DOCUMENT RESUME

ED 208 005 TM 810 685

AUTHOR Wiley, David E.; And Others
TITLE Test Validity and National Educational Assessment: A Conception, a Method, and an Example.
INSTITUTION CEMREL, Inc., Chicago, Ill. ML-GROUP for Policy Studies in Education.; Northwestern Univ., Evanston, Ill.
SPONS AGENCY National Inst. of Education (ED), Washington, D.C.
PUB DATE 81
GRANT NIE-G-78-0155
NOTE 74p.

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Educational Assessment; Elementary Education; Error of Measurement; *Latent Trait Theory; Maximum Likelihood Statistics; *Models; National Competency Tests; *Reading Comprehension; Reading Skills; *Standardized Tests; Testing Problems; *Test Validity
IDENTIFIERS Empirical Analysis

ABSTRACT

This paper brings to first fruition an analytic schema based on four elements which involve a conception of skills independent or particular testing devices: (1) the development and application of a class of statistical models incorporating qualitative definitions of skill, distorted in item response by errors conceived as misclassifications; (2) a critique and reformation of the concept of test validity--making more concrete and specific the implications of invalidity; and (3) an integration and fusion of these concepts which allows meaningful empirical analyses of item response data. This conception/model is exemplified as contributing to the clarification of previously intractable technical and policy issues in the testing field. (Author/GK)

Test Validity and National Educational Assessment:
A Conception, A Method, and An Example

David E. Wiley
Northwestern University

Edward Haertel
Stanford University

Annegret Harnischfeger
Northwestern University
and
CEMREL, Inc.

1981

CEMREL, Inc.
227 Sheridan Road
Kenilworth, Il. 60043

Abstract

This paper brings to first fruition an analytic schema based on four elements. These involve a conception of skills independent of particular testing devices: the development and application of class of statistical models incorporating qualitative definitions of skill, distorted in item response by errors conceived as misclassifications; a critique and reformation of the concept of test validity--making more concrete and specific the implications of invalidity; and an integration and fusion of these concepts which allows meaningful empirical analyses of item response data. We believe that this conception/model will contribute to the clarification of previously intractable technical and policy issues in the testing field.

Acknowledgements

This report is an outcome of research on technical and policy issues relating to testing for competency and curriculum evaluation. It took place in the ML-GROUP for Policy Studies in Education while it was a part of CEMREL, Inc. and at Northwestern University under a subcontract from CEMREL.

The first author benefited greatly from a conversation with Lee J. Cronbach prior to the writing of this report. This conversation altered the conception and the course of argument of the report and we would like to express our gratitude for that stimulation.

The work was supported by the National Institute of Education, U.S. Department of Education (Grant Number NIE-G-78-0155). The opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no endorsement by that agency should be inferred.

Warm thanks to Mary Bucholtz for enthusiastic and exemplary efforts in preparing this manuscript.

Table of Contents

1. The Test Scene: Standards of Performance, Test Instruments, and Educational Assessment 1

2. Validity Reconsidered 4

3. Valid and Meaningful National Estimates of Reading Comprehension Skill 10

4. Study Design: Model, Data, Analysis 13

4.1 The Model: Latent States, Latent Responses, and Misclassification 13

4.2 The Data: Sample and Testing Design 16

4.3 The Analysis: Item Selection and Model Specification 20

5. Grade Progression in Reading Skill: Tentative Assessments 24

5.1 The Models: How Well Do They Fit? 24

5.2 Response Validity: Matches and Mismatches between Manifest Response and Skill Level 26

5.3 Grade Progressions: Direct Estimates, Precision, and Extensions 30

6. Conclusions 34

References 39

Appendix A A1-A13

Appendix B B1-B9

Appendix C C1-C2

List of Displays

4.1 Population and Sample Design 17

4.2 Test Form and Levels Administered and Sample Sizes, By Grade 19

4.3 Items and Passages Selected for Analysis, by Test Level 22

4.4 Skill-Level Proportions Directly Estimable, by Grade 23

5.1 Comparisons of Two and Three Latent State Models 25

5.2A Misclassification Parameter Estimates, by Item and Grade 28

5.2B Misclassification Parameter Estimates, by Item and Grade ... 28

5.2C Misclassification Parameter Estimates, by Item and Grade ... 29

5.2D Misclassification Parameter Estimates, by Item and Grade ... 29

5.3 Estimates of Latent State Probabilities, by Grade and Item Set 31

5.4 Estimates Variances and Covariances of Latent State Probability Estimates 33

5.5 Fitted State Probabilities, by Grade and Item Set 34

1. The Test Scene: Standards of Performance, Test Instruments, and Educational Assessment

Historically, the purpose of educational and psychological test instruments has been to ground decisions about individuals; mainly to sort individuals into groups of relatively homogeneous intelligence, ability, performance, or achievement. The use of achievement tests for program or system evaluation is relatively new. It has been strongly advocated only during the last decade. Accordingly, tests that were originally designed to compare and sort individuals, such as standardized achievement tests, have been and are currently also widely used in the evaluation of educational programs and systems. Increasingly widespread state testing programs commonly use standardized tests for assessing pupil performance statewide and at district levels, but often they also provide test score information to schools and teachers about their pupils so as to ease and improve local decisions about pupil instruction.

Standardized, norm-referenced tests, primarily designed to position pupils relative to one another and to typical performance levels ("norm" distributions) on an achievement continuum, are still the most common test type in use, both for such individual assessments and for program or school system evaluation. This type of test, almost exclusively, is also used to predict future performance of individuals.

College entrance examinations, such as the Scholastic Aptitude Test (SAT), general aptitude batteries, and standardized intelligence tests are among the most common, if problematic, devices for such individual performance predictions. These norm-referenced tests have severe shortcomings, however. With growing concern about educational goals, their accomplishment through specific programs, and the assessment of their attainment by individuals, displeasure with norm-referenced tests has increased. Dissatisfactions have arisen because these tests do not address specific, defined goals and objectives and their mastery by individuals.

Objective-, domain-, and criterion-referenced tests, all of which focus on specific content, objectives, goals, and achievements to be reached, have emerged. The development of such tests was also impelled by the increasing resources, human and material, available to teachers, allowing them to individualize instruction with respect to content and goals, which in turn necessitated individually tailored assessments of pupil achievement. A third movement, born because of dissatisfaction with the achievements of high school graduates, has adjoined itself and together they have compelled the development of tests linked directly to educational goals. The need for minimal performance standards for graduation and promotion has promoted new test types.

All of these evolvements have initially concentrated on the assessment of individuals, primarily within single classrooms. Objective- and domain-referenced instruments are designed to allow concrete specification of the goals of measurement. The only current extension of objective-referenced testing beyond the classroom is attempted by the National Assessment of Educational Progress and similar state testing programs. The National Assessment measures performance of a nationwide sample of pupils in various content areas by means of formal specification of educational objectives. But their test reporting has severe limitations: Their reports do not permit, thus far, summarization of performance on test items into levels or patterns, allowing potential comparisons to performance standards.

New purposes of testing require the rethinking and modification of old procedures. And meaningful use of educational test data for nationally or regionally representative assessments of the proportions of individuals meeting educationally relevant standards would demand combinations of existing concepts in new operational forms.

Criterion-referenced tests have been constructed with narrow content ranges, because of their use for instructional decisions about individuals in classrooms or courses. Standardized, norm-referenced tests cover broader ranges of content, because of requisites for nationwide applicability and their less frequent administration to individuals, at most once or twice during a school year. Objective-referenced instruments, used in the National Assessment and intended for extensive evaluation of American education, encompass still wider ranges of accomplishment within content areas. This breadth of scope is made possible by the absence of the usual requirement of accurate measurement for every individual.

So, if we are to use the concept of a standard or performance criterion for more general purposes than individual assessment, new varieties of testing devices must be developed or important modifications of existing instruments and procedures undertaken. Thus, either criterion-referenced tests and the standards they assume need extension to broader content areas without losing the meaning of specificity of their criterion levels, or wider ranging tests must be equipped with such standards in order to serve new purposes.

It is possible to set performance standards and compare them to performance on tests which were not specifically designed for that purpose. This is surely not the most desirable state of affairs, but may be the wisest one at the beginning when we are exploring the best ways to accomplish our new goals. The intent of this paper is, in fact, to use existing--nationally representative--standardized test data to estimate the proportions of elementary school pupils in educationally meaningful performance categories.

2. Validity Reconsidered

Most recent psychometric work on validity-related matters has focussed on the use of tests for selection decisions. This work has been strongly stimulated by legal concerns about the fairness of selection procedures; primarily those used in the employment process. The focus of this research has not been on the nature of the tests themselves or the measurements deriving from them, but on the social selection procedures that incorporate these tests. Thus, the implications of the work for changes in the process relate only to the ways in which the scores of individuals with different non-test characteristics are incorporated into the criteria for selection, not to such issues as item content, item format, method of scoring, etc.

As a general perspective, this orientation fragments the validity concept --as tests are used in different ways--and it forecloses whole classes of questions that relate to item and test format, content selection, scoring and scaling. From our perspective, the new work does not focus on test validity at all. It primarily is a conceptual framework and a set of standards for assessing the social worth of selection procedures incorporating any criteria that are (a) quantitative, and (b) measured with error. Problematically, it focusses primary attention on external criteria and allows those who should be forced to attend to important concerns about the validity of their devices to ignore them.

Almost all other psychometric research, until recently, has been focussed on issues of error and reliability rather than on bias and validity. The theoretical framework for the analysis of measurement errors has become conceptually sophisticated, elaborate and full of concrete detail. It has progressed to the point that primitive correlational indices are no longer scientifically respectable as having clear meaning and where the conceptual and analytic frameworks for test items and responses to them are fully integrated with those for test scores.

On the other hand, the conceptual orientations to validity of tests are diffuse, fragmented and fundamentally incomplete. The widely accepted rubric of "construct validity" (Cronbach & Meehl, 1955) is abstractive enough so that it gives little or no guidance in the choice of operational procedures or the allocation of investigative resources. The decision-theoretic analysis of selection decisions (Cronbach & Gleser, 1957), is not integrated in any fundamental fashion with the construct framework. The recent theoretical work on selection bias builds on the decision frame but again ignores the "construct" issues. In fact, the whole issue of test "bias"--at its heart a phenomenon of differential validity--has never been linked to the core theoretical concepts of validity.

Finally, in this area, the frameworks for item assessment have never been fundamentally integrated with those for tests. Thus, "item bias" has no bearing on "test bias" and "content validity," which, at the operational level, seems to mean the sampling or selection processes for the items which make up the test, has no relational to test validity, which at the operational level, seems to mean a relation to a single external criterion in the (implicit or explicit) context of a selection decision. The fact these non-overlapping processes can be tenuously linked via the vagaries of "construct validity" does not imply that they could actually be integrated.

Inherently, the notion of test validity must rest on two conceptions: (a) that which a test ought to measure and (b) that which a test does measure. It is the discrepancies between the two, somehow defined, that bear on validity. Central theoretical and practical problems for psychometrics are (1) the mode of specification of the ought and (2) the form of expression of the discrepancy. Recent discussions of the validity concept in the psychometric literature (Cronbach, 1971; 1980) have focussed on the word interpretation as the entity which is validated. However, a central interpretation of "interpretation" has, at least since Cronbach and Meehl (1955), centered on the idea of a definition or theoretical conception of what is intended to be measured (i.e., the "construct")

--our ought. The problem with the specification of the ought is that, if it occurs at all in the actual world of test construction--beyond an undefined label--it is formulated in ways that make it difficult to separate valid from invalid components of the measurements.

Crohbach (1971) gives a salient example of a specification of an intent of measurement which highlights this issue of separation:

> Consider further reading comprehension as a trait construct. Suppose that the test presents paragraphs each followed by multiple-choice questions. The paragraphs obviously call for reading and presumably contain the information needed to answer the questions. Can a question about what the test measures arise? It can, if any conterinterpretation may reasonably be advanced. Here are a few counterhypotheses (Vernon, 1962):
>
> 1. The test is given with a time limit. Speed of reading may contribute appreciably to the score. The publisher claims that the time limit is generous. But is it?
>
> 2. These paragraphs seem abstract and full. Perhaps able readers who have little motivation for academic work make little effort and therefore earn low scores.
>
> 3. The questions seem to call only for recall of facts presented in simple sentences. One wants to measure ability to comprehend at a higher level than word recognition and recall.
>
> 4. Uncommon words appear in the paragraphs. Is the score more a measure of vocabulary than of reading comprehension?
>
> 5. Do the students who earn good scores really demonstrate superior reading or only a superior test-taking strategy? Perhaps the way to earn a good score is to read the questions first and look up the answers in the paragraph.

> 6. Perhaps this is a test of information in which a well-informed student can give good responses without reading the paragraphs at all.
>
> These miscellaneous challenges express fragments of a definition or theoretical conception of reading comprehension that, if stated explicitly, might begin: "The student considered superior in reading comprehension is one who, if acquainted with the words in a paragraph, will be able to derive from the paragraph the same conclusions that other educated readers, previously uninformed on the subject of the paragraph, derive." Just this one sentence separates superior vocabulary, reading speed, information, and other counterhypotheses from the construct, reading comprehension. The construct is not identified with the whole complex practical task of reading, where information and vocabulary surely contribute to success. A distinctive, separate skill is hypothesized. (pp. 463-464)

Cronbach's example implies several things in this context. First, it makes clear that reading comprehension as an intent of measurement is not all things to all persons; it is not speed, vocabulary, test-wiseness, or prior information, regardless of whether these "constructs" contribute to success on test task itself, other tasks given contemporaneously, or future tasks. If we take this fact and realize that such sources of invalidity in the assessment of reading comprehension are (a) themselves valid intents of measurement with other instruments and are (b) irremovable sources of variation in test performance for many "constructs"[1] then two further implications flow

--the problem of test validation, whether focussed on the notion of "interpretation" or not, cannot be shifted entirely to an analysis of test use, and that

[1] E.g., vocabulary knowledge is a logical prerequisite for appropriate performance on comprehension test tasks. Although variation in performance due to differences in vocabulary can be suppressed by experimental training or selection of common words, it cannot be removed as a source of extraneous (invalid) variation in practical test situations.

--the labeling of the test or the description of what it is intended to measure must be sufficiently precise to allow the separation of components of invalidity from valid variations in performance.

Also, we must note that these sources of invalidity are often positively related to the characteristic that is the intent of measurement. Thus, in the Cronbach example, those who have the skills necessary for "comprehension" of passage content or derivation of correct conclusions, given adequate vocabulary, will also be more likely to have previously acquired that vocabulary knowledge.

Our ongoing program of research, of which this study is a part, is fundamentally affected by these issues. For example, a "reading comprehension" test might produce scores which strongly correlate with vocabulary knowledge for several distinguishable reasons:

--those individuals who have good comprehension skills also generally have extensive vocabulary knowledge and vice versa, i.e., reading comprehension skill(s) is (are) highly correlated with vocabulary knowledge and

(a) the test primarily measures reading comprehension or

(b) the test primarily measures vocabulary knowledge

--those individuals who have good comprehension skills do not necessarily have extensive vocabulary knowledge, i.e., reading comprehension and vocabulary knowledge are not highly correlated and

(c) the test primarily measures vocabulary knowledge.

If someone were to use the test for a predictive purpose where, at least on the surface, the test label was not considered of basic importance, that person might be inconcerned about which of these were actually the case. However, if one were engaged in placement of individuals in remediation programs in reading one might hot be concerned about (a) or (b) but (c) would be troublesome. And if one were evaluating a curriculum which might change the relation between reading comprehension and vocabulary knowledge or engaging in a national social assessment of reading comprehension abilities then only (a) would constitute a satisfactory state of affairs.

As this study is focussed on the latter issue--social assessment of competencies is reading comprehension for a national population--these validity issues are critical. In order to generate valid estimates of the proportion of individuals, nationally, possessing specific levels of reading skill, we must be able to remove variations and biases deriving from other, distinct, characteristics--whether they be vocabulary knowledge or test-wiseness.

3. Valid and Meaningful National Estimates of Reading Comprehension Skill

At an earlier stage of this project, Haertel (1980) conducted a study of standardized reading comprehension tests using large national samples of response data. Three of those samples are here analyzed along with three additional ones. In the earlier study Haertel attempted to differentiate among a set of distinctively defined skills based on a linguistic analysis of the reading comprehension test tasks. These skills were defined so that each test item required a specific subset of the skills. An item response model was formulated so that individuals were assumed to belong to either a group possessing all of the skills in the subset--the "can solve" group--or a group not possessing all of those skills--the "cannot solve" group. Individuals in these two categories were not assumed to respond uniformly with correct and incorrect answers, respectively. Instead, non-matching responses were allowed to occur, with specific probabilities--so-called false negatives and false positives. Statistical analyses of the response data using the model then yielded estimates of two distinctive types of quantities

a) proportions or numbers of individuals with various combinations of skills ("latent state probabilities") and

b) proportions of mismatching responses deriving from each item ("misclassification probabilities").

The major findings of the research were that

a) the models fit the data extremely well--extensive exploration of potential lack of fit resulted in no evidence of systematic deviations and the analyses showed that the models were at least as adequate as previous psychometric models with more parameters.

b) The reading comprehension tests analyzed were not sensitive enough to allow differentiation of subskills --i.e., the models fit the data well with only one generic skill specified for each test level. Thus, a single common dichotomy (can and cannot solve) was sufficient to account for differences in the reading comprehension skills assessed by all items in a test at a specific level.

These results led us to two conclusions:

1. Standardized reading comprehension tests may not have the discriminating power attributed to them by those who focus primarily on available reliability coefficients. I.e., if, as discussed above, such tests can only grossly discriminate between two gross skill categories, then there must be large elements of the reliable variance in such tests which are invalid, and

2. If such invalid components are actually "stripped off" by the models used, then perhaps analyses could be conducted which would yield valid and meaningful national estimates of reading comprehension skills, defined at least in the broad terms corresponding to the test levels.

In the study reported in this paper, we implement the methodology and the conceptual framework applied by Haertel (1980) using six nationally representative samples of elementary school pupils--one for each of grades one through six. For these samples we estimate

a) the proportions of individuals in each grade at particular skill levels, and

b) the proportions of matching and mismatching responses for each item at each grade level.

Because the tests used at each grade level were repeated at adjacent levels, we are then able to trace changes in the proportions of skilled individuals over grades and observe systematic modifications in validity of the observed responses.

4. Study Design: Model, Data, Analysis

4.1. The Model: Latent States, Latent Responses, and Misclassifications

If students' responses to items reflected only the skills they possessed and the skills the items required, it would be possible to establish just which patterns of responses to a set of items should occur, and which should not. For any combination of skills possessed, items requiring these skills (or some of them) and no others would be answered correctly, and items requiring skills not possessed would be answered incorrectly. Only a small number of the possible response patterns would be expected to occur. For example, for five items involving only three skills, there are 32 possible patterns of correct and incorrect responses, but only 8 possible patterns of presence and absence of skills. Thus, if each combination of skills possessed determined a specific pattern of correct and incorrect item responses, at most 8 of the 32 possible response patterns would be expected to occur. If hypothesized skill hierarchies ruled out some of the 8 skill combinations, even fewer than 8 item response patterns would be expected. Of course, an item's predicted skill requirements do not completely determine which students will get it right. Each item also entails unique processes, not represented by its skill requirements. Moreover, carelessness, lapses of attention, errors in recording a response, etc. may lead to incorrect responses by students who possess all the skills an item requires, while successful guessing or elimination of distractors may lead to correct responses by students lacking one or more requisite skills. In summary, students' responses to test items are imperfect indicators of the skills they possess and the skills items require. Students possessing the requisite skills for an item may give incorrect, "false negative" responses, while students lacking one or more of the skills an item requires will sometimes give correct, "false positive" responses.

The method used in this study for the validation of skills and their relationships explicitly accounts for these imperfections. The actual responses students mark on their answer sheets are termed "manifest responses," and are distinguished from a hypothetical set of "latent responses" reflecting only the skills items require and students possess. The pattern of latent responses shows which items would be answered correctly if false positives and false negatives never occurred. There is a set of latent responses for each permissible skill combination. Thus, all students possessing a given combination of skills have the same latent responses. They are said to conform to the same latent state. In examining any set of items, the possible latent states and the latent response pattern for each state are derived prior to the computer analysis, solely on the basis of hypothesized hierarchies among skills, and the different items' skill requirements. Often, for students conforming to a given latent state (i.e., possessing a given combination of skills) the most likely manifest response pattern is the same as the latent response pattern for that state. Manifest response patterns differing for only one item from the latent response pattern are usually less likely, manifest patterns differing for two items are still less likely, etc. Each discrepancy between latent and manifest response patterns, either a false positive or a false negative, is termed a misclassification. Full details on this class of models is given in Appendix A.

The mathematical and statistical procedures used in this study--maximum likelihood methods--yield numerical estimates of the probabilities of each possible misclassification for each item. Since every manifest response to an item is either a correct classification or a misclassification, the probability of a correct classification (a manifest response matching the latent response to an item) can be calculated as one minus the misclassification probability.[2]

At the same time as it generates estimates of misclassification probabilities, the mathematical procedure produces estimates of the proportion of the students in each latent state. These are referred to as estimates of structural parameters. Every student is assumed to possess one of the permissible combinations of skills, i.e., conform to one of the latent states. Therefore, the sum of the proportions in all of the latent states equals one. The statistical procedures used in this study to estimate the parameters and assess the precision of the estimates are fully described in Appendix B.

[2] In reporting the results of all analyses, a "true positive rate" and a "false positive rate" are given for each item. The true positive rate is the probability of a "correct" manifest response, given that the latent response is "correct." This is one minus the item's false negative misclassification probability. The false positive rate is the probability of a "correct" manifest response, given that the latent response is "incorrect," i.e., the item's false positive misclassification probability.

4.2. The Data: Sample and Testing Design

This research required data from a large sample of elementary school children. Not only are the maximum likelihood methods used based on large-sample theory, but in addition, to obtain stable estimates of population proportions for the many response patterns which can occur across even a few items, numerous respondents are needed. In addition to having many respondents, it is desirable to have a large pool of test items from which to draw. This facilitates item modeling by providing more small sets of items which vary systematically in their skill requirements. Finally, the data used in this research represent well-defined populations, so that estimates of population parameters and their standard errors can be meaningfully interpreted.

The Sustaining Effects Study, carried out by System Development Corporation, included the collection of achievement test data from a large nationally representative sample of pupils in grades one through six. Data were collected in fall of 1976 using tests of vocabulary, reading comprehension, mathematics concepts, and mathematics computation. The sampling design and procedures employed in this extensive data collection are described in Sustaining Effects Study Technical Report Number 1 (Hoepfner, Wellisch, and Zagorski,1977). For a representative subsample, The Participation Study, of the same pupils, Decima Research collected extensive, detailed information on home background and economic status (Breglio, Hinckley, and Beal, 1978). The population and sample definitions for this data base are given in display 4.1.

Display 4.1. Population and Sample Design

Population: All 20,881,979 public elementary school pupils enrolled in grades 1 through 6 in the 50 United States, during the 1976-77 school year (62,534 schools).

Sample Design: 2-stage, stratified random cluster sample, implemented with replacement schools to adjust for non-cooperation.

Strata: 10 Federal districts

x 3 LEA Sizes

x 3 LEA Poverty levels

Yields: 90 Strata

- 6 Strata without schools

Yields: 84 Strata

Clusters: 3 schools per stratum

Yields: 252 schools = 84 strata times 3 schools per stratum

-10 lost without substitution

Yields: 242 schools

Units: 18,000 pupils

- 362 lost or moved

Yields: 17,368 pupils

- 2

Yields: 17,366 pupils on final data file

The Sustaining Effects/Participation Study provides data on the reading comprehension scales of the Comprehensive Test of Basic Skills (CTBS) form S, at levels A (grade 1) through 3 (grade 6). These tests were given, in pairs, at each grade level. Each test level and the grades at which it was given are exhibited in Display 4.2.

As discussed in the section on the problem of design effects (Appendix B), the theory on which the chi-square test and asymptotic standard errors are based requires a simple random sample from the population. Data from the Sustaining Effects Study, however, represents a stratified cluster sample. In this study, a universe of schools was first defined, and all schools in the universe were divided into strata according to size, location and other demographic characteristics. For the Sustaining Effects Study, the universe included public schools with some of grades 1 through 6. Once strata were defined, some schools were randomly sampled within each stratum, and the students tested were all clustered within these selected schools. In the Sustaining Effects Study, students were randomly sampled within schools, and the number tested was determined by the school's size.

A preliminary analysis was conducted to estimate the effective sample size. The fifth grade was chosen for this analysis, and four representative items were selected from the level 1 test. Each fifth grade student's response pattern across these four items was tabulated, and the variance of estimated proportions in each of the 16 response categories was computed using the ultimate cluster estimate of the rel-variance for ratios (Hansen, Hurwitz, & Madow, 1953, pp. 316-321). To obtain the standard error of each estimate, the square

Display 4.2. Test Form and Levels Administered and Sample Sizes, by Grade

Test: CTBS - Form S - Reading Comprehension (including Sound Matching)

Levels and Their Characteristics

Level	Title	No. of Passages	No. of Items	Sentences/ Passages	Response options per item
A	Sound Matching	0	28	-	3
B	Reading Comprehension	24	24	1.4	3
C	Read. Comprehension: Passages	6	18	7.8	4
1	Reading Comprehension	7	45	9.1	4
2	Reading Comprehension	7	45	11.9	4
3	Reading Comprehension	7	45	11.4	4

Level/Grade Match and Sample Sizes

	Level		Sample Size		
Grade	Below Grade Level	At Grade Level	Before edit	After edit	Effective
1	A	B	3103	2598	799
2	B	C	2750	2188	884
3	C	1	2753	2395	986
4	1	2	2638	2327	919
5	1	2	2737	2520	1005
6	2	3	3385	3017	1127

root of the rel-variance was multiplied by the estimated proportion. In a simple random sample, the standard error of a proportion, p, is simply the square root of p times one minus p, divided by the sample size. Using this formula, the effective sample size could be computed for the estimate of each proportion by determining the size of the simple random sample which would yield the same standard error as that actually obtained. To arrive at a single estimate of the effective sample size for use in the study, the harmonic mean of the 16 effective sample sizes was computed, weighting each according to the corresponding estimated proportion. Once the grade 5 effective sample size was obtained, effective sample sizes for other grades were estimated by calculating the size of a simple random sample which would yield tne obtained standard error, given the obtained standard deviation. Since the ratio of the actual sample size to the obtained sample size should be relatively invariant across grade levels, effective sample sizes for the grades could then be estimated using the fifth grade effective sample size, the fifth grade actual sample size, and the actual sample size at the other grade levels (Display 4.2).

4.3. The Analysis: Item Selection and Model Specification

In designing the analyses, we selected a series of items from each test-level. The CTBS-Forms test levels chosen for analysis were: B; C, 1, 2. Within each level, we chose three items under the constraint that each relate to a different reading passage.[3] Thus, a total of 12 items were originally selected

[3] Haertel, in the earlier study (1980), found that selecting more than one item relating to the same reading passage resulted in dependencies which distorted the generality of the skill, defining it in a passage-dependent context, fixing vocabulary and other passage characteristics.

from the four test levels. Additionally, a set of 12 items were independently selected using the same constraints. In the following section, the first set is referred to as the "X-items" and the second set as the "Y-items." The full specification of these items is given in Display 4.3.

The two sets were then used to produce two separate "chains" of linked analyses. The analyses were specified by fitting a two-state model--can solve vs. cannot solve--for each grade in which only one test level was analyzed: Grade 1 - level B and Grade 6 - level 2. For the other grades, three-state models were fitted. Display 4.2 specifies the level combinations analyzed in these grades.

The skill combination states specified for these latter analyses were formalized as possession of (a) neither of the skills corresponding to the analyzed test levels; (b) the skill corresponding to the lower-level test, but not the upper-level one, and (c) the skills corresponding to both test levels.

The implications of these state definitions for misclassification proportions are:

> State (a)--All correct responses are false positive and all incorrect responses are true negatives.
>
> State (b)--All correct responses to items on the lower form are true positives while all incorrect responses to these items are false negatives. All correct responses to items on the higher form are false positives while all

Display 4.3. Items and Passages Selected for Analysis, by Test Level

Level	X-items		Y-items	
	Item No.	Passage No.	Item No.	Passage No.
B	1 4 15	1 4 15	2 6 7	1 4 15
C	1 6 18	1 2 4	2 15 16	1 3 4
1	1 16 29	1 3 5	11 18 28	2 3 5
2	6 14 33	2 3 6	13 21 26	3 4 5

incorrect responses to these items are true negatives.

State (c)--All correct responses are true positives and all incorrect responses are false negatives.

The complete chain of X-item analyses over grade levels was then replicated with the Y-items, producing two alternate sets of estimates of the "latent state" parameters (Display 4.4). Finally, a simple scaling model was used to extend the estimates of the proportions of individuals at each skill level over all grade levels.

Display 4.4. Skill-Level Proportions Directly Estimable, by Grade

Grade	Skill-Level Proportion			
	less than B	less than C	less than 1	less than 2
1	*			
2	*	*		
3		*	*	
4			*	*
5			*	*
6				*

5. Grade Progression in Reading Skill: Tentative Assessments

This section displays and discusses the results of the statistical analyses outlined in Section 4.3. It is organized into four subsections which focus, in turn, on the fit of the models, the estimates of misclassification rates, the changes over grades in proportions of individuals possessing various levels of reading skill, the precision of the grade-change estimates, and a preliminary extension of those estimates.

5.1. The Models: How Well Do They Fit?

The empirical study which preceded this one (Haertel, 1980) strongly supported the conclusion that standardized tests of reading comprehension--at least those intended for elementary school pupils--can only grossly differentiate the skill levels of such pupils. In fact, using the class of statistical models that are fitted here, the earlier study found that distinctions beyond the dichotomy "can solve-cannot solve" were not attainable within a specific test level. Thus, in investigating grade-level progressions in skill, the first issue to resolve was whether distinct test levels required distinct skills, or--more accurately--whether the skill differences manifested between the test levels were detectable with the sample sizes and methods used in this study.

Display 5.1. organizes and exhibits the evidence bearing on this issue. The two states (can solve-cannot solve) used earlier were fitted to the data from pupils at grade levels 2, 3, 4, and 5. The three state models described in Section 4.3. were also fitted to these data. The left hand columns of the display exhibit the grade levels, test-level combinations, and item sets for

Display 5.1. Comparisons of Two and Three Latent State Models

Grade	Levels	Item Set	2-state model		3-state model		difference		sig.
			χ^2	df	χ^2	df	χ^2	df	
2	BC	X	93.08	50	34.15	49	58.93	1	<.001
		Y	80.32	50	39.26	49	41.06	1	<.001
3	C1	X	77.88	50	54.03	49	23.85	1	<.001
		Y	68.24	50	36.07	49	32.17	1	<.001
4	12	X	64.15	50	58.23	49	5.92	1	.015
		Y	54.69	50	48.92	49	5.77	1	.016
5	12	X	44.57	50	42.02	49	2.55	1	n.s.
		Y	39.77	50	35.45	49	4.32	1	.05

which models were fitted. The remainder of the table contains the likelihood ratio chi-square values for the models, together with the difference between them.[4] The letter statistic yields an assessment of the value of the third state in explaining the responses of the individuals. Thus it informs us about whether skill level differences are manifested, in a detectable form, between the test levels.

The evidence clearly supports test level differences, especially among the earlier ones. And none of the three-state models display more than chance levels of lack of fit. The two-state models clearly do not fit well for the early grade levels, with the fit improving in higher grades. Thus, levels

[4] The difference χ^2 is merely the difference in χ^2 values resulting from the two estimation procedures. Under the hypothesis that the two-state model is correct, it is distributed as (central) χ^2 with one degree of freedom.

B and C manifest clearly distinctive skills. This is to be expected, as these test levels are quite different (Display 4.2). Level B contains only one item per passage and each passage averages only 1.4 sentences in length. On the other hand, Level C was constructed with three items per passage and the passage lengths average almost eight sentences. Clearly different skill levels are required and they are amply manifested in the data.

Differences, likely smaller but still clear, are exhibited between Level C and Level 1 as well. Such differences, however, become difficult to detect when we compare Levels 1 and 2. For the Fourth grade group, there is some evidence but it is considerably weaker than at lower test levels and no evidence of such distinctiveness is apparent at Fifth grade. In what follows, we will maintain the Level 1-Level 2 distinction but the proportion of individuals estimated to be in the Level 1--intermediate--state is uniformly small.

All in all, there are no obvious differences between the two item sets (X and Y) in the evidence they provide and the three-state models all fit the data well.

5.2. Response Validity: Matches and Mismatches between Manifest Response and Skill Level

Rates of valid correct responses. Display 5.2 (A and C) exhibits estimates of the rates at which individuals in the various grades respond correctly to each of the items[5] when they actually possess the reading skill appropriate to the

[5] Note that there are 30 item/grade-level combinations for each item set.

particular test level. Of considerable importance is the fact that all these values (except two) are estimated to be less than one. And the vast majority of these values are precisely enough estimated to be clearly distant from one in fact. This implies that there is an appreciable probability that an individual possessing the relevant skill will manifest an incorrect response.

Note, however, that all values but one exceed 0.5, which is surely a baseline of minimal validity, and also that of the thirty-six potential differences in parameter values across adjacent grade levels, thirty-two display increases.[6] This implies that, for particular items, factors which cause skilled individuals to respond incorrectly diminish in their impact over grades.

Rates of invalid correct responses. Display 5.2 (B and D) also exhibits estimates of the rates at which individuals in the various grades respond correctly to each of the items when they actually do not possess the reading skill appropriate to the test level. In more simplified models of the response process, these rates are termed "guessing" probabilities and are sometimes "corrected" via "formula" scoring. Note that forty-eight of the sixty estimates exceed the nominal (equi-probable) "guessing" values.[7] Note also that of the thirty-six potential differences in parameter values across adjacent grade levels, thirty-four display increases. This implies that, for particular items, factors which

[6] Note that items from Levels 1 and 2 were repeated in three grades while those from Levels B and C were only repeated in two grades.

[7] "Guessing" probabilities are usually estimated by the reciprocal of the number of response options. Thus, the nominal values are 1/3 for Level B and 1/4 for Levels C, 1, and 2 (Display 4.2).

Display 5.2.A Misclassification Parameter Estimates, by Item and Grade

True Positive - Set X

				Grade			
Level	Item	1	2	3	4	5	6
B	1	1.000	0.999				
	4	0.757	0.998				
	15	0.734	0.983				
C	1		0.960	0.994			
	6		0.922	0.967			
	18		0.743	0.945			
1	5			0.887	0.945	0.925	
	16			0.865	0.964	0.962	
	29			0.849	0.947	0.975	
2	6				0.946	0.963	0.976
	14				0.672	0.719	0.858
	33				0.571	0.633	0.758

Display 5.2.B Misclassification Parameter Estimates, by Item and Grade

False Positive - Set X

				Grade			
Level	Item	1	2	3	4	5	6
B	1	0.280	0.654				
	4	0.507	0.678				
	15	0.360	0.442				
C	1		0.414	0.601			
	6		0.240	0.441			
	18		0.244	0.277			
1	5			0.391	0.580	0.687	
	16			0.274	0.457	0.484	
	29			0.222	0.379	0.401	
2	6				0.393	0.484	0.599
	14				0.214	0.259	0.230
	33				0.318	0.328	0.341

Display 5.2.C Misclassification Parameter Estimates, by Item and Grade

True Positive - Set Y

		Grade					
Level	Item	1	2	3	4	5	6
B	2	0.842	0.983				
	6	1.000	0.999				
	7	0.612	0.999				
C	2		0.898	0.970			
	15		0.965	0.984			
	16		0.912	0.992			
1	11			0.392	0.502	0.517	
	18			0.631	0.911	0.912	
	28			0.903	0.951	0.964	
2	13				0.822	0.835	0.854
	21				0.888	0.928	0.973
	26				0.897	0.941	0.976

Display 5.2.D Misclassification Parameter Estimates, by Item and Grade

False Positive - Set Y

		Grade					
Level	Item	1	2	3	4	5	6
B	2	0.373	0.592				
	6	0.344	0.635				
	7	0.319	0.507				
C	2		0.303	0.423			
	15		0.439	0.562			
	16		0.246	0.396			
1	11			0.220	0.230	0.238	
	18			0.243	0.331	0.407	
	28			0.351	0.471	0.520	
2	13				0.455	0.441	0.448
	21				0.307	0.391	0.356
	26				0.296	0.381	0.497

cause unskilled individuals to respond correctly increase in their impact over grades.

Factors contributing to invalidity. It is instructive to recall here the threats to validity of reading comprehension inferences which Cronbach (1971) took from Vernon (1962). Of the six threats which he summarized, three would affect the rate of true positive responses and three the rate of false positive responses. Those falling in the first category include: speed, motivation, and vocabulary. If an individual possessed the requisite reading comprehension ability but a) took longer to read and respond than the time allowed, b) found the material sufficiently foreign to his experience or interest to try hard, or c) had insufficient vocabulary to exercise his comprehension skills, then he might respond incorrectly. These factors, however, would have no impact on the rate of false positive response.

On the other hand, recognition/recall, test-wiseness, or prior information would have no impact on the rates of true positive responses. However, if a) the item tested recall or recognition rather than comprehension, b) the individual had the skill to eliminate inappropriate response options without comprehending the passage, or c) if he knew the answer without reading the passage, the unskilled individual could attain a correct response at a rate above the base guessing probability. -

5.3. Grade Progressions: Direct Estimates, Precision, and Extensions

Estimates. The estimates of the proportions of individuals, within each grade, who possess skills below each test level are given in Display 5.3. As the

Display 5.3. Estimates of Latent State Probabilities, by Grade and Item Set

Item Set	Grade	Cumulative Probability of State			
		$<B$	$<C$	<1	<2
X	1	0.843	-	-	-
	2	0.270	0.508	-	-
	3	-	0.333	0.411	-
	4	-	-	0.444	0.504
	5	-	-	0.289	0.325
	6	-	-	-	0.295
Y	1	0.789	-	-	-
	2	0.314	0.481	-	-
	3	-	0.274	0.489	-
	4	-	-	0.498	0.555
	5	-	-	0.345	0.391
	6	-	-	-	0.263

test level/grade level matches were not complete, estimates are missing for the lower test levels in the higher grades and vice versa. As the proportions are cumulative, they increase over skill levels within a grade. These increases result from the definition of the proportions and are not empirical findings. The proportions decrease over grade levels for a particular skill column. This is an empirical finding and signals the increase in the proportion of those attaining particular skill levels over grades. The only exception to this occurs between the first and second entries in the third column and these differences are small and probably reflect the fact that the skills reflected in test levels 1 and 2 are difficult to distinguish (see Section 5.1.).

The corresponding values estimated from the two item sets (X and Y) are approximately equal and the general findings are consistent and clear. The percentages of individuals who possess the most minimal comprehension skills (level B) increase from about 20 percent at the beginning of grade one[8] to about 70 percent at the beginning of grade 2. At higher skill levels (level 1 or above), the percentage of skilled individuals increases from about 25 percent in grade three to substantially more than 70 percent by grade six.

Precision. Estimates of the variances and covariances[9] of the values given in Display 5.3 are exhibited in Display 5.4. These estimates are organized by item set and grade level. Because the grade-level samples are constituted of different individuals, parameter estimates for distinct grade levels do not covary. Thus, covariances are displayed for estimates pertaining to common grade levels only. The "first" and "second" designations in the column headings refer to the first and second entries in the corresponding row of Display 5.3.

Values of the first and sixth grade variances are larger than the other values because two-state models were fitted to data deriving from only one test level. When three-state models are fitted to data from two appropriate test levels, individuals are more finely differentiated and standard errors of estimates diminish even though estimates are unbiased, in either case, under the model. This is akin to the increases in precision accompanying an analysis of covariance. In the case of precision estimates, differences between item set-X values and item set-Y values are real because distinct item sets are differentially informative about the parameter values.

[8] $[(1. - .84 + 1. - .79)/2 \approx .19]$

[9] These sampling variances and covariances are based on the effective sample sizes rather than the actual ones and thus are adjusted for the sampling design's effect on precision (see Appendix B).

Display 5.4. Estimated Variances and Covariances of Latent State Probability Estimates*

		Estimated Dispersions ($X10^4$)		
Item Set	Grade	first variance	covariance	second variance
X	1	210.629	-	-
	2	5.083	2.297	7.125
	3	3.662	2.634	4.523
	4	7.352	4.979	8.974
	5	6.274	4.312	7.623
	6	26.276	-	-
Y	1	33.160	-	-
	2	2.083	0.971	2.227
	3	2.056	1.619	7.622
	4	3.297	1.838	2.726
	5	3.041	1.751	2.629
	6	7.251	-	-

*These dispersions should be referred to the cumulative probabilities given in Display 5.3. All values should be divided by 10^4.

Extensions. If data were available on the whole range of test levels for individuals in each grade, Display 5.3 could be extended to show how extensively skills at each level were mastered by those in each grade. Display 5.3 exhibits the results of an analysis which extended these values indirectly. This analysis was performed by scaling the values in Display 5.3 according to a model which assumed that the cumulative probabilities could be logistically transformed so that the values resulting were an additive function of parameters representing grade and test level.[10]

[10] Formally, the cumulative probabilities were transformed via $\lambda_{ij} = \ln[p_{ij}/(1-p_{ij})]$ and the model:

$$\lambda_{ij} = \mu + \alpha_i + \beta j + \varepsilon_{ij},\ i = 1,\dots,6,\ j = 1,\dots,4.$$

was assumed. The original λ_{ij} are given in Display C.1 (Appendix C). The baseline, grade, and test-level parameters are given in Display C.3. The (continued on next page)

Display 5.5. Fitted State Probabilities, by Grade and Item Set

		State Probabilities				
Item Set	Grade	<B	B	C	1	≥ 2
X	1	0.843	0.094	0.023	0.007	0.033
	2	0.270	0.238	0.112	0.047	0.333
	3	0.152	0.181	0.108	0.051	0.508
	4	0.153	0.183	0.108	0.051	0.505
	5	0.082	0.117	0.083	0.043	0.675
	6	0.072	0.045	0.077	0.041	0.705
Y	1	0.789	0.094	0.067	0.010	0.040
	2	0.314	0.167	0.220	0.043	0.256
	3	0.157	0.117	0.215	0.053	0.458
	4	0.162	0.119	0.217	0.053	0.449
	5	0.092	0.078	0.171	0.050	0.609
	6	0.053	0.049	0.122	0.039	0.737

The entries in Display 5.5 should be treated with caution in tracing skill gains as they are based on stringent assumptions about the uniformity of such skill gains over grade levels. They, however, do provide some baseline data for future studies of skill acquisition.

6. Conclusions.

There are two major thrusts of this study. One relates to the discussion of test validity undertaken in Section 2. That discussion attempted to lay out

variances and covariances of the logits are given in Display C.2. Estimates were derived by sequentially differencing the logits, beginning with grade 1 and averaging the sole pair of duplicated estimates. This "degree of freedom" was also used to "test" the model, using values from Display C.2. Resulting parameter estimates were used to reconstruct cumulative probabilities for all table locations and these were differenced to produce Display 5.4.

the ground for a reconception of validity based explicitly on the notions that tests have intents and that their characteristics never completely match those intents. The deduction from this specification was that what tests are intended to measure ought to be defined in a fashion that is, both verbally and formally, independent of the test instrument. Only such a definition will allow the use of the construct validity notion in a productive fashion, differentiating invalid from valid components of measurement, even when they are related to one another.

The methodology used and the data sets to which it was applied permitted us to explore (Haertel, 1980) and define analyses which came to empirically distinguish between reading comprehension and other, related, characteristics which standardized tests of reading comprehension measure. The distinction arrived at is surely incomplete, but the results are surely provocative enough to stimulate considerable further work. Because of our ability to distinguish between parameters which related only to the reading comprehension construct and other parameters which directly reflect components of invalidity, and because these latter parameters are further differentiated with respect to the particular variety of invalidity, we were able to trace changes in the validity of the reading comprehension test scores over grade levels. In doing this, we observed that some components of invalidity decrease while others increase as the grade level and thus the reading comprehension skill increases. We believe that the conceptual framework and the modes of analysis used here will eventually lead to a much more structural and surely more accurate analysis of the validity of tests such as those analyzed in this paper.

The second major challenge taken up in this study was that of estimating grade level changes in the reading comprehension skill attainment of American elementary school pupils. And we wished to do that in a fashion which would separate the valid components of reading comprehension measured by the tests from related, but invalid, components. We have done this. But how accurate and meaningful are these estimates? First, we are constrained by the tests in two distinct fashions:

(1) the items on these tests do not allow refined measurement of reading subskills actually addressed by them (Haertel, 1980), and

(2) these items may also miss major components of the reading process which are rightly called comprehension.

We do not view the former as problematic because we wished to address the reading comprehension process at a more general and socially meaningful level. The latter may be more an issue in the long run but we have no simple way of addressing it in the context of this study. A third threat to the accuracy and meaningfulness of the estimates relates to the above discussion concerning the components of invalidity and the accuracy with which the statistical procedures removes their influence on the comprehension estimates. This issue is not fully resolvable in the absence of further work but we are encouraged by our results.

Finally, assuming that our conceptions and models--at least in outline--are appropriately and correctly focused, what remains to be done? From our perspective, at least three lines of work have positive value:

(1) further theoretical and empirical work which will independently "validate" the components of invalidity which we believe we have "trapped" in our misclassification parameters. E.g., relating independent assessments of vocabulary knowledge to the true positive rates and test-wiseness assessments to the true negative ones;

(2) direct exploration of the implications of the models and analysis of further data to fully articulate the validities of existing tests and their consequences for biases in the assessments of individuals in particular groups or with specific characteristics;

(3) application of the techniques to existing data sets with more desirable characteristics in terms of item selection, age levels, subpopulations, e.g., NAEP data;

(4) the creation of new tests developed to minimize contamination by the components of invalidity isolated by our techniques.

This paper brings to first fruition an analytic schema based on four elements. These involve a conception of skills independent of particular testing devices: the development and application of class of statistical models incorporating qualitative definitions of skill, distorted in item response by errors conceived as misclassifications; a critique and reformation of the concept of test validity--making more concrete and specific the implications of invalidity; and an integration and fusion of these concepts which allows meaningful empirical analyses of item response data. We believe that this conception/model will contribute to the clarification of previously intractable technical and policy issues in the testing field.

REFERENCES

Breglio, V.J., Hinckley, R.H., and Beal, R.S. Students' economic and educational status and selection for compensatory education. Technical Report #2 of The Sustaining Effects Study. Santa Monica, Ca.: System Development Corporation, 1978.

Cochran, W.G. Some methods for strengthening the common chi-square tests. Biometrics, 1954, 10, 417-451.

Cronbach, L.J. Test validation. In Thorndike, R.L. Educational measurement, second edition. Washington, D.C.: American Council on Education, 1971.

Cronbach, L.J. Validity on parole: how can we go straight? New Directions for Testing and Measurement, 1980, 5, 99-108.

Cronbach, L.J. and Gleser, G.C. Psychological tests and personnel decisions. Urbana: University of Illinois Press, 1957.

Cronbach, L.J. and Meehl, P.E. Construct validity in psychological tests. Psychological Bulletin, 1955, 52, 282-300.

Haertel, E. A study of domain heterogeneity and content acquisition. Evanston, Il.: CEMREL, Inc., 1980.

Hansen, M.H., Hurwitz, W.N., and Madow, W.G. Sample survey methods and theory, Volume 1. New York: Wiley, 1953.

Hoepfner, R., Wellisch, J., and Zagorski, H. The sample for the sustaining effects study and projections of its characteristics to the national population. Technical Report #1 of The Sustaining Effects Study. Santa Monica, Ca.: System Development Corporation, 1977.

Murray, J.R. Statistical Models for qualitative data with classification errors. Unpublished doctoral dissertation, University of Chicago, 1971.

Proctor, C.H. A probabilistic formulation and statistical analysis of Guttman scaling. Psychometrika, 1970, 35, 73-78.

Rao, C.R. Linear statistical inference and its applications. New York: Wiley, 1965.

Vernon, P.E. The determinants of reading comprehension. Educational and psychological measurement, 1962, 22, 269-286.

APPENDIX A

Models fo~ Qualitative Data With Misclassifications

Edward Haertel

Stanford University

Models for Qualitative Data With Misclassifications

A.1. Latent States

The skills explored here are each regarded as dichtomous. It is assumed that with respect to each skill, students all belong to one of two categories: those who possess the skill and those who do not. Of course, a student's membership in one or the other category is not observable, but may be inferred from his item responses. These inferences are always subject to error. Thus, the two categories defined by each skill are said to be latent states, and may only be inferred from the student's manifest responses.

When more than one skill is considered at a time, each possible pattern of present and absent skills gives rise to a distinct latent state. As an example, consider two skills, A and B. These could give rise to four latent tates: (1) lacks A lacks B, (2) lacks A had B, (3) has A lacks B, and (4) has A has B. Every student would fall into one of these four patterns, and would conform to exactly one latent state. Just as two skills yield four latent states, three skills could give rise to eight latent states, four skills to sixteen, and so forth.

In general, some latent states may be excluded on theoretical grounds. That is, it may be hypothesized that there are some patterns of presence and absence of skills which will not describe any students at all. In this study, such constraints are expressed as hypothesized hierarchical relationships among skills. Where one skill is logically, psychologically, or chronologically posterior to another, the former is said to be hierarchically related to the latter. Suppose, in the two-skill example, that Skill B is logically dependent on the presence of Skill A. Then Skill B would be said to be hierarchically related to Skill A, and no student would be expected to belong in the latent state "lacks A has B." Under this assumption, only three rather than four states would be required to classify all students. In the absence of any hierarthical constraints, four skills would give rise to sixteen latent states. However, a strict skill hierarchy would prohibit all but five of these skill combinations.

The distribution of skills in a population of students can be described completely by the proportions of students in each latent state. Since every student is in exactly one latent state, these proportions must sum to exactly one.

A.2. Misclassifications

An item's skill requirement is whatever set of skills is required to solve that item. If a set of items with appropriate skill requirements is chosen, student's overt responses to the set of items may be used to classify them into one of a set of manifest states exactly corresponding to the latent states described above. To continue the earlier example, suppose the skill requirement for item 1 consists only of Skill A. Then students can be divided into two manifest states on the basis of their responses to item 1: "lacks A" and "has A." Suppose item 2 has as its requirement Skill B only. Then the four possible patterns of responses to items 1 and 2 would define four manifest states, corresponding to the four latent states earlier.

Obviously, a student's manifest state and his latent state need not correspond. This is due in part to the use of the multiple-choice question format; which affords students the options of guessing or of finding the correct answer by a process of elimination. Even if a free-response format were used, however, item responses would give imperfect information as to students' possession of underlying skills. This is because (1) every item entails unique processing requirements not captured by its skill description; (2) the treatment of skills as unitary entities is an imperfect approximation, thus a student's ability or inability to employ the specific processes required by a single item is an imperfect indicator of his ability to apply related processes; (3) even a student capable of employing the processes required by an item may fail to do so due to lapses in attention, carelessness, etc., and (4) errors in recording the response will sometimes occur, though they should be rare.

The relation between latent states and manifest states is probabilistic. In theory, it is completely described by the set of conditional probabilities of each manifest state being observed, given membership in each latent state. These conditional probabilities are presented in the form of a misclassification matrix.[1] The rows of this matrix correspond to manifest states, and the columns to latent states. The entry in the ith row and the jth column is the conditional probability of a response in the ith manifest state, given conformity to the jth latent state. For the two-item example described earlier,

[1] For a systematic development of misclassification matrices, their properties, and applications, see Sutcliffe (1965a, 1965b).

the misclassification matrix would be as shown in Table A1. Each entry in Table A1 represents the conditional probability of a manifest state given a latent state. For example, the entry in the first row, first column ("$P\ (\overline{A}\overline{B}|\overline{A}\overline{B})$") is read "probability of manifest lacks A, lacks B classification, given latent state lacks A, lacks B." Note that the diagonal entries of the misclassification matrix represent the probabilities of correct classification given each latent state. All off-diagonal entries correspond to misclassification (errors). The entries in each column of a misclassification matrix sum to one.

From Table A1, it would appear that, for the two-item example, three independent conditional probabilities could be specified for each of the four columns of the misclassification matrix. (The fourth entry in each column would be obtained by subtraction, since each column sums to one.) In practice, specification of the misclassification matrix is simplified substantially by the assumption of conditional independence. This assumption is required by virtually every statistical theory of test responses which distinguishes latent from manifest states. It is assumed that within any group of students in the same latent state, the (conditional) distributions of responses to different items are all independent of one another (Lord & Novick, 1968, p. 316). That is to say, the conditional probability of a correct response to any item, given a student's latent state, is the same regardless of his responses to all other items. It is a consequence of this assumption that within any column of the misclassification matrix, i.e., conditional upon any particular latent state, the probability of any pattern of item responses is simply the product of the conditional probabilities of the responses to the separate items.

TABLE A1--Misclassification Matrix for the Two-Item Example

Manifest State	Latent State			
	Lacks A, Lacks B ($\bar{A}\bar{B}$)	Lacks A, Has B ($\bar{A}B$)	Has A, Lacks B ($A\bar{B}$)	Has A, Has B (AB)
Lacks A, Lacks B ($\bar{A}\bar{B}$)	P ($\bar{A}\bar{B}\|\bar{A}\bar{B}$)	P ($\bar{A}\bar{B}\|\bar{A}B$)	P ($\bar{A}\bar{B}\|A\bar{B}$)	P ($\bar{A}\bar{B}\|AB$)
Lacks A, Has B ($\bar{A}B$)	P ($\bar{A}B\|\bar{A}\bar{B}$)	P ($\bar{A}B\|\bar{A}B$)	P ($\bar{A}B\|A\bar{B}$)	P ($\bar{A}B\|AB$)
Has A, Lacks B ($A\bar{B}$)	P ($A\bar{B}\|\bar{A}\bar{B}$)	P ($A\bar{B}\|\bar{A}B$)	P ($A\bar{B}\|A\bar{B}$)	P ($A\bar{B}\|AB$)
Has A, Has B (AB)	P ($AB\|\bar{A}\bar{B}$)	P ($AB\|\bar{A}B$)	P ($AB\|A\bar{B}$)	P ($AB\|AB$)

A second (and more substantive) simplifying assumption is also invoked in specifying misclassification matrices: Misclassification probabilities only vary with the unions of latent states, conforming to or not conforming to the skill combination required by any given item. That is to say, for any item the misclassification probabilities for different latent states depend only upon whether or not all the skills that item requires are present. If a set of latent states are defined using all the skills an item requires (and possibly others as well) then the item's skill requirements can be used to partition those latent states into two categories. In the first would be latent states for which all of the skills the item required were present. In the second would be all latent states for which one or more of the skills the item required were absent. Within each category, misclassification probabilities for all latent states would be the same. The presence or absence of skills not part of the given item's skill requirement is irrelevant, and if the entire set of skills the item requires is not present, it does not matter which or how many of the relevant skills are lacked.

Table A2 shows how the assumption of conditional independence permits simplification of the misclassification matrix for the two-item example. Note that each conditional probability is decomposed into a product of two factors, one for each item. New notation is introduced in Table 3, to make explicit the relation of manifest states to particular items. Again taking the entry in row 1 column 1 as an example, $P\ (\overline{A}_1|\overline{A}\overline{B}) \cdot P\ (\overline{B}_2|\overline{A}\overline{B})$ represents the probability of a manifest "lacks A" classification on item 1 (i.e., marking an incorrect alternative or omitting item 1) given latent state "lacks A, lacks B," times the probability of a manifest "lacks B" classification on item 2 (defined as

TABLE A2--Simplification of Misclassifications Introduced by the Assumption of Conditional Independence

	LATENT STATE			
Manifest State	Lacks A, Lacks B ($\overline{AB}$)	Lacks A, Has B ($\overline{A}B$)	Has A, Lacks B ($A\overline{B}$)	Has A, Has B (AB)
Item 1 - Lacks A, Item 2 - Lacks B ($\overline{A}_1\overline{B}_2$)	$P(\overline{A}_1\|\overline{AB}) \cdot P(\overline{B}_2\|\overline{AB})$	$P(\overline{A}_1\|\overline{A}B) \cdot P(\overline{B}_2\|\overline{A}B)$	$P(\overline{A}_1\|A\overline{B}) \cdot P(\overline{B}_2\|A\overline{B})$	$P(\overline{A}_1\|AB) \cdot P(\overline{B}_2\|AB)$
Item 1 - Lacks A, Item 2 - Has B ($\overline{A}_1B_2$)	$P(\overline{A}_1\|\overline{AB}) \cdot P(B_2\|\overline{AB})$	$P(\overline{A}_1\|\overline{A}B) \cdot P(B_2\|\overline{A}B)$	$P(\overline{A}_1\|A\overline{B}) \cdot P(B_2\|A\overline{B})$	$P(\overline{A}_1\|AB) \cdot P(B_2\|AB)$
Item 1 - Has A, Item 2 - Lacks B ($A_1\overline{B}_2$)	$P(A_1\|\overline{AB}) \cdot P(\overline{B}_2\|\overline{AB})$	$P(A_1\|\overline{A}B) \cdot P(\overline{B}_2\|\overline{A}B)$	$P(A_1\|A\overline{B}) \cdot P(\overline{B}_2\|A\overline{B})$	$P(A_1\|AB) \cdot P(\overline{B}_2\|AB)$
Item 1 - Has A, Item 2 - Has B (A_1B_2)	$P(A_1\|\overline{AB}) \cdot P(B_2\|\overline{AB})$	$P(A_1\|\overline{A}B) \cdot P(B_2\|\overline{A}B)$	$P(A_1\|A\overline{B}) \cdot P(B_2\|A\overline{B})$	$P(A_1\|AB) \cdot P(B_2\|AB)$

before) given latent state "lacks A, lacks B."[1] Within each column, the conditional probabilities for the item 1 manifest states ($\bar{A}_1$ and A_1) must sum to one, as must the conditional probabilities of $\bar{B}_2$ and B_2. Thus, in the matrix shown in Table 3, only two parameters rather than three must be specified for each column.

The effect of invoking the second simplifying assumption, that misclassification probabilities only vary according to the presence or absence of an item's full compliment of required skills, is shown in Table A3. Note that in item 1 the same probabilities appear in the first and second (lacks A) columns, and in the third and fourth (has A) columns. The presence or absence of skill B is irrelevant. Likewise, item 2 factors are the same for the first and third columns, and for the second and fourth. In the first row, for example, the second simplifying assumption implies that $P(\bar{A}_1|\bar{A}\bar{B}) = P(\bar{A}_1|\bar{A}B) = P(\bar{A}_1|\bar{A})$, $P(\bar{A}_1|A\bar{B}) = P(\bar{A}_1|AB) = P(\bar{A}_1|A)$, $P(A_1|\bar{A}\bar{B}) = P(A_1|\bar{A}B) = P(A_1|\bar{A})$, and $P(A_1|A\bar{B}) = P(A_1|AB) = P(A_1|A)$. The conditional probabilities for responses to item 2 ($\bar{B}_2$ and B_2) are similarly simplified. Only four values need be specified to determine the entire matrix illustrated in Table A3. One possible set would be $P(A_1|\bar{A})$, $P(A_1|A)$, $P(B_2|\bar{B})$, and $P(B_2|B)$.

At this point, an algebraic simplification may be introduced. The four-by-four matrix shown in Table A3 turns out to be the Kronecker product of two

[1] Note that where the skill requirements of two or more items overlap, a single response pattern may include conflicting manifest classifications. For example, if a third item requiring only skill B were analyzed along with items 1 and 2, one possible manifest response would be $\bar{A}_1$ $\bar{B}_2$ B_3, i.e., state "lacks A" on item 1, state "lacks B" on item 2, state "has B" on item 3. Given latent state $\bar{A}\bar{B}$, the probability of this manifest state would be $P(\bar{A}_1|\bar{A}\bar{B}) \cdot P(\bar{B}_2|\bar{A}\bar{B}) \cdot P(\bar{B}_3|\bar{A}\bar{B})$, by the assumption of conditional independence.

TABLE A3.--Simplification of Misclassifications Introduced by the Assumption of Invariance Across Irrelevant Skills

	LATENT STATE			
Manifest State	Lacks A, Lacks B ($\bar{A}\bar{B}$)	Lacks A, Has B ($\bar{A}B$)	Has A, Lacks B ($A\bar{B}$)	Has A, Has B (AB)
Item 1 - Lacks A, Item 2 - Lacks B ($\bar{A}_1\bar{B}_2$)	$P(\bar{A}_1 \mid \bar{A}) \cdot P(\bar{B}_2 \mid \bar{B})$	$P(\bar{A}_1 \mid \bar{A}) \cdot P(\bar{B}_2 \mid B)$	$P(\bar{A}_1 \mid A) \cdot P(\bar{B}_2 \mid \bar{B})$	$P(\bar{A}_1 \mid A) \cdot P(\bar{B}_2 \mid B)$
Item 1 - Lacks A, Item 2 - Has B ($\bar{A}_1 B_2$)	$P(\bar{A}_1 \mid \bar{A}) \cdot P(B_2 \mid \bar{B})$	$P(\bar{A}_1 \mid \bar{A}) \cdot P(B_2 \mid B)$	$P(\bar{A}_1 \mid A) \cdot P(B_2 \mid \bar{B})$	$P(\bar{A}_1 \mid A) \cdot P(B_2 \mid B)$
Item 1 - Has A, Item 2 - Lacks B ($A_1\bar{B}_2$)	$P(A_1 \mid \bar{A}) \cdot P(\bar{B}_2 \mid \bar{B})$	$P(A_1 \mid \bar{A}) \cdot P(\bar{B}_2 \mid B)$	$P(A_1 \mid A) \cdot P(\bar{B}_2 \mid \bar{B})$	$P(A_1 \mid A) \cdot P(\bar{B}_2 \mid B)$
Item 1 - Has A, Item 2 - Has B (A_1B_2)	$P(A_1 \mid \bar{A}) \cdot P(B_2 \mid \bar{B})$	$P(A_1 \mid \bar{A}) \cdot P(B_2 \mid B)$	$P(A_1 \mid A) \cdot P(B_2 \mid \bar{B})$	$P(A_1 \mid A) \cdot P(B_2 \mid B)$

two-by-two matrices, each containing parameters for one item.[1] These component matrices are shown in Table A4. Each is itself a misclassification matrix, with columns representing latent states and rows manifest states, the conditional probabilities in each column summing to one, and the diagonal representing correct classifications. Note that, since the two conditional probabilities in each column sum to one, specifying either value in a column determines the other. Thus the entire matrix in Table A3 can be specified given just one value from each column of the matrices in Table A4.

In the general case, the misclassification matrix for any set of two or more items is constructed by forming the Kronecker product of misclassification matrices for the individual items. The dimensionality of these matrices depends upon the scoring used. Since in this study skill specification has focused on the correct response alternative only, items are scored dichotomously (correct/incorrect) and misclassification matrices for individual items have just two rows and two columns.

A.3. The Complete Model

The proportions of students in different latent states and the conditional probabilities in the misclassification matrices together determine the

[1] The Kronecker product (direct product) has wide application in formal algebra, and is used in statistics to represent a variety of factorial structures (Bock, 1975, pp. 273-283; Haberman, 1974, pp. 150-166). Either of these works provides a technical discussion. In the present context, to form the Kronecker product of the two-by-two misclassification matrices for items 1 and 2 and obtain the four-by-four matrix shown in Table A3, the entire matrix for item 2 is multiplied in turn by each element of the matrix for item 1, and the four resulting two-by-two matrices are adjoined in the same arrangement as the elements from the item 1 matrix.

TABLE A4.--Examples of Misclassification Matrices for Two Items

Item 1 (Requires Skill A)			Item 2 (Requires Skill B)		
	Latent State			Latent State	
Manifest State	Lacks A($\bar{A}$)	Has A(A)	Manifest State	Lacks B($\bar{B}$)	Has B(B)
Item 1 - Lacks A ($\bar{A}_1$)	$P(\bar{A}_1 \mid \bar{A})$	$P(\bar{A}_1 \mid A)$	Item 2 - Lacks B ($\bar{B}_2$)	$P(\bar{B}_2 \mid \bar{B})$	$P(\bar{B}_2 \mid B)$
Item 1 - Has A (A_1)	$P(A_1 \mid \bar{A})$	$P(A_1 \mid A)$	Item 2 - Has B (B_2)	$P(B_2 \mid \bar{B})$	$P(B_2 \mid B)$

overall probability of each possible pattern of responses. To illustrate, numerical values will be chosen arbitrarily for the parameters in the two-item example, and the probabilities of each possible pattern of responses will be derived. These arbitrary values are presented in Table A5. Note that the hierarchical relationship between Skill A and Skill B has been assumed in specifying the latent states. Note also that only six numerical values in Table 6 were freely chosen--all others were obtained by subtraction.

Four patterns of responses to items 1 and 2 are possible. Using "+" for correct and "0" for incorrect, these are "00," "0+," "+0," and "++." Consider the probability of a "00" response. For a student in latent state "lacks A, lacks B" the probability of a "00" response is .4550. For the "has A, lacks B" state, the conditional probability of a "00" response is .0325. For the "has A, has B" state, it is .0075. Since the proportions of students in thes three states are .40, .50 and .10 respectively, the overall probability of a "00" response is .40 x .4550 + .50 x .0325 + .10 x .0075, or .1990. Similarly, the probabilities of the "0+," "+0" and "++" response patterns are .1110, .4010, and .2890, respectively.

In the same general fashion, overall probabilities of every possible pattern of responses could be computed for any set of items, given the proportions in each latent state and the matrix of misclassification probabilities.

TABLE A5.--Illustrative Values for Misclassification Parameters in Two-Item Example

Latent State	Proportion
Lacks A, Lacks B	.40
Lacks A, Has B	0[a]
Has A, Lacks B	.50[b]
Has A, Has B	.10[b]

Item 1 Misclassification Matrix

	Latent State	
Manifest State	Lacks A	Has A
Lacks A	.70	.05
Has A	.30[b]	.95[b]

Item 2 Misclassification Matrix

	Latent State	
Manifest State	Lacks B	Has B
Lacks B	.65	.15
Has B	.35[b]	.85[b]

TABLE A5.--Continued

Complete Misclassification Matrix

Manifest State	LATENT STATE			
	Lacks A, Lacks B	Lacks A, Has B	Has A, Lacks B	Has A, Has B
Lacks A, Lacks B	.4550	.1050	.0325	.0075
Lacks A, Has B	.2450	.5950	.0175	.0425
Has A, Lacks B	.1950	.0450	.6175	.1425
Has A, Has B	.1050	.2550	.3325	.8075

Computed Probabilities of Each Possible Manifest Response Pattern

Manifest State for Item 1	Manifest State for Item 2[a]	
	Item 2 - Lacks B	Item 2 - Has B
Item 1 - Lacks A	.1990	.1110
Item 1 - Has A	.4010	.2890

[a] Fixed by hierarchical constraint assumed for Skills A and B.

[b] This value was freely chosen. All values not lettered were obtained by subtraction.

APPENDIX B

Parameter Estimation, Hypothesis Testing, and Precision Assessment for the Models

Edward Haertel

Stanford University

Parameter Estimation, Hypothesis Testing, and Precision Estimation for the Models

B.1. The Estimation Procedure

Once a model for some set of items has been formulated, numerical values can be estimated for the proportions of students in each latent state and for the conditional probabilities in the classification matrices. This is accomplished by the method of maximum likelihood. A detailed description of the procedure used is given in Murray (1971); briefly, it is as follows: As illustrated above (Appendix A), any set of values for the model parameters generates a set of probabilities that students will mark each of the possible (coded) response patterns. Since students' responses are assumed to be (conditionally) independent of one another, the probability that students in particular skill categories will respond in various patterns is simply the product of their separate probabilities of so responding. Thus, using any set of parameter estimates, we can compute the overall probability, or likelihood, of a set of observations.

As an example, suppose that there were two items and four possible response patterns: wrong-wrong, wrong-right, right-wrong, and right-right, wnere rignt represents "has skill" and wrong represents "lacks skill." Suppose also that some set of parameter estimates generated probabilities of .4, .1, .3, and .2, respectively, for these patterns and that when ten students were tested, the frequencies in each pattern were 5, 0, 4, and 1. Probabilities of these students responding as they did are .4 for each of the 5 "wrong-wrong" students, .3 for each of the 4 "right-wrong" students, and .2 for the "right-right" student. The overall likelihood of the obtained data, given the set of parameter estimates generating these probabilities, is $(.4)^5 \cdot (.3)^4 \cdot .2 =$ .000016589. As this likelihood is a function of the parameter estimates, any set of estimates will yield a unique value for the likelihood. The procedure in maximum likelihood estimation is to find a set of parameter estimates which maximizes the value of this function, or, what is usually done, minimizes the negative of the log of the function. Under specifiable conditions, as the number of respondents increases, this strategy will yield--with increasing probability--values which are unique and which have statistically desirable features (Rao, 1965, pp. 289-302).

The maximum likelihood procedure yields several useful statistics in addition to the parameter estimates themselves. These include the likelihood ratio chi-square and the asymptotic covariance matrix of the estimates, from which it is possible to compute their standard errors (Rao, 1965). In large samples, like those used in this study, these statistics can be used to assess the overall fit of the model, and to construct confidence intervals for the value of the parameters. The use of these statistics is further described below, in the section on establishing criteria for goodness of fit.

Finding the maximum likelihood estimates for a given model is, in technical terms, a linearly constrained non-linear function minimization problem. The linear constraints are that all conditional probabilities and latent state proportions must be between zero and one, and that certain subsets of these parameters must sum to unity. The problem is non-linear because a given increment or decrement in the value of a particular parameter will produce quantitatively different changes in response pattern probabilities, depending upon the values of that parameter and others. An algorithm for solving problems of this kind was published by Shanno (1970a, 1970b) and was implemented in the computer program used by Murray (1971). The same program, with minor modifications, was used in this study. Because the number of possible response patterns increases very rapidly with the items considered, the number of items that can be simultaneously analyzed is sharply limited. Experience with the program has indicated that models with up to four items are completely tractable. Models with five items are roughly six times more costly to analyze, but do not exceed the capacity of the program. Models with six items and no more than three skills (eight latent states) can be analyzed, but only at substantial cost, and models with more than six items cannot be solved by the program in its present form.

B.2. Establishing Criteria for Goodness of Fit

As described by Rao (1965), the maximum likelihood estimation procedure yields, if the model is valid, a likelihood ratio chi-square, which is asymptotically distributed as a chi-square on k-1-p degrees of freedom, where k is the number of possible response patterns and p is the number of non-redundant parameters[1] estimated in fitting the model. The chi-squared fit statistic

[1] Certain sets of parameters must sum to unity, e.g., the conditional probabilities of a true positive and of a false negative on the same item, or the probabilities of being in each possible latent state. Since given all but one of the parameters in such a set the last may be obtained by subtraction, each such set is said to contain one redundant parameter. The choice of which parameter to regard as redundant is arbitrary. The number of non-redundant parameters is the number which could be freely chosen.

assesses the likelihood of obtaining the observed data, given that the statistical model is a correct and complete representation of the process giving rise to the data. A large value indicates large departures of the observed data from likely values given the model. Thus, in this application a small chi-square is desirable. The size of the chi-square will also depend upon the size of the sample used, since the likelihood of discrepancies of a given size should be less if more persons are tested. That is to say, if the model specified were correct and complete, as more and more persons were tested the observed proportions of persons manifesting each possible response pattern would come closer and closer to the proportions predicted by the model.

The chi-squared test is sensitive to any lack of correspondence between the predicted and observed proportions for all response patterns. For purposes of this study, however, not all sources of such lack of fit are of equal importance. Incomplete or inaccurate specification of either the latent states or the classification matrices may result in statistically significant lack of fit. Adequate specification of misclassifications was investigated empirically in a preliminary study (Haertel, 1980), and guidelines were developed to minimize lack of fit due to misspecification of misclassifications. Incomplete modeling of latent states, however, may be inevitable.

It is to be expected that in our current state of knowledge, as it relates to the modeling of latent states, some skills will be omitted. The substantive model, which was implemented in our earlier work (Haertel, 1980), included nine skills, and was clearly simpler than the actual processes of reading comprehension. The model used in this study includes only one skill per test level, an additional oversimplification of the total set of processes, but one which apparently reflects what the tests are capable of measuring.

While the effects of an omitted skill required by only one of a set of items may be absorbed into the misclassification specification for that item any omitted skills common to two or more items and possessed by some but not all students will contribute to the lack of fit. Some such effects can be avoided, e.g., by not using more than one item from a passage. Others, however, will remain. The experience of researchers with models of this kind has been that with samples as large as those to be used in this study, non-significant chi-squares are rarely obtained (e.g., see Murray, 1971; Proctor, 1970). In this study, lack of fit may arise not only as a consequence of omitted reading comprehension skills, but also due to failure of the skills included to function as underlying dichotomies. In addition, substantively trivial departures from the predicted response pattern proportions may arise, due to response biases on the part of some children (e.g., a tendency to guess the fourth choice), sex or racial/ethnic differences in the interest level of individual passages, or any other systematic influence upon the responses of a segment of the student population, operating across items. As described below in the section on design effects, data from stratified cluster samples, like those used in this study, can only approximate the characteristics of a simple random sample. While an adjustment for this effect is made, it is necessarily imperfect, and departures from the theoretical assumption of simple random sampling may also perturb the fit statistics in this study.

The sensitivity of the overall chi-square test to omitted skills and the difficulty of obtaining non-significant chi-squares with large samples requires that additional criteria be established for judging the fit of the models. One such criterion is that discrepancies between the fitted and observed response pattern proportions, i.e., residuals, be small. Criteria for the acceptable magnitude of residuals, established on the basis of two early sets of analyses, were used in earlier analyses (Haertel, 1980). In addition to simple differences between observed and predicted proportions (raw residuals), a standardized residual proposed by Cochran (1954) is employed in establishing these criteria. This standardized residual is asymptotically distributed as a normal deviate with zero mean and unit variance.. While it will increase with sample size in much the same way as the likelihood ratio chi-square, it can provide information on whether lack of fit is due to large residuals for a few cells (response patterns) or to moderate residuals in many cells. In the former case, patterns of residuals can provide valuable information on the sources of lack of fit, and can aid in revising the model to bring the overall chi-square down.

B.3. <u>Testing Individual Parameters</u>

In this study, the major hypotheses addressed the existence of specific skills and that specified hierarchical relationships held among them. These hypotheses can be formulated as specifying that certain parameters are or are not equal to zero. A rigorous procedure is available for testing hypotheses of this form.

If two skills are hierarchically related, no student should possess the second who does not possess the first. Thus, the proportions of students in any latent states including the second skill but not the first should be zero.

The hypothesis that two skills are hierarchically related is equivalent, therefore, to a hypothesis that parameters representing proportions of students in latent states corresponding to certain combinations of skill states are equal to zero. If one of the skills used in defining the latent states does not describe a difference among items and among students, then pairs of latent states differing only with respect to that skill may be collapsed. This is (mathematically) equivalent to setting the proportions of students in all latent states for which that skill is present (or absent) to zero. Thus, the hypothesis that a given skill exists can be considered equivalent to the hypothesis that the value is not zero for parameters representing proportions of students in at least one latent state including (or not including) that skill.

To test whether one or more parameters are zero, two models are fitted. In the first model, the parameters to be tested are permitted to take on any values. The second model is exactly like the first, except that the parameters to be tested are forced equal to zero. Since the second model is simply the first with certain constraints, it must necessarily yield a chi-square greater than or equal to that obtained with the first model. It will also have more degrees of freedom--one more degree of freedom for each parameter constrained to equal zero. The arithmetic difference of the likelihood ratio chi-squares for these two models is known as a difference chi-square. It is asympototically distributed as a chi-square on as many degrees of freedom as there were additional constraints imposed in the second model. Even if the overall chi-squares for the two models are both significant, the difference chi-square need not be. It tests the specific hypothesis that the specified parameters are all equal to zero, under the assumption that the other aspects of the model are correct.

B.4. The Problem of Design Effects

The theory on which maximum likelihood estimation and the associated statistics is based assumes a simple random sample from the population of interest. Data to be used in this study, however, represent stratified cluster samples. In obtaining each of these data sets, a universe of schools was first defined, and all schools in the universe were divided into strata according to size, location and other demographic characteristics. Once strata were defined, some schools were randomly sampled within each stratum, and the students tested were all clustered within these selected schools. In comparison to a simple random sample of students, stratification could yield increased precision. The effect of clustering, however, is to reduce precision. This is because observations on students in the same school are correlated. Thus, additional observations taken in the same school contain less new information than observations on students selected at random from the population. In the data used in this study, the net effect of stratification of schools and clustering of students within schools was to decrease precision. As a result, proportions of students manifesting different response patterns are not expected to approximate population proportions as closely as they would in a simple random sample of the same size. While this has no systematic effect on the parameter estimates, it results in an inflation of the likelihood ratio chi-square, and a reduction in the estimated standard errors of the parameters.

A simple method is used to adjust for this effect. In practice the variability of estimates based on a stratified cluster sample of a given size is almost proportional to that of estimates based on a simple random sample of the same size, and very close to that of estimates from a simple random sample of somewhat smaller size. The size of a simple random sample yielding the same precision as the actual stratified cluster sample is called the <u>effective sample size</u>. By substituting the effective sample size for the actual sample size in these analyses, the correct values of chi-squares and standard errors can be approximated. Estimation of the effective sample size for the data to be used in this study is in the main text (4.).

APPENDIX C

Intermediate Estimates for Scaling of Cumulative Skill Level Proportions

Display C.1. Logits of Estimated Latent State Probabilities

Item Set	Grade	≤ B	≤ C	≤ 1	≤ 2
		Logits of Cumulative Latent State Probabilities			
X	1	1.681			
	2	-0.995	0.032		
	3		-0.695	-0.237	
	4			-0.225	0.016
	5			-0.900	-0.731
	6				-0.871
Y	1	1.319			
	2	-0.781	-0.076		
	3		-0.974	-0.044	
	4			-0.008	0.221
	5			-0.641	-0.443
	6				-1.030

Display C.2. Estimated Variances and Covariances of Logits of Latent State Probability Estimates*

Item Set	Grade	first variance	covariance	second variance
		Estimated Dispersions (10^3)		
X	1	1202.55	-	-
	2	13.09	4.66	14.06
	3	4.91	4.90	7.72
	4	12.07	8.68	14.36
	5	14.86	9.57	15.84
	6	60.74	-	-
Y	1	119.66	-	-
	2	4.49	1.81	3.57
	3	5.20	3.26	12.21
	4	5.28	2.98	4.47
	5	5.95	3.24	4.64
	6	19.30	-	-

* The estimated covariance of the logits (f) of two probability estimates is

$$\hat{cov}(f(\hat{p}_1), f(\hat{p}_2)) = \left(\frac{1}{\hat{p}_1(1-\hat{p}_1)}\right)\left(\frac{1}{\hat{p}_2(1-\hat{p}_2)}\right) \quad cov(\hat{p}_1, \hat{p}_2).$$

Display C.3. Logistic Scale Values for Grades and States

		Item Set	
Scale Value	Parameter	X	Y
Baseline	μ	1.681	1.319
Grade			
1	$\alpha 1$	0.000	0.000
2	$\alpha 2$	-2.676	-2.100
3	$\alpha 3$	-3.403	-2.998
4	$\alpha 4$	-3.391	-2.962
5	$\alpha 5$	-4.102	-3.611
6	$\alpha 6$	-4.242	-4.198
State			
<8	$\beta 1$	0.000	0.000
<0	$\beta 2$	1.027	0.705
<1	$\beta 3$	1.485	1.635
<2	$\beta 4$	1.690	1.849

9