio chi-square provides a sensitive omnibus test of model fit. Hypotheses of the form $\lambda_j = 0$ also are readily tested by difference chi-squares contrasting nested models. The distributions of these statistics are derived assuming simple random sampling. Because NAEP data are obtained from deeply stratified cluster samples, likelihood ratio and difference chi-squares will

[1]Additional degenerate cases in which $\phi_i = \tau_i$ for some i or where one or more of the $\lambda_j = 0$ do not arise in practice, and are not discussed. Also omitted from this discussion are cases where k is too large relative to n.

be inflated. An approximate adjustment for the effects of the sampling plan is given by an inflation factor known as the design effect, introduced by Kish (1967). For the NAEP data, the design effect for item p values is estimated to be roughly two (Folsom, 1977). Thus, an estimate from the NAEP data of the proportion of examinees manifesting a given response pattern should be equivalent in precision to an estimate obtained from a simple random sample roughly half as large.

Each examinee's data were weighted (inversely to the net probability of that examinee's selection), using the calculated weights provided on NAEP public-use tapes. Likelihood ratio and difference chi-squares were calculated assuming an effective sample size equal to one-half the actual sample size. For all booklets, this effective sample size was over 1,000. The effects of weighting on the obtained parameter estimates were small, and the design effect adjustment had no effect whatsoever on parameter estimates, serving only to reduce the values of chi-squares by a factor of 2 and to increase estimated standard errors by a factor of $\sqrt{2}$. Because the design effect correction is approximate, rigid adherence to precise significance levels in the interpretation of fit and test statistics is not appropriate.

Actual calculations were carried out using MLMN, a general purpose Fortran program for maximum likelihood estimation, which is not generally available.[2] However, identical results can be obtained using the MLLSA program (Clogg, 1977), which employs an efficient iterative procedure by Goodman (1974). MLMN is substantially more expensive to run than MLLSA, but provides asymptotic standard errors of estimates, which are not available from the present version of MLLSA.

Procedure

Latent class modeling of algebra and prealgebra skills required first the definition of an item domain and the selection of exercises (items) to be modeled. Separate analyses were then carried out for each distinct sample of items, from different booklets, answered by independent samples of examinees. After fitting latent class models to each item set, the final step was to fit a single comprehensive model to data from the separate samples.

Selection of Exercise Sets

It was decided for this study to use sets of just six items from each booklet, providing 64 correct/incorrect response patterns and assuring at least 45 degrees of freedom for testing goodness of fit. Five items would yield at most 20 degrees of freedom for testing model fit, while seven items would require fitting 128 response patterns. By taking only one 6-item set from any given exercise booklet, statistical independence was assured, since each examinee responded to only a single booklet. This simplified the testing of the final composite model.

NAEP mathematics exercises are each designed to measure a specific objective. Objectives in turn are classified into content areas (numbers and numeration; variables and relationships; shape, size, and position, etc.) as well as process categories (knowledge, skill, understanding, application). Process categories are divided into subtopics (recall facts, translate statements, routine problems, etc.), and some subtopics are subdivided still further. Independent of this hierarchical scheme, other classifications such as "consumer problem" are also imposed.

The goal of these analyses was to cut across the many specific objectives sampled in the age 17 assessment, not by retreating to higher taxonomic levels, but by determining empirically the number of

[2]MLMN was written by Richard Wolfe, presently at the Ontario Institute for Studies in Education. It was enhanced slightly by the present author for use in this study.

skill patterns (latent classes) required to account for performance in some broad domain. It initially appeared too ambitious to consider all 443 exercises administered at age 17, so attention was restricted to exercises that appeared to require an understanding that letters could represent variable quantities in mathematical statements and expressions. Candidate items were first identified by scanning the brief "exercise texts" in the documentation files provided with the NAEP public-use tapes, and the full texts of those that appeared usable were then reviewed on microfiche. Only the "A" parts of multipart exercises were used to avoid spurious statistical dependencies. The 75 exercises initially identified included some classified as algebraic manipulation (solving equations, simplifying and factoring, plotting, graphs), mathematical skills and computation, numbers and numeration, understanding, translation, and other topics. When full texts were examined, roughly 20 were rejected for not relying on the common skill. This left just 6 of the 12 Age 17 booklets with at least six acceptable exercises, Booklets 1, 2, 3, 4, 7, and 9. Six exercises were drawn from each of these booklets for analysis.[3]

Within-Booklet Analyses

The first analysis for each booklet was to fit the simplest possible model, with only two latent classes. These may be referred to as the "null" and "full" classes, and were included in all subsequent runs as well. Their latent response patterns are "000000" and "111111," respectively. This initial run yielded a nonsignificant chi-square for only one of the six booklets (Booklet 1: $\chi^2 = 59.3$, $df = 50$). For the remaining five booklets, chi-squares for the initial run ranged from 72.5 to 132.5.

Following this initial run, the residuals (observed minus fitted proportions for each of the 64 response patterns) were examined to determine what additional models should be tried. Any such additional model would include the null and full latent classes plus one or more additional classes with latent response patterns containing from 2 zeros and 4 ones to 4 zeros and 2 ones. (Classes with response patterns containing a single one or a single zero would not be identified.) There are $\binom{6}{2} + \binom{6}{3} + \binom{6}{4}$ or 50 possible three-class models, and thousands of models with four or more latent classes. Therefore, some systematic procedure is essential to guide the exploration of more complex models. If the current model fits, residuals are expected to be small and nonsystematic. Otherwise, systematic patterns in the residuals usually offer clues to more complex models that might fit better. Simple inspection of the residuals for all 64 response patterns is generally not informative, but the following procedure has proven useful.

Each latent class that might be introduced corresponds to a subset of items that some examinees can solve and the remainder cannot, namely, the items with "1" in its latent response pattern. Residuals are summed over just the response patterns in which all these items are answered correctly. If the sum is relatively large and positive, it suggests that correct responses to all of these items occurred together more frequently than the current model permits and that the new latent class may be needed. When that latent class is introduced, the fit is generally improved. To facilitate selection of subsequent runs, this calculation is done routinely after each run for all possible additional classes.

The new model is compared to the previous model by calculating a difference chi-square, and if it is statistically significant, the previous model is discarded in favor of the new, more complex model. Typically, several three-class models, several four-class models, and perhaps one or more five-class models were tried for a given item set. Although difference chi-squares cannot be calculated to compare two models with the same number of latent classes, among those models for a given item set with the same number of latent classes, the one with the lowest chi-square was considered to fit best. Generally,

[3]The exercise numbers, item difficulties, content, and format of each exercise analyzed are available upon request from the author.

Table 1
Best-Fitting Models with Two to Four Latent Classes

Booklet	Number of Latent Classes	Chi Square	df	Latent Response Patterns[a]
1	2	59.29	50	
	3[b]	49.91	49	010111
2	2	84.20	50	
	3[b]	66.56	49	100111
	4	62.19	48	100110, 100111
3	2	101.09	50	
	3[b]	66.18	49	101100
	4	52.37	48	101100, 101101
4	2	132.49	50	
	3	77.01	49	011101
	4[b]	63.08	48	010101, 011101
7	2	72.48	50	
	3[b]	62.08	49	000110
9	2	102.24	50	
	3	86.71	49	010011
	4	74.34	48	010011, 001111

[a] In addition to latent response patterns shown, all models included classes associated with the patterns 000000 (null) and 111111 (full).

[b] Model included in final, cross-booklet analysis.

this best-fitting model was markedly superior to the next best; there was little ambiguity in the model selection for individual booklets.

Results of these analyses are shown in Table 1. For Booklets 1, 2, 3, 4, and 7, latent response patterns for all best-fitting models formed Guttman scales—consistent with the hypothesis that the various skills the items required formed a scale defining several distinct levels of content mastery. Alternative four-class models that did not form scales were tried and did not fit as well. For Booklet 9, however, the latent response patterns for the best-fitting four-class model did not form a Guttman scale. According to this model, some examinees could solve the third and fourth items but not the second, while others could solve the second but not the third and fourth. Moreover, Booklet 9 was the only booklet for which no model could be found that gave a nonsignificant chi-square; numerous models with up to seven latent classes were considered. When the second exercise was removed from this six-item set, a satisfactory fit to the remaining five items was obtained with a four-class model ($\chi^2 = 19.45$, $df = 18$), but the latent classes for this model again failed to form a scale (0_0000, 0_1011, 0_0101, 1_1111).

No post hoc explanation for these anomalous Booklet 9 results may be taken as definitive, but inspection of the six exercises from Booklet 9 suggests a possible explanation. Several exercises in this set are nonstandard problems that may depend fairly strongly on general mental ability, especially fluid-analytic ability, as well as instruction in algebra. Although isolated exercises in the other sets are of this type, only Booklet 9 contains two or more such problems. Thus, the Booklet 9 exercise set may be measuring a distinct reasoning ability in addition to the crystallized ability imparted through instruction. Items that are not unidimensional would not be expected to conform to a Guttman scale pattern. Given the failure to fit Booklet 9 satisfactorily with any model, hierarchical or otherwise, only the remaining five booklets were included in the subsequent analysis.

Analysis Across Booklets

It was an empirical finding of the within-booklet analyses that linear scales best described performance on each of the first five booklets. Two- and three-class models necessarily form such Guttman scales, but four-class models may not. This finding led to a search for a single comprehensive scale underlying performance across samples of exercises and of examinees.

At least one exercise set, Booklet 4, required the inclusion of four latent classes to obtain a satisfactory fit. Thus, a comprehensive model would require a minimum of four classes. More than four classes might be required if no single six-item set sampled all of the skill levels in the domain, or if the lowest or highest skill level sampled within a set was represented by only a single item. In either case, two or more latent classes would be represented by a larger pooled latent class in that six-item set, as explained above.

Discovery of the simplest possible comprehensive model proceeded as follows. Only "best-fitting" models (Table 1) that yielded nonsignificant chi-squares were considered. For each of Booklets 4 and 7, there was only one such model. For each of Booklets 1, 2, and 3, there were two models from which to choose. All of these models are shown graphically in Figure 1. Each bar in this figure shows the cumulative proportions of examinees in the identified classes (or pooled classes) for one within-booklet model. Also shown are the final combined model and the discrepancies between it and the separate models for each booklet. In selecting models for Booklets 1, 2, and 3, the two possible choices for each booklet were first compared to one another. In each case, the more complex of the two closely resembled the simpler model, except that one class was split in two. For Booklets 2 and 3, this splitting resulted in one latent class characterizing less than 6% of the population. Even though these were statistically significant (as shown by difference chi-squares), they were judged to be of little importance, especially since they were not replicated across booklets. For Booklet 1, the additional class emerging in the more complex model included 15% of the population, and corresponded closely to a class found using Booklets 3, 4, and 7. Thus, the simpler (three-class) models were selected for Booklets 2 and 3, and the more complex (also three-class) model was chosen for Booklet 1. It should be noted that either the two-class or the three-class model for Booklet 1 is consistent with the final model fitted across booklets.

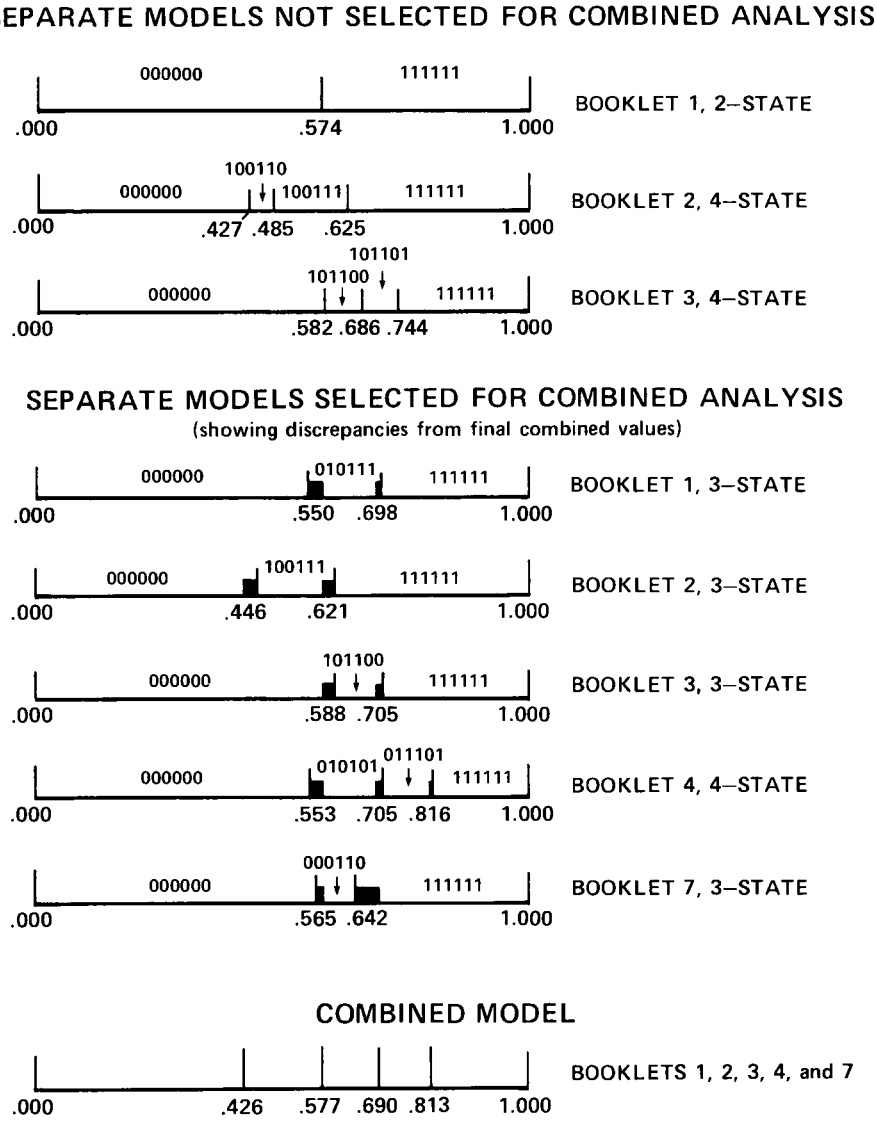
Inspection of Figure 1 suggested five latent classes common to the separate booklets. The distinction between Classes 1 and 2 was shown by Booklet 2 only. The distinction between Classes 2 and 3 emerged in all five booklets, that between Classes 3 and 4 in all but Booklet 2, and that between 4 and 5 in Booklet 4 only. The next step was to confirm that the separate models were consistent both empirically and substantively with such a pooled model.

The ideal method for fitting a pooled model would be to reestimate all item parameters and latent class proportions simultaneously by the method of maximum likelihood. This would involve the estimation of 5 (booklets) $\times$ 6 (items per booklet) $\times$ 2 (parameters per item) or 60 item parameters plus 4 common latent class parameters. No software was available to accomplish such a single joint estimation. The two-stage procedure used instead was (1) to estimate the common latent class parameters by a weighted least squares method, and (2) to reestimate item parameters within each of the five booklets separately by the method of maximum likelihood, constraining the latent class parameters to the established values. The first-stage analysis was accomplished using standard package software, as described below. The second stage employed Clogg's (1977) MLLSA program. Results of these estimations for all five booklets are shown in Table 2, which also presents latent response patterns for each latent class. Note that the unconstrained estimations sometimes produced estimates only for pooled sets of latent classes.

For each of the five booklets, the constrained estimation resulted in a larger likelihood ratio chi-square than the original unconstrained estimation. For a significance test of the constraints imposed across booklets, increments in the chi-squares were pooled across booklets.

The logic and results of this procedure were as follows. If the common latent class parameters were

Figure 1
Cumulative Latent Class Proportions
for Models of Separate Booklets and for Combined Model



determined independent of a given booklet, the difference of the likelihood ratio chi-squares for the constrained and unconstrained estimations could be tested against a chi-square distribution with $k - 1$ degrees of freedom. Tests of these separate difference chi-squares are not appropriate because data from all five booklets were used to establish the common latent class parameter values.

Across all five booklets, 11 separate latent class parameters were estimated. In the constrained runs, just four parameters were estimated. Thus, the pooled difference chi-square had 7 degrees of freedom.

Table 2
Final Models Fitted to Six-Item Sets

Chi Square[a]	Latent Class	$p_{unconstrained}$	$p_{constrained}$	Latent Response Pattern					
				Item Level					
Booklet 1				3	2	4	1	2	2
$\chi^2_{49} = 49.91$	1	{.550}	.426	0	0	0	0	0	0
$\chi^2_{51} = 50.95$	2		.151	0	0	0	1	0	0
	3	.148	.113	0	1	0	1	1	1
	4	{.302}	.123	1	1	0	1	1	1
	5		.187	1	1	1	1	1	1
				Item Level					
Booklet 2				1	4	2	1	1	1
$\chi^2_{49} = 66.56$	1	.446	.426	0	0	0	0	0	0
$\chi^2_{51} = 68.52$	2	.175	.151	1	0	0	1	1	1
	3,4	{.379}	.236	1	0	1	1	1	1
	5		.187	1	1	1	1	1	1
				Item Level					
Booklet 3				2	3	1	2	3	3
$\chi^2_{49} = 66.18$	1	{.588}	.426	0	0	0	0	0	0
$\chi^2_{51} = 66.80$	2		.151	0	0	1	0	0	0
	3	.117	.113	1	0	1	1	0	0
	4,5	.295	.310	1	1	1	1	1	1
				Item Level					
Booklet 4				4	2	3	2	4	2
$\chi^2_{48} = 63.08$	1,2	.553	.577	0	0	0	0	0	0
$\chi^2_{51} = 64.67$	3	.152	.113	0	1	0	1	0	1
	4	.112	.123	0	1	1	1	0	1
	5	.184	.187	1	1	1	1	1	1
				Item Level					
Booklet 7				3	3	3	2	2	3
$\chi^2_{49} = 62.08$	1,2	.565	.577	0	0	0	0	0	0
$\chi^2_{51} = 67.60$	3	.077	.113	0	0	0	1	1	0
	4,5	.358	.310	1	1	1	1	1	1

[a] Chi-squares are shown first for unconstrained and then for constrained analyses.

Its value was 10.73, $p > .15$. The ideal simultaneous estimation of latent class and item parameters by maximum likelihood would minimize the value of this pooled difference chi-square. Thus, any departures from the maximum likelihood estimates due to the use of the two-stage procedure actually followed would increase the value of the test statistic. For this reason, the obtained chi-square of 10.73 is an upper bound

to the difference chi-square that would be found if the theoretically correct but computationally infeasible maximum likelihood estimation were carried out.

The common latent class parameters were estimated by the following weighted least squares regression procedure. Each of the original (unconstrained) within-booklet analyses yielded independent estimates of examinee proportions in single latent classes and/or (pooled) sets of classes. Denote these estimates $\hat{y}_{ij}$ for booklet i and class j ($j = 1, \ldots, 5$) or pooled classes (1 and 2; 4 and 5; or 3, 4, and 5; $j = 6, \ldots, 8$). Let w_{ij} be inversely proportional to the variance (squared standard error) of the corresponding estimate $\hat{y}_{ij}$. The pooled estimates $\tilde{y}_j$ were obtained by minimizing the sum of squared deviations $\sum_i \sum_j w_{ij} (\hat{y}_{ij} - \tilde{y}_j)^2$, subject to the constraints

$$\begin{aligned} \sum_{j=1}^{5} \tilde{y}_j &= 1 \\ \tilde{y}_6 &= \tilde{y}_1 + \tilde{y}_2 \\ \tilde{y}_7 &= \tilde{y}_4 + \tilde{y}_5 \\ \tilde{y}_8 &= \tilde{y}_3 + \tilde{y}_4 + \tilde{y}_5 \quad . \end{aligned} \tag{3}$$

In practice, this estimation was readily accomplished using standard software. The GLM procedure in the SAS statistical software system was used with a WEIGHT variable and an option (NOINT) to force the regression line through the origin. The vector of $\hat{y}_{ij}$ formed the dependent variable, except that for $j = 5$, 7 or 8, $1 - \hat{y}_{ij}$ was substituted for $\hat{y}_{ij}$. There were four binary independent variables coded as shown in Table 3. Regression weights yielded $\tilde{y}_1, \ldots, \tilde{y}_4$ directly; $\tilde{y}_5$ was obtained by subtraction.

Results

For the unified model to serve as a comprehensive framework for describing skills in the domain sampled required more than statistical evidence of satisfactory fit to the data. It was also necessary that items soluble by examinees at a given level cohere in some way, that they define discernable levels of skill, or content mastery. Skill level definitions were developed, corresponding to the empirically determined scale.

Table 3
Coding of Binary Variables for Estimation of $\tilde{y}_j$

j	Latent Class(es)	v_1	v_2	v_3	v_4
1	1	1	0	0	0
2	2	0	1	0	0
3	3	0	0	1	0
4	4	0	0	0	1
5[a]	5	1	1	1	1
6	1+2	1	1	0	0
7[a]	4+5	1	1	1	0
8[a]	3+4+5	1	1	0	0

[a] Dependent variable $\hat{y}_{ij}$ was replaced by $1 - \hat{y}_{ij}$.

Items in the five booklets, 30 items in all, were divided into four categories, as indicated by the latent class models. These four categories defined the five levels of content mastery distinguished in the final pooled model. Due to problems of model identifiability discussed above, items at the lowest or highest levels for a given booklet could sometimes be placed in either of two adjacent categories. Where a booklet contained only a single item at the lowest level, a parameter representing its latent class and the adjacent class pooled together was estimated, and the false positive parameter for the single item was inflated. At the high end of the scale, the corresponding identification problem could lead to depression of an item's true positive parameter. Thus, in the lowest group, if any item had a markedly higher false positive probability than was typical, it was tentatively placed at a still lower level, while a high-group item with a low true-positive probability was tentatively moved up.

Inspection of the content of items in the four categories clearly showed a trend toward greater complexity at higher levels, but the specific content of exercises at each level showed some diversity. Roughly three items for which levels were ambiguous were moved to adjacent levels. The following rough interpretation was then made of the levels. Recall that the defining attribute of exercises in the domain was the appearance of a letter representing some variable quantity in an expression, equation, or inequality. At Level 1 (Prealgebra), this use of a letter appeared largely incidental to the substance of the problem. Most Level 1 exercises required interpretation of letters only in contexts typically encountered before any formal instruction in algebra.

Specifically, the exercises required an understanding of the number line, interpretation of symbols for inequalities ($<$, $\geqslant$, etc.), or direct substitution of given numerical values into simple expressions or equations. An example of the latter type of problem would be, "What is the value of $X + 2$ when $X = 4$?" One exercise testing understanding of the multiplicative property of zero also appeared.

Level 2 (Translation) exercises almost all required (1) translation of prose statements into symbolic form, (2) selection of the simple algebraic expression, equation, or inequality corresponding to a given prose statement, or (3) simplification of inequalities stated in prose. The last of these types included problems of the form, "If $X-5$ is equal to or greater than 7, then X must be equal to or greater than what number?" Also in the second category were a multiple-choice question about the sign of n in an expression like $(-3444/n = +246)$; a multiple-choice question requiring selection of one of four equations correctly expressing the result of adding or multiplying by 0 or 1 (a typical distractor was "$X + 0 = 0$"); a multiple-choice problem requiring the student to choose the correct simplification of an inequality like "$3y + 5 > 20$," and one problem similar to number series completion items, in which the student had to fill in a blank in a table representing a simple functional relationship between X and Y.

Exercises appearing at Level 3 (Linear) required the solution of linear equations or inequalities in one unknown, or factoring (but not solution) of quadratic equations in one unknown. The exercises also required simplifying more complex algebraic expressions, drawing the graph of a linear equation, and recognizing that a pair of equations in x and y defined parallel lines. None of these items provided prose statements of expressions, equations, or inequalities. They primarily tested content introduced during the latter part of the first semester of algebra instruction.

The highest level, 4 (Quadratic), contained four exercises requiring solution of two simultaneous linear equations (integer solution), solving quadratic equations (both roots required for credit), or determining the equation of a line from its graph. These were all second- or third-semester algebra topics.

To confirm these impressions and to locate different item types in the typical curricular sequence in algebra, problems similar to each of those at Levels 2, 3, and 4 were located in the widely-used text, *Modern Algebra, Book 1* (Dolciani, Berman, & Freilich, 1962). The timetable in the Teacher's Edition permitted correlation of page numbers with semesters of algebra study. (Level 1 exercises either did not appear or appeared in the first few pages of the text.) Level 2 exercises appeared in the first 163 pages, or roughly the first eight weeks of a first course in algebra. Level 3 exercises were of types introduced

in pages 78–368, corresponding to the latter part of a first-semester algebra course.[4] The highest level, 4, included only second- or third-semester topics introduced on pages 370, 467, or in Book 2 of the series.

According to the comprehensive model, 43% of the population of 17-year-olds were unable to solve any of the exercises. Correct responses to any item by a member of this group would be attributed to guessing. The remaining 57% could solve at least the Level 1 (Prealgebra) exercises, and 42% could solve Level 2 (Translation) problems, 31% could solve Level 3 (Linear) problems, and 19% could solve even the Level 4 (Quadratic) exercises.

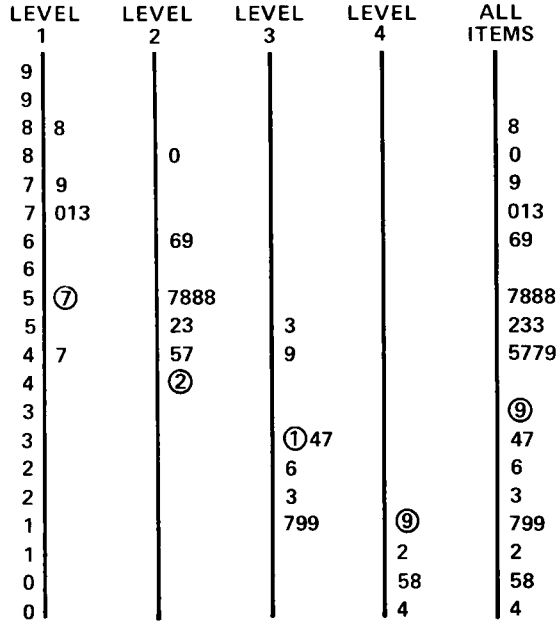
Discussion and Conclusions

Item difficulties (proportion correct) are not adequate to describe the distribution of skills in an examinee population. The proportion of correct responses reflects not only the extent of skill mastery, but also, and to an unknown degree, the influences of item format and specific item content. Aggregation of item difficulties across items permits generalization from a broader content sample, but it cannot eliminate systematic biases due to format. The model demonstrated in this paper distinguishes the extent of skill mastery, reflected in latent class proportions, from the confounding influences of item format, content of distractors, and so forth, which are reflected in the misclassification probabilities for each separate item. These proportions and probabilities are estimated jointly as distinct parameters of the model. The median false positive (guessing) probability for free response items was .06, while that for multiple-choice items was .32. The single item requiring the respondent to draw a graph had a false positive probability of only .02. These format-related differences in susceptibility to guessing are not accounted for in summaries of item difficulties. Because multiple-choice items predominate at lower skill levels, mean proportions correct for these data overstate the extent of mastery of simpler skills relative to more complex performances. This is shown in Figure 2, which presents stem-and-leafs of item difficulties for each skill level, along with the proportions of skill masters for each level given by the latent class parameter estimates.

The restricted latent class models demonstrated in this study offer an alternative to latent trait models for assessment data (Bock, Mislevy, & Woodson, 1982). The two types of models share some of the same advantages but entail different sets of assumptions about the structure of the content domain and of examinee abilities. Both latent trait and latent class models provide criterion-referenced descriptions of examinee performance that are independent of the particular items given. With latent trait models, this is accomplished by locating each calibrated item on scale used to describe the distribution of examinee abilities. With the latent class models used here, examinees and items are referenced not to a common score scale but to a common empirically determined set of latent classes. Latent trait approaches require a priori classification of items into "indivisible curricular elements" (Bock et al., 1982, p. 8), each of which is unidimensional. Latent class models require no a priori classification and no assumption about dimensionality. They do, however, require that students' abilities and items' requirements be represented by a set of discrete skills, rather than continuous variables. Latent class models are applicable with smaller sets of items. They include distinct parameters to represent the skills items are intended to measure versus incidental item features (e.g., format and wording) that also influence item difficulty. The two approaches

[4]One additional exercise at Level 3 required the student to select the correct description of the set of integers shown on a number line. While this type of item appears on p. 48 of Dolciani et al. (1962), the format of this particular item presented special problems, including the unconventional use of open circles on the number line to depict points. (An open circle usually represents the end of an interval not containing the circled point.) Performance on this five-choice item (four distractors plus "I don't know") was well below chance at $p = .19$.

Figure 2
Stem-and-Leafs of Item Difficulties by Category, Also Showing Estimated Proportion Able to Solve Each Item Type (Circled Leaves)



should be regarded as complimentary, offering useful alternative characterizations of item domains and examinee abilities.

Not surprisingly, items were found to cohere in ways that reflected the organization of the curriculum. Although many NAEP exercises are highly innovative and use novel formats, virtually all exercises in the domain defined and sampled were of types found in the text, *Modern Algebra* (Dolciani et al., 1962). Page numbers of corresponding exercises in that text were closely related to item skill levels. This suggests that skill level in algebra is largely determined by specific curricular exposure, rather than general maturational or background influences. What is more intriguing is the break between Levels 2 and 3. According to the model, the 15% of examinees in latent class 3 were able to solve Level 2 (Translation) items, testing content from the first eight weeks of algebra instruction, but were unable to solve Level 3 (Linear) content from the latter part of a first-semester course. It is unlikely that 17-year-olds tested in March, April, and May are in the middle of first-semester algebra courses. However, it may be that students conforming to the third latent class took algebra and learned the terminology and rules for translation and simplification of expressions, but they failed to learn (or to retain) procedures for applying those skills to solve new problems. Overall, these analyses reveal a generally poor level of algebra skills among 17-year-olds. It is distressing that nearly half are unable to solve even the simplest items included, and less than a quarter can solve exercises testing second- or third-semester content.

In summary, these analyses demonstrate the utility of latent class models in abstracting policy-relevant generalizations about examinee abilities from matrix-sampled data. Items can be grouped according to skill level in a defensible empirical manner, and mastery proportions for each level can be estimated. It was found that, for NAEP mathematics items involving algebraic variables, a unidimensional scale comprised of four content categories defining five mastery levels fit the response data from five

exercise booklets satisfactorily. Idiosyncratic features of separate items were captured in their respective misclassification parameters, while the population distribution of underlying skills was characterized by latent class parameter estimates.

References

Bock, R. D., Mislevy, R., & Woodson, C. (1982). The next stage in educational assessment. *Educational Researcher, 11(3),* 4–11, 16.

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.

Dayton, C. M., & Macready, G. B. (1976). A probabilistic model for validation of behavioral hierarchies. *Psychometrika, 41,* 189–204.

Dolciani, M. P., Berman, S. L., & Freilich, J. (1962). *Modern Algebra: Structure and Method, Book 1.* Boston: Houghton Mifflin.

Folsom, R. E. (1977). *National Assessment approach to error estimation* (Sampling Error Monograph 250-796-5). Research Triangle Park NC: Research Triangle Institute.

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.

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

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

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. New York: Houghton Mifflin.

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

Proctor, C. H. (1970). A probabilistic formulation and statistical analysis for Guttman scaling. *Psychometrika, 35,* 73–78.

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

Acknowledgment

The work on which this publication is based was performed pursuant to Grant NIE-G-80-0003 of the National Institute of Education. It does not, however, necessarily reflect the views of that agency. The author is indebted also to Nel Noddings for her advice and assistance in reaching the substantive interpretation of skill levels presented.

Author's Address

Send requests for reprints or further information to Edward Haertel, School of Education, Stanford University, Stanford CA 94305, U.S.A.