Some statistical models and computational methods that may be useful for cognitively-relevant assessment

Brian Junker[1]
Department of Statistics
Carnegie Mellon University
232 Baker Hall
Pittsburgh PA 15213
brian@stat.cmu.edu

Prepared for the Committee on the Foundations of Assessment, National Research Council, November 30, 1999

[1]On leave at Learning Research and Development Center (LRDC), 3939 O'Hara Street, University of Pittsburgh, Pittsburgh PA 15260. I would like to thank a number of people, both for general conversations on cognitively relevant assessment models, and in some cases also for specific and copious comments on an earlier draft of this report. These include Gail Baxter, Pam Beck, Lou DiBello, Al Corbett, Karen Draney, Bob Glaser, Ken Koedinger, Bob Mislevy, Peter Pirolli, Lauren Resnick, Steve Roudenbush, Bill Stout, and Mark Wilson. In addition, I was very fortunate to be an occasional visitor to a graduate seminar on intelligent tutoring systems at LRDC, led by Kurt VanLehn, during the Fall 1999 semester. Two unpublished manuscripts, Louis Roussos' *Summary and review of cognitive diagnosis models* and Mislevy, Steinberg and Almond's *On the design of complex assessments*, were very influential as I began the task of writing this report; I also appreciated and benefitted from the comments and reactions of members of the Committee on the Foundations of Assessment when I presented a draft of this report to them on October 1, 1999. Many of the people people listed here will be surprised to see their names at all, and others will wonder why they expended so much effort on communicating their views to me, given the small effect it seems to have had on this report. Nevertheless, any special insights contained in this report are due to the influences of these people, and the many errors that remain are mine. Finally I would like to thank the NRC staff, including Naomi Chudowski, Jane Phillips, and John Shepard, for their patience with my on-deadline (and usually past-deadline) writing habits.

Contents

1 Introduction **4**

2 On Designing an Assessment **7**
2.1 Purpose of Assessment 7
2.2 Student Representations 9
2.2.1 Some General Considerations 9
2.2.2 Some Specific Types of Representation 11
2.2.3 Validity, Reliability, Granularity 21
2.3 The Data 22
2.4 Making and Reporting Inferences About Students 26

3 Some Basic Statistical Assessment Models **31**
3.1 Two Simple Discrete-Attributes Models 31
3.1.1 A Deterministic Assessment Model 31
3.1.2 A Simple Stochastic Assessment Model 32
3.2 Item Response Theory 33
3.3 Two-Way Hierarchical Structure 35

4 Extensions of the Basic Models **38**
4.1 The Linear Logistic Test Model 38
4.2 Modeling Multidimensional Proficiencies 39
4.2.1 Multidimensional, Compensatory IRT Models 39
4.2.2 Multidimensional, Non-Compensatory IRT Models 41
4.3 Other Discrete-Attribute Approaches 42
4.3.1 Latent Class Models 42
4.3.2 The Haertel/Wiley Restricted Latent Class Model 43
4.3.3 The HYBRID Model 45
4.3.4 DiBello and Stout's "Unified Model" 45

5 Case Study: Two Approaches to Cognitive Assessment **46**
5.1 The Corbett/Anderson/O'Brien Model 47
5.2 The Draney/Pirolli/Wilson Model 50

6 Mixtures of Strategies **52**

7 Some Concluding Remarks **55**

8 References **57**

Appendices **67**

A Some Estimation Methods for the LLTM and other IRT-like Models **67**
A.1 Conditional Maximum Likelihood 67
A.2 Marginal Maximum Likelihood and the E-M Algorithm 67
A.3 Markov Chain Monte Carlo 69

B Talk given to the Committee, 1 October 1999 **70**
B.1 Preface 71
B.2 Outline of talk 71
B.3 Bayes and MCMC: The Cliff's Notes 72
B.3.1 Data 72
B.3.2 Parameters 72
B.4 Item Response Models 73
B.4.1 Item Response Models: LLTM 74
B.4.2 Item Response Models: Simple IRT vs. LLTM 75
B.5 Latent Class Models 76
B.5.1 Latent Class Models 76
B.5.2 Latent Class Models: Haertel and Wiley 78
B.6 Corbett/Anderson/O'Brien 78
B.7 Closing Thoughts 80

1 Introduction

This report surveys a few statistical models and computational methods that might be used to underlie an assessment system designed to yield inferences that are relevant to a rich cognitive theory of instruction, learning or achievement. That is to say, we are interested in a deeper account of features of student performance than which items or tasks a student got right, and we are interested in a richer description of these features than a number-right score.

The models discussed here mostly arose in the effort to build a practical methodology for cognitive diagnosis in the information-processing style of cognitive psychology developed by Pittsburgh-based cognitive scientists Newell, Simon, and Anderson, and their intellectual descendants. There are two reasons for this focus: first, many examples of this style are within my ready reach. Second, this style—especially in its application to intelligent tutoring systems—is particularly aggressive in making explicit the links between an underlying theory of performance and the observable data that we can obtain by watching students perform tasks. This aggressive explication makes it easy to lay bare some issues in assessment and link them to statistical issues, both of which are essential for us understand as we face the more complex demands of cognitively relevant assessment.

The reader who is not an "information processor" should not be put off by the heavy emphasis on assessment and ITS's in this style. Probabilistic reasoning has been highly successful here, in explicating issues of what it is important to assess, what sources of uncertainty in assessment might be present, what the expected patterns in data are, etc. The same kind of efforts that led to these advances in cognitive diagnosis can be brought to bear on assessment from any other psychological perspective as well. The nature of the student representation, the kinds of evidence, and the details statistical models and methods may be very different in their particulars. Regardless of one's psychological perspective, the challenges of designing an assessment are the same: one must consider how one wants to frame inferences about students, what data one needs to see, how one arranges situations to get the pertinent data, and how one justifies reasoning from the data to inferences about the student.

All of the models discussed in this report, from factor analysis to item response theory (IRT) models, latent class models, Bayesian networks and beyond, should be thought of as special cases of the same hierarchical model-building framework, well-illustrated recently by the textbook of Gelman, Carlin, Stern and Rubin (1995). They are models with latent variables of persistent interest, which may vary in nature—continuous, ordered, dichotomous, etc.—as well as in relationship to one another depending on the cognitive theory that drives the model, that are posited to drive probabilities of observations, which may also vary in nature and interrelationship. The functions that link observations to latent variables vary as appropriate to the nature and interrelationships of the observed and latent variables, and the models are made tractable by many assumptions of conditional independence, especially between observations given latent variables.

In the past these various models have been treated and discussed as if they were quite distinct. This is due in part to the fact that until recently computational methods have lagged far behind

model building so that for years the sometimes ideosycratic estimation methods first seen as making each model tractible in applications have "stuck" to the model, enhancing the appearance of distinction among the models; and in part to the historical accident that these various models were originally proposed to solve rather different-sounding problems. But as Roudenbush (1999) handily illustrated in his written materials for the Committee at its October 1999 meeting, models with rather dissimilar names and apparently different domains of application can and do amount to the same mathematical object when decontextualized from estimation method and original domain of application. Rapid progress over the past two decades in computational methods fueled by faster computing machinery, and a better sense of the wide applicability of a core methodology to problems from human demography to theoretical physics fueled by a revolution in communication within and between the disciplines, has encouraged and confirmed the view that most statistical models arise from a single framework in which the model is built up from a hierarchy of conditional probability statements.

Two quick examples illustrate the flexibility that this view of modeling provides. Fienberg, Johnson and Junker (1999; see also Dobra, 1999) illustrate how IRT models, well-known in large scale educational assessments, can be readily applied to solve difficult problems in multiple-recapture censuses, for example in assessing the size of the World Wide Web by comparing the "hits" at various search engines for the same queries. Patz, Junker and Johnson (1999) adopt the entire conditional-probability structure implicit in generalizability theory and replace that theory's traditional normal-distribution assumptions with assumptions appropriate to fallible scoring of individual items by multiple judges *according to discrete scoring rubrics*. The resulting model for analyzing multiple ratings of performance items can better account for per-item uncertaintly in the rating process, and can produce a "consensus" rating on each performance, using the same scoring rubric that the judges themselves use, weighted according to the biases and internal reliabilities of the particular judges who saw that performance.

Within the hierarchical model building framework just sketched, this report tries to illustrate the continuum from IRT-like statistical models, that focus on a simple theory of student performance involving one or a few continuous latent variables coding for "general propensity to do well" on the one hand, to statistical models embodying a more complex theory of performance, involving many discrete latent variables coding for different skills, pieces of knowledge, and other features of cognition underlying observable student performance, on the other. Some of the most interesting work on these latter types of models has been done in the context of intelligent tutoring systems (ITS's), and related diagnostic systems for human teachers, where a finer-grained model of student proficiency is often needed to guide the tutor's next steps. However, this extra detail in modeling can mean that no one inference is made very reliably. While this may be an acceptable price to pay in an ITS where the cost of underestimating a student's skills may be low (perhaps costing only the time it takes the student to successfully complete one or two more tasks accessing a particular skill, to raise the ITS's confidence that that particular skill has been mastered), low reliability is not acceptable in high-stakes, limited testing-time assessments, in which e.g. attendance in summer school, job or grade promotions, and entrance into college or other educational programs, may

be affected by inferences about the presence or absence of particular skills and knowledge. We shall also see that the more detailed models often increase the computational burden of inferences, whatever their reliability.

This report deals mostly with full specifications of reasonably complex statistical assessment models. Taken at face value, fitting these models and drawing inferences from data with them requires at least a powerful personal computer and rather complex software; and may in some cases require complex coding of students' behavior as they perform assessment tasks. This would seem to constrain the possible applications of the models. In-class and other embedded assessments that require immediate, teacher-scored feedback, for example, might not be able to use the full power of such models. And in some situations, the technical complexity of the relationship between task performance and inferences about skills and knowledge possessed by students, would create potential political and litigation hurdles in their implementation. I focus on these fully-realized models for two reasons. First, these more complex models come closer to what we may be aiming for when we try to incorporate a more complex theory of performance into the measurement problem. Considering them in full detail allows us to ask what inferences are possible or reliable, taking such a theory seriously. Second, it is certainly true that in many situations—especially the teacher-scored embedded assessments alluded to above—simpler summaries of student behavior would be necessary. These summaries might be based on qualitative scoring rubrics, for example, that are not directly related to the models considered in this report; or they might be based on simplified versions of these complex models, that allow us to use the data in a more straightforward way to make less refined inferences. In either case, these simpler summaries derive validity only to the extent that they reflect variation predicted by the underlying theory of performance; the more complex statistical models considered here provide a vehicle for formulating such predictions.

In developing this report I have made two simplifying assumptions in the statistical modeling of student performance, to ease my own expositional burden. First, I have assumed that all features of student performance in a (generic) cognitive model can be coded by binary present/absent variables. Thus, a piece of knowledge such as the name of the capital of Brazil, or a skill such as factoring the exponent of a sum into a product of the exponentiated summands, is coded as either present or absent, instead of present to some degree, in the models. Second, and perhaps more seriously, I have assumed that all models—IRT-based, cognitively motivated, or otherwise—are only sensitive to binary outcomes, that is, whether the student got each task or subtask right or wrong. As I shall argue at some points below, models built on only these binary task performance measures may not always be able to gather enough evidence from a typical-length student exam, to allow reliable inferences about complex underlying features of student performance, from particular skills and knowledge to students' choices of strageties from problem to problem or over a whole sequence of problems.

Neither of these simplifying assumptions—"binary skills" and "binary outcomes"—is sacrosanct and both might be sacrificed to meet the features of the task domain and underlying theory of cognition, as well as the the inferential substance and purposes, of a particular assessment system. On the other hand, relaxing these assumptions may lead to a model with more complex com-

putational needs, or one in which inferences are less reliable. Balancing the detail of modeling suggested by the purposes of assessment and theory of student performance on the one hand, with the computational tractibility and inferential reliability of the resulting model on the other, is one of the arts of assessment modeling. This report illustrates some considerations that inform that art.

2 On Designing an Assessment

We begin our survey of statistical models and computational methods for cognitively-relevant assessment by briefly considering several major decisions in designing the assessment to which these models and methods may be applied. It may seem strange to start a survey of statistical models and computational methods with assessment design, but this is exactly the right place to start. Statistical models for assessment data vary greatly in complexity, and computational methods for these models also vary in both complexity and running time. Relatively simple assessment goals, even in the context of cognitive modeling, require only relatively assessment designs and simple models and computational methods; this often means that inferences are available almost as soon as we have the data—"in real time". On the other hand, complex assessment goals require more complex assessment designs, and hence more complex models and algorithms; consequently inferences will only be available "offline"—too long after the data has been generated and collected to provide immediate feedback to students and teachers, for example.

Design of an assessment involves answering at least the following questions: What is the purpose of the assessment? What kinds of inferences do we wish to make about the student? What data will we need to see to make these inferences, and how can we arrange to get it? How will we make and report these inferences? In the next several subsections we will consider briefly some aspects of these questions. This discussion will not be complete but is intended to highlight some important connections between assessment design, assessment model complexity, computation for model fitting, and reliability of inferences.

2.1 Purpose of Assessment

The traditional purposes assessments to which psychometric methods have been applied have been for linear ranking and related mastery and selection decisions, using a single measure of proficiency like a number-correct score or a latent trait estimate from an item response model. Johnny is "more proficient" than Suzy, and Suzy is more proficient than Bill. Johnny's proficiency is so high that he has achieved mastery of the subject; and while Suzy has not achieved mastery, she and Johnny both make good candidates for training program X. Bill's level of proficiency, however, is too low for admission to this program.

Although mastery in the sense of having a level of proficiency that is above some threshhold is certainly part of the traditional psychometric toolkit, both formal and informal analysis of the

attributes[2] needed to achieve mastery shows considerable complexity in what the student[3] actually brings to the exam. Figure 14.5 of Tatsuoka (1995), for example, lists frequent knowledge states corresponding to θ and scaled SAT scores on a form of the Scholastic Aptitude Test (SAT) Mathematics Exam. Tatsuoka's work particularly shows (see also Section 2.4 below) how data deviations from a conventional psychometric (IRT) model can exhibit considerable structure with respect to putative skills underlying the exam problems themselves. This suggests a more nuanced version of mastery, in which we attend to standards of performance on particular skills underlying the exam problems, instead of number-right or patterns of rights and wrongs on the problems themselves. The more nuanced version of mastery in turn demands a more nuanced and complex model for assessing student performance.

Which of the above "conceptions" of an assessment model—a single continous proficiency variable or many discrete student attributes working in concert—is right? The answer to this question depends partly on what is "really going on in our heads", illuminated for example by basic research on the neural basis of cognition. But it also depends on the purpose of the assessment. If our purpose is to produce linear rankings along one or a few scales—for selection or certification, for overall program evaluation with pre- and post-tests, or more generally to map the progress of students through a curriculum via several equated tests given at various intervals throughout the curriculum, etc.—then a multitude of discrete student attributes probably gets in the way of the story we want to tell; moreover inferences for a single latent proficiency variable will be more reliable, given the same amount of data from a well-constructed test, than inferences for a large collection of discrete student attributes.

On the other hand, assessment can also be used to inform instruction, in order to enhance student learning. Indeed there is substantial argument and evidence, as summarized for example by Bloom (1984), that part of what distinguishes higher student achievement in "mastery learning" and individualized tutoring settings as opposed to the conventional classroom, is the use of frequent and relatively unobtrusive formative tests coupled with feedback for the students and corrective interventions for the instructor, and followup tests to determine how much the interventions helped. This approach continues to be advocated as part of a natural and effective apprenticeship style of human instruction (e.g. Gardner, 1992), and it is the basis of many computer-based intelligent tutoring systems (ITS's; e.g. Anderson, 1993; and more broadly Shute and Psotka, 1996). This kind of assessment often requires the most complex models of student performance, and may also require complex estimation methods.

Another purpose of assessment is to refine our understanding, not of individual differences in

[2]Different cognitive models call the attributes that affect task performance different things. Some models simply call them rules or subgoals, and in some cases subgoals are collections of rules. Many, notably the tutors described by Anderson, Corbett, Koedinger, and Pelletier (1995), make a procedural/declarative distinction. Others, for example Gertner, Conati and van Lehn (1998), make a finer distinction. To minimize terminology I will refer to all of these as simply "student attributes" or occasonally "skills and knowledge".

[3]Not every person who takes an exam is a student, so "examinee" might be a more precise word choice. However in this report I will use the terms "student" and "examinee" interchangeably.

student performance per se, but rather of how a particular set of problems measure mastery of a particular set of skills that we are interested in, and how we might construct problems that measure particular skills of interest. That is to say, we may be most interested in test design (Mislevy, Sheehan, and Wingersky, 1993; Embretson, 1994, 1995b; Baxter and Glaser, 1998; Nichols and Sugrue, 1999) or in real-time generation of novel exam problems (e.g., Embretson, 1998, 1999). To the extent that our model of student performance is complex, statistical models for this purpose will be complex also.

Psychometric methods from true score theory (Lord and Novick, 1968) to item response theory (IRT; van der Linden and Hambleton, 1997) have been honed to characterize and minimize the error in linear ranking, mastery and selection decisions. For these purposes, a new approach to assessment modeling is not needed. For purposes that require a more complex model of individual student attributes underlying performance, we must at least consider the computational burden and reliability of inferences that such models imply, especially if real time feed back is intended. Exact forms of computational methods for complex models may run too slowly to be of use (e.g. Hrycek, 1990), even if the feedback does not have to occur in real time. ITS's and other embedded assessment systems often avoid this difficulty by either running simpler approximations to the "right" model, or by precalibrating parts of the model (this occurs in Computerized Adaptive Testing as well; see for example Wainer et al., 1990; and Sands, Waters, and McBride, 1997). These methods can run the risk of being more "sure" of the student's knowledge state than is in fact justified, but are offset by either the low cost of uncertainty (e.g. in ITS's) or the availability of a wide variety of informal corroborating contextual information (e.g. in classroom-embedded assessment).

More detailed consideration of purposes of assessment that might require cognitively rich modeling is given by Lesgold, Lajoie, Loga, and Eggan (1990). They list five specific purposes for testing: to develop a description of student capabilities; to explain a student's level of performance; to adapt instruction to a student's competences; to screen, for the purposes of warning or remediation, low-achieving students; and to screen, for the purposes of selection, high achieving students. The specific details of one's map of assessment purposes are not as important as that one considers the purpose of the assessment and tries to determine what modeling complexity/effort is useful for that purpose.

2.2 Student Representations

2.2.1 Some General Considerations

As Hunt (1995, p. 415) notes, the first step in developing an assessment is to decide on a representation of the examinee's knowledge, capabilities, etc: All inferences that we can make from the assessment are ultimately framed in term of this representation.

Hunt and other authors going back to at least Cronbach (1957) distinguish between *psychometric* and *cognitive science* representations[4]. Psychometric representations are usually characterized

[4] Pellegrino, Baxter and Glaser (1999) review the history of these two approaches to psychology, and various

by a general trait orientation to what test items tell us about underlying student achievement, reflected by rather coarse, usually continuous, measures of student proficiency; see for example the edited volume on IRT modeling of van der Linden and Hambleton (1997). By contrast, cognitive science representations are often characterized in terms of theory-driven lists of student beliefs and capabilities, and correspondingly fine-grained, discrete measures of student attributes (e.g. presence of absence of a particular bit of knowledge, success or failure mastering this or that skill, etc.); well represented discrete skills models often implemented with Bayesian networks of discrete variables; see for example the edited volume on cognitively diagnostic assessement of Nichols, Chipman and Brennan (1995).

This dichotomy between continuous proficiency approaches associated with traditional psychometrics on the one hand, and discrete attributes approaches recently associated with recent attempts at cognitively diagnostic assessments on the other, is perhaps too sharply drawn to capture the ways in which these representations have stretched toward each other, but it is useful because it allows us to make several important observations. First, statistical methods and probabilistic reasoning are useful regardless of one's psychological perspective—from the trait and behaviorist perspectives with which the continuous proficiency approach of IRT models and their relatives have been associated, to the cognitive diagnosis perspective for which the discrete attributes approach of discrete Bayesian networks seem natural—because much of educational and psychological assessment is a signal-detection problem: we wish to draw inferences about constructs that may or may not be directly observable from behavior that is directly observable, but may noisy or incomplete, or both, as an operationalization of the construct we wish to learn about.

The second important observation to make is one that we began in Section 2.1: the choice between continuous proficiency and discrete attribute representations is driven at least as much by the purposes of the assessment as it is by some physical or psychological reality. Despite the persistence of the "latent trait" terminology in their work, few psychometricians today believe believe that the latent continuous proficiency variable in an IRT model has any deep reality as a "trait"; but as a vehicle for efficiently summarizing, ranking and selecting based on performance in a domain, latent proficiency can be quite useful. By the same token, computer-based intelligent tutoring systems often implement a kind of ongoing student assessment in terms of discrete attributes, usually neither as fine-grained as a fully-developed theory of performance nor as efficient a summary as an overall proficiency measure would be; yet these "middle-level" models may be well-tuned for making inferences about what the tutor should do next with a particular student. There is in fact a wide range of options between the continuous proficiency and discrete attributes extremes of assessment modeling, providing a variety of choices for a variety of assessment purposes.

Our third observation is that the activity of developing a student representation, as an informed collaboration between cognitive psychologist and psychometrician/statistician, provides fertile ground for cognitive psychology to provide a stronger substantive foundation for psycho-

attempts to blend them; see also Snow and Lohman, 1989, p. 264, for a brief discussion. The same distinctions are echoed today for example by O'Connor (1992, pp. 16–21), Embretson (1998), and Nichols and Sugrue (1999).

metrics, and for psychometrics to provide principles of efficient summarization and reliability of inference for cognitively valid assessment. As Snow and Lohman (1989, p. 266) outline, cognitive psychology can contribute to test development by providing a more nuanced picture of what makes a test item difficult and what produces noticable individual differences in student performance; by elaborating our understanding of target aptitudes, achievements and content domains that might suggest alternative data collection and measurement strategies; and by developing new theory-based perspectives on the goals, requirements, demands and conditions of educational environments. Conversely, Mislevy (1994) gives many examples of the ways in which psychometrics and the broader statistical enterprise on which it is based can contribute to the sharpening of construct development, operationalization and measurement in cognitive psychology. These include issues such as the decreasing reliability of inferences when the assessment model fails to track important variation in the underlying performance data (e.g. Mislevy, 1994, p. 468); tradeoffs between increasing complexity of the student model and decreasing precision of measurement, due either directly to model complexity (as is clear to anyone who has watched the standard error of a regression coefficient increase as additional regressors are added to the model) or indirectly to the difficulty of precisely measuring covariates needed in the more complex model (Mislevy, 1994, esp. p. 474); and the tools to identify deviations from a model that accounts of the main components of variation in multi-faceted measurement, and to examine deviations from the model for regularities which suggest either elaboration of the model or quality control issues in the data collection and measurement design (Mislevy, 1994, pp. 476-480).

Despite great progress that has been made since Snow and Lohman's (1989) somewhat speculative invitation to psychometricians, and even more recently since Mislevy's (1994) demonstration of some of the synergy that is possible when a statistician becomes deeply involved in cognitive measurement problems, there is still much work to do. Questions that are simple in other contexts, such as sample size to produce suitably precise inferences, be only approachable by approximate analysis or lengthy simulation, and may depend delicately on details of the statistical model intended to more nearly reflect a complex cognitive theory of performance. And selecting the right "granularity" both in the student attributes underlying performance and in the performance data itself, is still more art than science. Thus for the near future at least, we are not in a position to be able to recommend that one take a "standard" model "off the shelf"—there is no collection of standard models, and perhaps not even a shelf yet. Progress in developing student representations that have both sufficient cognitive validity and adequate psychometric utility will continue to require the full intellectual participation of both the cognitive specialist and the statistician.

2.2.2 Some Specific Types of Representation

We now turn to a brief menagerie of student representations that can form touchstones for the work to follow. As indicated above, these do not represent any sort of "shelf" of "standard" models, they are simply some extremes, and variations on the extremes, that suggest both the variety of models available and the points of commonality among them.

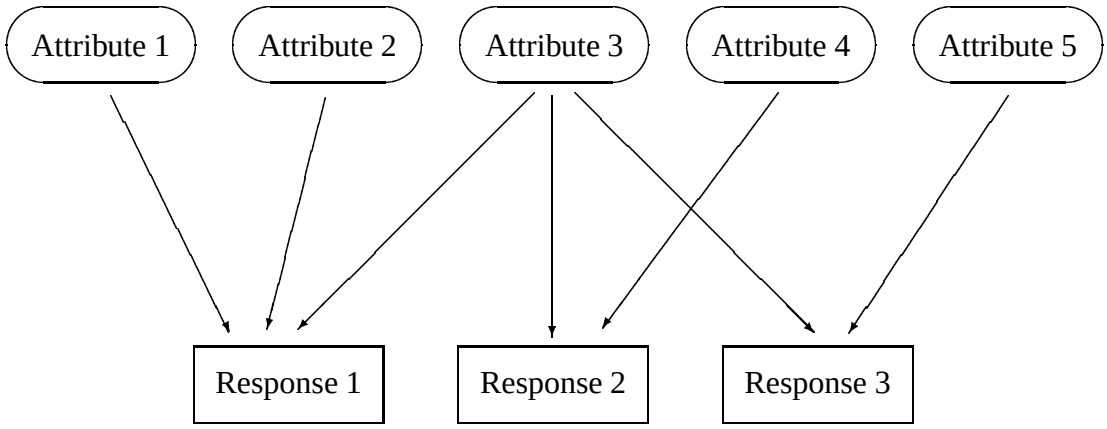


Figure 1: Typical discrete-attributes assessment model; boxes indicate observable student variables, ovals indicate latent student variables. Quality of student responses provides evidence for the presence/absence, or degree of presence, of each of various student attributes. The Q-matrix described in equation (1) indicates presence ($q_{jk} = 1$) or absence ($q_{jk} = 0$) of an edge connecting attribute k to response j.

Our first representation, illustrated by Figure 1, is what I consider to be the basic "discrete attributes" representation suggested by any theory of performance that hypothesizes a discrete set of student attributes underlying a student's responses in a series of assessment tasks. If the attributes are coded in binary fashion (present/absent), the representation depicted in Figure 1 allows for $2^5 = 32$ student knowledge states, corresponding to each possible combination of present or absent attributes. Arrows indicate dependency of responses on student attribute: response 1 for example depends on the first three attributes. The responses themselves may be in one-to-one correspondence with distinct assessment tasks, they may be nested within tasks so that Response 1 and Response 2 relate to the same task but Response 3 relates to a different task, etc.

How the attributes combine to produce a response is left unclear. Two simple possibilities are *pure conjunctions* in which all antecedent attributes have to be present together in order to produce a correct response; or *pure disjunctions* in which any one or more attributes is sufficient to produce a correct response. For ease of exposition, many of the examples presented in this report will involve conjunctive models relating binarily-coded attributes (present/absent) to binarily-coded responses (right/wrong), but there is no compelling theoretical reason—though there may be practical implementation issues—restricting models to conjunctive form or binary coding.

Figure 1, interpreted as a conjunctive model with binary coding of attributes and responses, already illustrates three important points about measurement in this framework. First, when more than one attribute is involved in a response, there is an inherent "credit/blame" problem: For ex-

ample, suppose that we observe only Response 3 for a particular student, and that response is incorrect. We do not know whether to blame Attribute 3, Attribute 5, or both (similarly, in a purely disjunctive model we would not know whether to credit Attribute 3 or Attribute 5 for a correct Response 3). A related problem is the "hiding" of one attribute behind another. Suppose we know the student does not possess Attribute 5. A wrong Response 3 tells us nothing about Attribute 3, the other attribute required for this task. Both problems may be alleviated by observing a correct Response 2. This tells us that Attributes 3 and 4 are present, so that Response 3 is being driven completely by Attribute 5. However, an incorrect Response 2 does not help, regardless of whether we have also observed Response 1 or not. This suggests both that careful design of the assessment may be able to alleviate some of these problems, and that such careful design may be difficult.

Second, the relationships between responses and attributes displayed in the figure can be compactly described in matrix form, by a matrix Q with elements $q_{jk} = 1$ if attribute k is needed for response j, and $q_{jk} = 0$ if not. For example, the Q matrix for Figure 1 would be

$$Q = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \end{bmatrix} , \tag{1}$$

indicating directed edges (arrows) connecting attribute nodes (columns of the Q-matrix) with response nodes (rows of the Q-matrix). The q_{jk}'s might be 0's and 1's as shown here, or they might be given different values indicating the strength of association between the attribute and the response. The Q-matrix is thus essentially an accounting device that describes the "experimental design" of the tasks or responses in terms of underlying attributes that the responses are intended to be sensitive to. Although it has gained prominence in recent years as a tool for task analysis in the work of Tatsuoka (e.g. Tatsuoka, 1990, 1995), it or something like it, would be present in any well specified model of task performance in terms of underyling student attributes or task features.

Third, the arrows in Figure 1 may represent deterministic connections between attributes and response variables, or probabilistic ones. The direction of the arrows in these models indicates the flow of information when we use the model to predict response data. Thus, under a purely conjunctive version of Figure 1 we might predict from the knowledge that a particular student possessed Attributes 1, 2, and 3, but not Attributes 4 and 5, that only Responses 1 and 2 would be correct (or would be correct with high probability), and Response 3 would be incorrect (or correct with low probability). In applying the model to assessment data, the information flows in the other direction. This is especially evident when the arrows represent deterministic connections: if Response 1 is correct and the model is purely conjunctive we may be sure that the student has Attributes 1, 2 and 3. It also works when the connections are probabilistic, and the graph is interpreted as a Bayesian network[5] (which in this case means that all responses are conditionally independent given all attributes and all attributes are marginally independent); in that case we

[5] each variable is conditionally independent of all of its non-descendants, given its parents (Pearl, 1988, Corollary 4, p. 120).

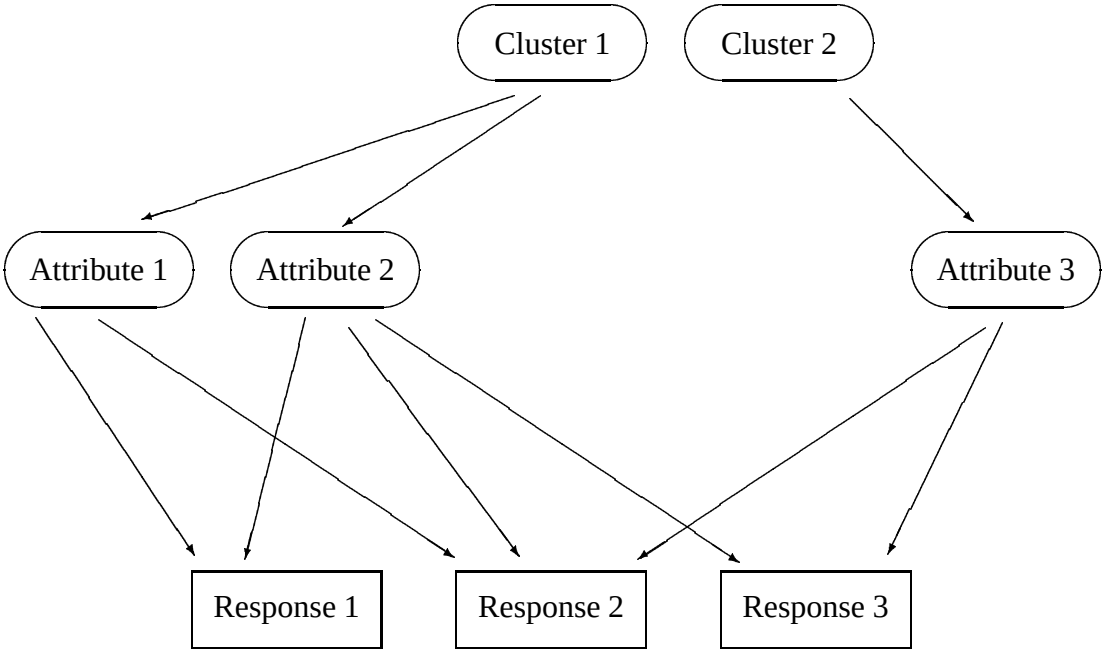


Figure 2: Multi-layered discrete-attributes assessment model; boxes indicate observable student variables, ovals indicate latent student variables. Presence/absence (or degree of presence) of attributes gives evidence about higher-order clusters or meta-attributes within each student. The Q-matrix and hierarchical modeling frameworks described in the text can accomodate this structure as well.

---

start with base rates or prior probabilities for possession of each attribute, and apply the rules of probability, and especially Bayes' rule (see Appendix B for details), to adjust each attribute's base rates upward (e.g. in the case of correct responses) or downward (e.g. in the case of incorrect responses), to obtain posterior probabilities that each particular student posseses the attribute.

When the arrows represent probabilistic connections, the relevant probabilities might be assigned on the basis of prior theory or experience, or estimated directly from pilot study data, or even from the same assessment data used to evaluate student performances. The probabilities might be further structured by a model of how attributes combine to produce responses (not only the conjunctive/disjunctive distinction made above but also whether for example Attribute 5 is easier to acquire if Attribute 3 is already in place).

Figure 2 displays a natural variation on the discrete student attributes representation. Let us interpret the edges of the graph probabilistically; again we assume that all responses are independent given all attributes, and in this case also all attributes are independent given all cluster variables. The idea of the cluster variables is to tie together attributes that tend to be learned together, for

example Cluster 1 in the figure might represent "mixed number skills", Attribute 1 might be the skill "separate whole number from fraction" and Attribute 2 might be the skill "simplify improper fraction to mixed number form"[6]. We might speculate that Attribute 2 is present within a student, then there is a greater likelihood that Attribute 1 is present as well, and conversely—perhaps because these two skills are taught in the same unit in most mathematics curricula. If we observe that Response 3 is correct for a particular student, successive applications of the rules of conditional probability would increase the probability that that Attribute 2 is present, and hence also that the probability that the "mixed numbers skills" cluster is present—perhaps the student has seen a mixed numbers unit in school—which in turn increases the likelihood that Attribute 1 is present, before ever seeing Responses 1 or 2. More direct relationships between attributes can be obtained by connecting skills directly in the graph—say, placing an arrow from Attribute 2 to Attribute 1 and (perhaps) removing Cluster 1 from the graph (thus, knowledge about Attribute 2 directly influences our inferences about Attribute 1 [and vice-versa] without going through a higher-order cluster variable). An example of a model incorporating both kinds of dependence between skills is given in Figures 9 and 10 of Mislevy (1994, pp. 464–465).

As alluded to above, a common interpretation of graphs such as Figure 1 and Figure 2 is as a *Bayesian network*. A Bayesian network is a directed graph representation of a probability distribution in which each variable (each node in the graph) is conditionally independent of all of its non-descendants, given its parents; see Pearl, 1988, Corollary 4, p. 120. In Figure 1 this meant that the responses are conditionally independent of one another given the attributes, and the attributes were independent of one another. In Figure 2 not only are the responses conditionally independent of each other given the attributes, but also the attributes are independent of each other given the clusters, and the clusters are independent of one another. Of course, Bayesian networks are in widespread use in expert systems and as student and user models in ITS's (see for example Jameson, 1995). There is no loss in having more layers in the network than the two in Figure 2, and in fact ITS student models may employ fairly complex multi-layered Bayesian networks; see for example Gertner, Conati and VanLehn (1998). The hierarchical modeling introduced in Section 1 and considered in somewhat more detail in Section 3.3 below, on the one hand, and Bayesian networks on the other, are essentially just two names for the same modeling framework. In artificial intelligence and ITS applications Bayesian networks often consist of all discrete nodes, but there is no reason that they must. All hierarchical models (in the sense if Section 3.3, and all Bayesian networks, have the useful property that information from observing one outcome or subset of outcomes can propagate throughout the network, via the rules of conditional probability, to modify our predictions about outcomes not yet observed.

Figure 3 depicts the student representation of a typical unidimensional IRT model. The discrete attributes of Figures 1 and 2 have been replaced by a single continuous proficiency variable. IRT models have typically been employed in sitations in which the observable responses correspond to

[6] Clearly, there are more than these two skills involved in arithmetic operations with mixed numbers, but these will suffice for illustration.

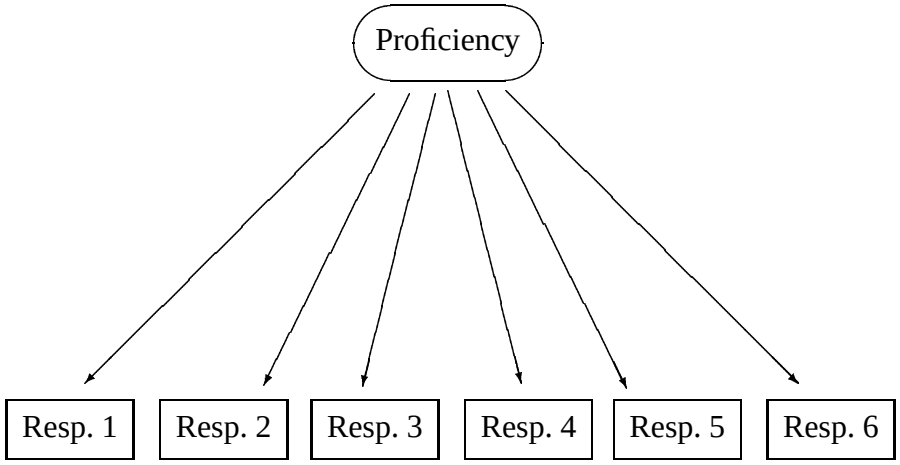


Figure 3: Unidimensional item response theory (IRT) assessment model; boxes indicate observable student variables, oval indicates latent student variable. Discrete student attributes have been replaced with a single continuous "proficiency" variable (proficiency is also a student attribute).

separate assessment tasks or test items, but the observable responses could also be nested within tasks, so that Responses 1 and 2 are related to one task, Responses 3, 4, and 5 to another, etc. Standard IRT models are Bayesian networks, that is, responses are conditionally independent given the proficiency variable. A standard monotonicity assumption—that an increase in the value of the proficiency variable increases the probability of correctly responding to any item—means that proficiency will be closely related[7] to a number-of-responses-correct score.

Figure 4 depicts the student representation of a multidimensional IRT model. The model differs from the IRT model of Figure 3 in that now there are two or more continuous proficiency variables, that may be connected to observable response variables in fairly arbitrary ways. Qualitatively there is little difference between Figure 4 and Figure 1; the underlying reason for this is that both represent one-layer Bayesian networks. The differences are in the details: (a) whether the latent variables are discrete attributes or continuous proficiency variables; (b) the number of latent variables, which is traditionally less in IRT models than in Bayesian network based student models in ITS's for example; and (c) the nature of the connections between latent variables and observable responses. IRT models tend to exhibit more "simple structure", that is, there tend to be clusters of items that depend on only one proficiency variable per cluster (though this would also provide a way out of the credit/blame problem discussed in connection with Figure 1 above), and the predominant multidimensional model is disjunctive—called *compensatory* in the IRT literature—rather than

[7] In the case of binary responses the relationship is a strong probabilistic relationship called *stochastic ordering* (Grayson, 1988; Huynh, 1994); in the case of polytomously-coded responses the relationship appears to be considerably more complex (Hemker, Sijtsma, Molenaar, and Junker, 1997).

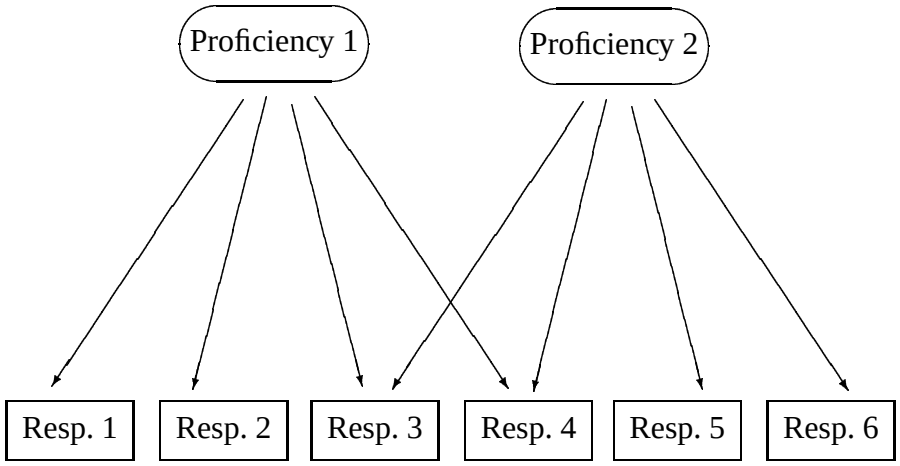


Figure 4: Multidimensional item response theory (IRT) model; boxes indicate observable student variables, ovals indicate latent student variables. Discrete student attributes have been replaced with two or more continuous "proficiency" variables. Qualitatively there is little difference between this figure and Figure 1; this is why a statistician may see little difference between IRT and discrete-attributes models. The primary differences are in the details: (a) whether the latent variables are continuous or discrete; (b) the number of latent variables; and (c) the nature of the connections between latent variables and (coded) observable responses.

conjunctive—called *noncompensatory* in the IRT literature (e.g. Reckase, 1985; Wilson, Wood and Gibbons, 1983; Fraser and MacDonald, 1988; Muraki and Carlson, 1995). A typical trait interpretation of Figure 4 might take Proficiency 1 as proficiency in a part of mathematics, Proficiency 2 as proficiency in verbal tasks, items 1 and 2 as "pure math", items 5 and 6 as "pure verbal" and items 3 and 4 as "word problems". In the compensatory framework, a great deal of mathematical proficiency might overcome verbal deficiencies, or vice-versa, on the "word problems".

Embretson (e.g. Embretson, 1985) has explored the conjunctive version of this model extensively, which she calls the "multicomponent latent trait model" (MLTM). In these models, task performance represented by response 4 for example might involve execution of two different actions[8]. The proficiency variables index the proficiency with which each action is expected to be executed; these proficiencies are translated into probabilities which are then multiplied together to obtain the probability that response 4 is correct. Thus, modeling student performance on a word problem might require sufficient levels of *both* mathematical *and* verbal proficiency, rather than just enough of one to offset a lack of the other.

Finally, Figure 5 illustrates one variation on a class of models that is extremely important for thinking about how to extend the student representations of Figures 1 through 4 to account for situations in which there is substantial qualitative heterogeneity among students performing the assessment tasks. The figure is labelled to suggest a progression of competencies into which we might divide a student population, or through which a single student might pass over time. In this model a student can be in only one of three states of competance. Within a state of competance, the student only has access to the student attributes associated to that state; and can only apply those attributes to each task response variable. Thus a student in the "low competance" state would only have the tools to respond correctly to the third task; a student in the "medium competance" state could respond correctly to all three tasks but would require a different attribute for each task; a student in the "high competance" state could use a single attribute to do all three tasks, but the first task has a twist that requires an extra attribute. Some states of competance might share attributes, but this would unneccessarily complicate the figure. The restricted latent class model of Haertel (1989) and Haertel and Wiley (1995) is similar in structure to this. Clearly, the low/medium/high competance labels on the three knowledge states in Figure 5 could be replaced with less judgemental labels, and the same figure would illustrate the modeling of multiple strategies or states of knowledge that are of roughly equal worth for performing tasks in a particular domain.

The "facets" models of physics performance of Minstrell and Hunt (Levidow, Hunt and McKee, 1991; Minstrell, 1998) are similar to Figure 5, positing up to ten general levels of competance within each of several areas of high school physics; within each competance level Minstrell and colleagues have catalogued[9] "facets", which are specific skills or bits of knowledge that may or may not be sufficiently general to explain a particular physical phenomenon. The large number

[8] Or the presence of two different student attributes, in which case the proficiency variables code the strength of each attribute within the student, for example.

[9] This use of the word "facets" is completely unrelated to the later use of the word in structuring IRT models.

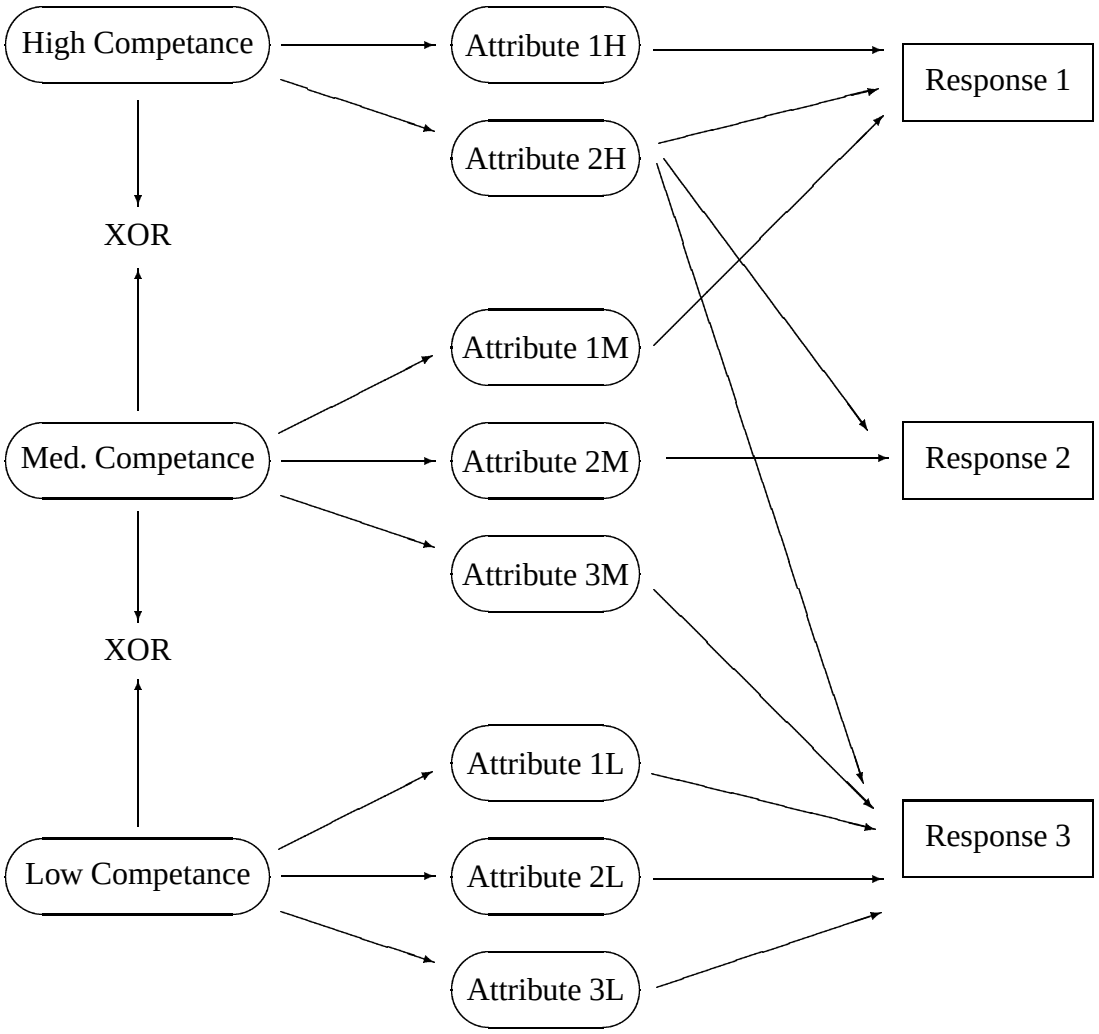


Figure 5: An example of a "multiple-strategies" assessment model. In this model a student can be in only one of three states of competance. Within a state of competance, the student only has access to the student attributes associated to that state; and can only apply those attributes to each task response variable. Thus a student in the "low competance" state would only have the tools to respond correctly to the third task; a student in the "medium competance" state could respond correctly to all three tasks but would require a different attribute for each task; a student in the "high competance" state could use a single attribute to do all three tasks, but the first task has a twist that requires an extra attribute. Some states of competance might share attributes, but this would unnecessarily complicate the figure.

of levels of competance and of student attributes within levels would make this model unwieldy, except that data collection through their computer program DIAGNOSER provides very direct evidence for each attribute and side-steps hiding and credit/blame problems, a point to which we shall return below.

Also note that within each knowledge state in Figure 5 is a version of the discrete attributes model of Figure 1. Each of these discrete attributes models could be replaced with any of the other models of Figures 2 (the multi-layered discrete attributes model), 3 (the unidimensional IRT model) or 4 (the multidimensional IRT model). This allows us to consider radically different ways of approaching problems. For example, a version of Yamamoto's HYBRID model (e.g. Yamamoto and Gitomer, 1993) might add a fourth knowledge state in which the student representation was a unidimensional IRT model. Thus, students who do not fit the low/medium/high mixture of discrete attributes models, might still follow an IRT model in which increasing general proficiency leads to more items right. The IRT model thus provides an interpretation (albeit a cruder one) for at least some of the students who don't fit the competancy progression of Figure 5.

A potential drawback of the structure in Figure 5 is that it does not efficiently allow for changing strategy from one task to the next; rather it postulates that students "come to the exam" with a fixed strategy that will be applied to all tasks. A different structure is required for strategy switching from task to task; essentially the strategy selection process coded by low, medium and high competance on the left in Figure 5 would be replicated within each task or task response, possibly depending on some features of the task itself (so that a student might select a strategy based on which one looks easier for that particular problem. The main challenge here is not in constructing the model, which locally has the same features—may be built out of the same building blocks—as models here for example. Rather, the challenge is to collect data that would be sufficiently informative about student strategy to make per-task estimation of strategy realizable.

From a statistical modeling-building point of view there is little difference between continuous proficiency and discrete attributes representations represented by Figures 1 through 5 (see Section 3.3 and equations (11), and especially Figure 8, below; see also Mislevy, 1994), and indeed there are many assessment models which represent gradations between these extremes (see Sections 3 and 4; as well as Roussos, 1994). The differences arise from details about the granularity of the latent variable representation—for example, a few continous latent variables representing student proficiency in IRT, vs. many discrete latent variables representing specific skills or other student attributes in discrete Bayesian networks—and the consequent assumptions about the conditional distributions that make up the models. A good statistician would not choose either of these extremes—nor, necessarily, any one of the models surveyed below in Sections 3 and 4, nor necessarily any particular one of the models surveyed by Roussos (1994). All of these models give sharp examples of building blocks that one might combine in various ways to accomodate the specifics of a student model, assessment purpose, data availability, etc. in a statistical model of assessment.

2.2.3 Validity, Reliability, Granularity

Two touchstones of traditional psychometrics are validity, the extent to which we are measuring what we intend to measure, and reliability, the extent to which observable responses are well-determined by underlying student variables (attributes or proficiencies), and conversely the extent to which inferences about these latent student variables would be stable if the assessment were repeated. These considerations do not disappear when we move to one of the non-IRT models listed in Section 2.2.2, but how we measure and manipulate them will change, just as the reliability coefficient of classical test theory has given way to item discrimination in some IRT models.

DiBello, Stout and Roussos (1995) list four reasons that an assessment model based on analysis of tasks into component attributes of student performance (skills, bits of knowledge, beliefs, etc.) may not adequately predict students' task performance. The first two may be at least partially interpreted in terms of *validity* issues: Students may *choose a different strategy* for peformance in the task domain than the strategy presumed by the task analysis; or the task analysis may be *incomplete* in that not all attributes required for task performance were uncovered in the task analysis or incorporated into the model. These are threats to a kind of construct validity, to the extent that they indicate areas where the statistical assessment model may not map well onto the underlying cognitive constructs.

The second two reasons may be interpreted at least partially in terms of *reliability*. DiBello, Stout and Roussos (1995) say that a task has *low positivity* for a student attribute, if there is either a high probability that a student who possesses the attribute can fail to perform correctly when the attribute is called for (in this report we will call this probability a *slip probability*), or a high probability that a student who lacks the attribute can still perform correctly when it is called for (we will call this probability a *guessing*[10]*probability*). Other unmodeled deviations from the task analysis model (conceived of as leading to incorrect task performance, such as transcription errors, lapses in student attention, etc.) are collected together under a separate category of slips. Note that both the phenomenon of low positivity and the separate slip category suggest hiding and credit/blame problems: for example, if the guessing probability is high and a student acts correctly when the corresponding attribute is called for, we do not know if it was due to a "guess" or due to correct application of the attribute.

A standard problem in cognitive modeling is the *granularity* of the model. Discussions from two different but complementary perspectives can be found in DiBello et al. (1995) and Martin and VanLehn (1995). Both sets of authors emphasize that, ultimately, grain size is or should be a function of the purpose of the assessment; a crude illustration of this principle is given in Section 2.4 below. DiBello et al. (1995) point out that finer levels of granularity can make the model unwieldy

[10] DiBello and Stout would probably object to this terminology; correct performance by a student who lacks full mastery of a modeled student attribute may not be a guess at all: it may either be the application of an alternate strategy, or the application of a less general version of the attribute, that works in the present case. Similarly, incorrect performance when a student possesses the skill may not be due to a "slip" or transient student error related to the attribute, but rather to poor question wording, transcription errors, etc.

for inference; Martin and VanLehn (1995) develop a Bayesian network-based assessment system at a maximum level of granularity and then provide a reporting interface that allows a human user to coarsen the granularity of the assessment report in fairly arbitrary ways. Another problem not explicitly addressed by these authors is the limit of information available to estimate model parameters as the granularity becomes finer. There are self-correcting mechanisms in ITS's—and perhaps in many formative assessments—that make the low reliability of inferences that one obtains in finely-grained models more tolerable: for example if a student is given too high or too low a score on a particular skill, the student is sent to an inappropriate module that he or she might quickly exit by performing the skill at his/her true level. Similar low reliability is not tolerable in higher-stakes assessments. We shall return to the problem of weak inferences in finely grained models briefly in Section 2.3.

DiBello and Stout view positivity and (in-)completeness as valuable ways to think about manipulating the granularity of the student representations. To increase the grain size of a model based on a careful task analysis, there are two obvious choices. On the one hand, we can drop some student attributes from the model; this leads to incompleteness. Whether this turns out to be important depends on how the purpose of assessment is related to the attributes that were dropped. On the other hand, we can merge together some attributes into a single coarser attribute. This can degrade positivity: if the component attributes of the coarser attributes can still be acquired separately for example, then for tasks that only require one ot two of these components, the "guessing" probability will be higher. Whether this turns out to be important, for example in estimation and prediction, depends on how large the "guessing" parameter turns out to be, and on details fo the topology of the student representation. These manipulations are illustrated in the "unified model" of

DiBello, Stout and Roussos (1995) and DiBello, Jiang and Stout (1999), which incorporates most of the models reviewed in Section 2.2.2 as special cases, as well as all of the notions we have discussed here, as special cases.

As discrete-attributes and related models become more common, it will be necessary to broaden our notions of reliability and validity. The discussion here suggests that some aspects of reliability and validity may be tied up with model grainsize—at least in discrete attributes models that allow us to radically manipulate grainsize—in ways that have not been obvious in the context of simpler true score and low-dimensional IRT models.

2.3 The Data

In the process of selecting a student representation for an assessment system, one must also consider what data one needs to see in order to make reliable inferences in the framework of the student representation, and how one arranges to get this data.

We begin with an illustration of a point that we have touched on before, relating sample size to model complexity: the more heavily parametrized the model, the heavier the data requirements. We contrast two cases: standard unidimensional IRT models as illustrated by Figure 3 and discrete attributes models as illustrated by Figure 1. Note that the sample size that matters here is the

number of responses per student, which affects precision of inferences about student attributes and proficiencies, not the number of students, which affects precision of estimates of conditional probabilities for the edges connecting attributes and proficiencies to data; in both the IRT and discrete attributes cases it is assumed that all of these "edge parameters" are known with certainty[11].

On the one hand, a well-designed test of 25 binary (right/wrong) items, based on an IRT model like that depicted in Figure 3, may be expected to estimate a unidimensional latent proficiency variable to within ± 0.22 population standard deviations[12] near the middle of the proficiency distribution, and to within about ± 0.31 standard deviations near the upper tail (Hambleton, 1989, Figure 4.18, p. 192). At $+2.00$ standard deviations this represents a variation of about ± 1.25 examinee population percentage points above and below the true proficiency value; less well-designed tests might accomplish this with roughly twice as many items (e.g. Lord, 1980, pp. 86–88). Hence the estimated posterior distribution for a proficiency parameter will be both much tighter than the population (prior) proficiency distribution, and its mean can be quite far from the population (prior) mean. This is due to the concentration of all of the student's response data on a single proficiency parameter. On the other hand, a very plausible Bayesian network similar to that depicted in Figure 1, developed by VanLehn, Niu, Siler and Gertner (1998), relating 34 binary (right/wrong) physics items to 66 binary (present/absent) attribute (skill) variables, was able to produce only the weakest inferences for 46 of the 66 attributes; in particular there was very little difference between prior and posterior distributions for these 46 attributes.

This example does not mean that the Bayesian network approach is hopeless, but it does draw attention to the severe lack of information available to estimate parameters in these models from a sequence of more or less naturally-occuring items scored wrong/right. There are essentially three approaches to improving the information that can be obtained about individual skills in the discrete skills models.

First, we can try to increase the richness of the item scores we assign. Even when the scoring is binary, care must be taken that the "right" distinction between right and wrong answers is made; for example Sijtsma (1997) shows that when items on a developmental measure of transitive reasoning are scored as simply as wrong/right, they have very low reliability[13] within a unidimensional IRT modeling framework. When the same items are rescored such that the item is wrong unless both the correct answer and a correct justification of the answer is provided, the items form a very strong IRT scale that reflects a natural increase in complexity of the transitive reasoning tasks.

A more transparent way to enrich item scores is to score items into more than two categories. This nearly always improves proficiency estimates in IRT models (Hambleton, 1989, p. 158); an early demonstration of this phenomenon is given by Thissen (1976). Today a variety of item re-

[11] In practice, they never are. See Section 3.3 and also Appendix B for some notions of how to incorporate uncertainty about edge parameters into measures of uncertainty about student variables (attributes and proficiencies).

[12] These are crude 95% intervals for proficiency, based on the information function plot on p. 192 of Hambleton (1989).

[13] Suggesting a problem with multidimensionality or at least heterogeneity of response process across students, not unlike that alluded to by Mislevy, 1994, p. 468

sponse models for partial credit and unordered categories of response exist (see Chapters 2–9 of van der Linden and Hambleton, 1997) and are in use, especially for more complex constructed-response items. Huguenard, Lerch, Junker, Patz, and Kass (1997; see also Patz, Junker, Lerch, and Huguenard, 1996) apply an unordered, polytomous response[14] IRT model to an experiment relating human working memory (WM) and other factors to performance navigating a hierarchical audio menu system. They use the unidimensional IRT proficiency variable to "factor out" general aptitude differences among experimental subjects, in order to focus inferences on how WM loads, and other experimental conditions, affect the various ways in which subjects could fail to navigate a hierarchical audio menu system, from only 27 items administered to less than 100 subjects. Similar gains might be expected in well-designed discrete attributes models that take advantage of more than wrong/right scores.

A second approach to improving the information available to estimate parameters in an assessment model is to make use of auxiliary information, i.e. information outside the (partial) correctness or nature of error in a students' task responses. Mislevy and Sheehan (1989) discuss the prospects for incorporating auxiliary information into estimates of the conditional probabilities that related proficiencies to responses in IRT models; and a very extensive conditioning model had been built for the National Assessment of Educational Progress (NAEP) (e.g. Johnson, Mislevy and Thomas, 1994) exists because the auxiliary information in the conditioning model does improve ability distribution estimates. Somewhat more directly, there is ongoing research is focused on using response latency to improve estimates of proficiency in computerized adaptive tests, at least in domains for which speed of response is important, based on the model of Roskam (1997). Finally, it may be possible to query students directly, or examine their task performances, for information that makes inferences about student attributes or proficiencies less uncertain. One example of this occurs in ITS's (e.g. Anderson, Corbett, Koedinger and Pelletier, 1995; Hill and Johnson, 1995) that either directly ask students what strategy they are pursuing, or keep track of other environmental variables, to help disambiguate strategy choice within a particular task performance.

Another example is suggested by cognitive construct validity studies such as that of Baxter, Elder and Glaser (1996). These authors establish the validity of a number-correct score on the Electric Mysteries fifth-grade science performance assessment, by asking students auxiliary questions to elicit explanations of how an electric circuit works, and plans for distinguishing among six "black box" circuits by experimentation; and by observing student strategies and self-monitoring behavior as the assessment progressed. They show that low-performing and high-performing students possess qualitatively quite different attributes relating to conceptual understanding about circuits, organization and detail of an experimental plan, effectiveness of strategy, and quality of self-monitoring behavior; and they establish that middle-performing students each have individual mixtures of the low- and high-performance attributes. This is a verbal description of a model not unlike that of Figure 5. The various competance levels probably could not be estimated accurately on the basis of total score (number of circuits correctly identified), or even pattern of rights and

[14] That is, items scored in several categories, not just wrong/right.

wrongs, alone, but Baxter et al.'s work directly suggests which additional response variables to collect—explanation, plan, strategy and self-monitoring—to decrease the uncertainty with which we might assign students to each of the three performance or competance states.

A third approach to improving estimation of student attributes is by careful "experimental design" applied to the assessment data collection process. We consider two clear possibilities. First, careful arrangement of tasks in a discrete-attributes model like that of Figure 1 may make it possible to gather enough redundant information about all the attributes that we can make inferences about the presence or absence of particular attributes despite local hiding and credit/blame assignment problems; indeed, VanLehn et al. (1998) attribute strong and weak inferences in their experiment with a discrete-attributes problem to exactly such design issues. The algebra of task and attribute dependencies developed by Tatsuoka (1990, 1995), Haertel and Wiley (1993) and others may be helpful in this regard, at least in simple cases. In addition, in some domains, it may even be possible to generate tasks "on the fly" to reduce our uncertainty about whether a particular student possessed this or that attribute.

Alternatively, if the assessment is embedded within a curriculum it may be possible to arrange observations so that only a small portion of the assessment model need be considered at any one time, *and* so that the data collected is directly and independently relevant to each of the few student attributes operating in that portion of the assessment model. Both the detailed information-processing-based ITS's of Anderson et al. (e.g. Corbett, Anderson and O'Brien, 1995) and the elaborate model for facets of students' understanding about physics of Minstrell (1998) involve hundreds of student attributes. Andersonian tutors manage the complexity by constraining students to follow a fairly narrow solution path on every task; and collecting responses for observable subtasks of each task that are in one-to-one correspondence with the underlying student attributes. Similarly Minstrell's DIAGNOSER (Minstrell, 1998) asks students apparently thin questions that are linked in a clever way to underlying student attributes ("facets" of understanding): DIAGNOSER asks only multiple choice questions, but each option[15] is tied to a particular student attribute: now not just subtasks but each possible response to a subtask (multiple-choice question) is tied to the presence/absence of a student attribute. Both approaches provide very direct evidence for each student attribute, that sidestep hiding and credit/blame attribution problems, assuming the model is correct[16].

[15] Minstrell also includes a "write-in" option that is not tied to any attributes, so that he can collect data relevant to expanding his model of student performance.

[16] The Minstrell model is essentially a deterministic version of Figure 5 within each problem; the knowledge tracing model described by Corbett et al. (1995) is essentially a version of Figure 1 in which guessing and slip parameters moderate otherwise deterministic links beteween attributes and responses with guessing and slip probabilities; see Section 5.1 below. They also simplify the model in an apparently harmless way: it is assumed that student attributes are present or absent *a priori* independently. This means that even if the presence of one attribute makes another more likely, the models do not take this into account; this results in somewhat lower certainty about the presence of attributes than would be present in the "complete" model, but it also results in greatly simplified computation. In addition the Corbett et al. (1995) models acquiring a skill over time using a latent Markov learning model; see Section 5.1.

2.4 Making and Reporting Inferences About Students

It should be clear by now that the point of view of this report leans heavily on traditional probability modeling. There are essentially two reasons for this: first, most of the models that are relevant for assessment are already expressed in the unifying language of traditional probability modeling; and second, rules for updating our inferences and assessing uncertainty about inferences—especially via Bayes' rule—enjoy relatively widespread concensus about their appropriateness, and are mechanically well-understood. Jameson (1995) surveys Bayesian networks and two other methods of updating student variables given data: Dempster-Schafer Theory (DST), and Fuzzy Logic. These latter two, especially DST, are intriguing potential alternatives to traditional probability modeling, but they provide neither the unifying framework within which to discuss the assessment models considered in this report, nor the consensus on methodology that would be important in any high-stakes assessment.

A more important issue to raise is the fundamental granularity problem implicit in choosing between an IRT style and a discrete-attributes style student representation. We have seen above in the sample size illustration in Section 2.3 that increasing the complexity of the model in the interest of cognitive validity can seriously reduce the power of the data to influence inferences about student variables. This might suggest that a better approach to assessment would be to use a coarser student model—perhaps an IRT model—for formal inference about student proficiency, and then informally map this proficiency onto a cognitive theory of performance in the domain. This can be done—Baxter, Elder and Glaser (1996) did something like this, and we will see more examples below—but it also has a drawback, namely that the informal nature of the association of student attributes with levels of proficiency makes it difficult to say with certainty that a student at a particular proficiency level does or does not possess particular student attributes.

Suppose we have a reasonably strong theory of performance in the domain of fractions in elementary school mathematics, that can be encoded by a discrete attributes Bayesian network model like that in Figure 1. Three of the attributes are, in increasing degree of sophistication[17], the skills "Informal use of the term 'half' ", "Compare fractions", and "Percentage". Let us identify these with Attributes 1, 4 and 5 in Figure 1; Attributes 2 and 3 will simply represent other "fractions" skills that do not concern our example. Since the model has an explicit parameter for whether each student possesses the "Compare fractions" skill (Attribute 4), we can report from the model a specific posterior probability, for each student, that he or she possesses this skill; or, if the posterior probability rises above some agreed-upon threshhold, simply report that the student does possess the skill. Additional evidence can be obtained about Attribute 4 by asking more questions that depend on it, but evidence from questions that do not depend on this attribute does not affect our inferences about Attribute 4 (Attribute 4 is independent of all other attributes and responses in this Bayesian network, except for responses that are its descendants; see Pearl, 1988, p. 120).

[17] The order of these by sophistication probably depends on the mathematics curriculum to which students are exposed; I can imagine "Percentage" being a less-sophisticated skill, with the aid of a handheld calculator, than "Compare fractions", for example.

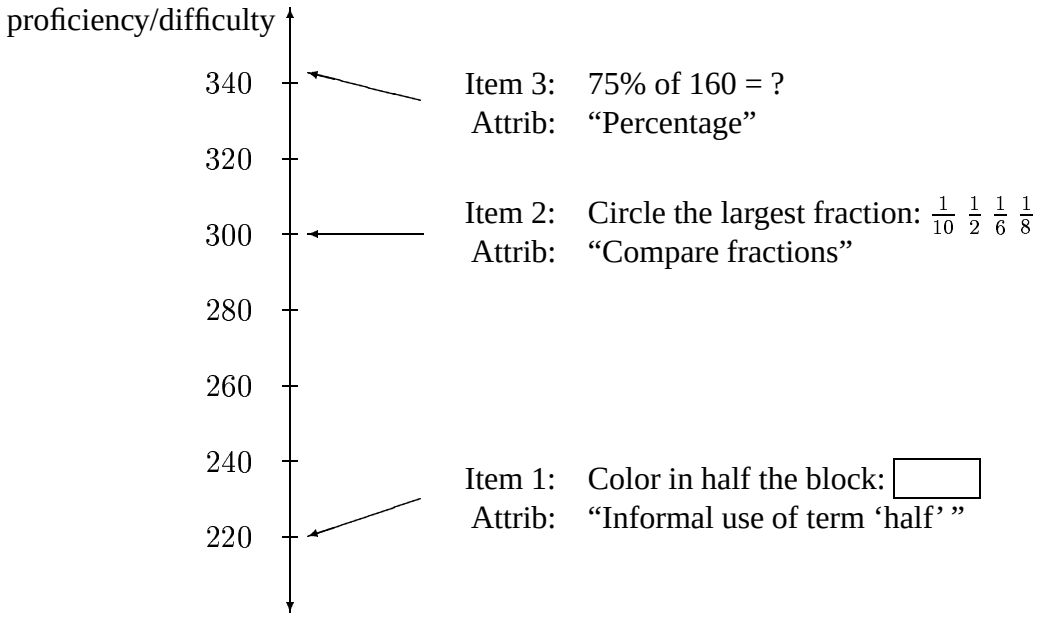


Figure 6: Scaling items and student attributes on a unidimensional IRT scale, as in Masters and Forster (1999) and Draney, Pirolli and Wilson (1995); this diagram is a greatly simplified version of Figure 2 of Masters and Evans (1986, p. 262). The vertical axis represents both student proficiency and item difficulty (more proficient students and more difficult items are located higher on the scale). When the item set is carefully constructed this can reveal an increasing progression of sophistication of domain understanding or increasing accumulation of attributes required to solve the items, at least for populations of students sharing a common curriculum. The student proficiency distribution can be plotted on the same scale to indicate where a population of students lies with respect to the items; or a single student's proficiency can be plotted, along with an indication of which items the student got wrong or right, as a kind of diagnostic screen.

Compare this with fitting a unidimensional IRT model to the same data, and displaying the items and the primary student attributes upon which they depend, on a linear scale as in Figure 6; such a display has been advocated by Masters and his colleagues for some time (e.g. Masters and Evans, 1986; Draney, Pirolli and Wilson, 1995; and Masters and Forster, 1999). When the item set is carefully constructed, this can reveal an increasing progression of sophistication of domain understanding or increasing accumulation of attributes required to solve the items, at least for populations of students sharing a common curriculum. Thus, in general, students whose proficiency is near 300 for example, have about a 12% chance of getting item 3 correct, a 50% chance of getting item 2 correct, and a 98% chance of getting item 1 correct. This is a statement about populations of students though, not any particular student. We could be as certain as we like that the student's proficiency score was 300[18]—and we will not know whether that student actually possesses the "Compare fractions" skill; we have to go outside the IRT model to assess this, for any particular student.

This is not to say that the graphical display of Figure 6 is useless; indeed it may work quite well on a number of levels, as both a communication device and as a kind of diagnostic screen. In many assessment settings it may only be feasable to report number-right or a similar gross measure of proficiency to a particular audience; a display like Figure 6 allows us to communicate what a student with each possible number-right score can be expected to do, much as NAEP achievment levels try to accomplish this interepretation of NAEP scores. As a guide to instruction, Figure 6 is useful also: a teacher might ask to plot the proficiency distribition of his or her students (e.g. as a rotated histogram) on the same scale to get an idea of what skills have been learned and which ones bear further teaching. A single student's proficiency score can be plotted, along with an indication of which items the student got wrong or right, as a kind of diagnostic tool: If the student's score is 340 and he/she got item 2 wrong, the teacher might probe the student directly about his/her facility with comparing fractions (i.e. go outside the IRT model as suggested above, which is not difficult in the classroom!). Or, several such plots might be compared side-by-side, to map the progress of students through a curriculum via several equated tests given at various intervals throughout the curriculum. The central point, however, is that despite its useful communication value, plots such as Figure 6 are not directly useful for assessing the presence or absence of discrete attributes in specific students.

Another example of the same kind of communication and diagnostic screening display is the

[18] Mischieviously, we could accomplish this by adding many more items of lesser sophistication and difficulty than Item 2—say, items near Item 1 for example. The unidimensional IRT proficiency/difficulty scale links performance on these items unrelated to the attribute "Compare fractions" to a proficiency estimtate "near" that attribute. Such mischief is impossible in the discrete-attributes Bayes network model of Figure 1, since as noted above "Compare fractions" (Attribute 4 in the figure) is stochastically independent of everything in Figure 1 except for responses that are its descendants in the graph. The fundamental problem that the mapping mapping of discrete student attributes onto the IRT proficiency/difficulty scale is an example of the well-known problem in IRT of mapping a multidimensional latent space onto a unidimensional one: many combinations of values in the multidimensional space correspond to the same unidimensional proficiency/difficulty value; no unique inference from the unidimensional space to the multidimensional space is possible.

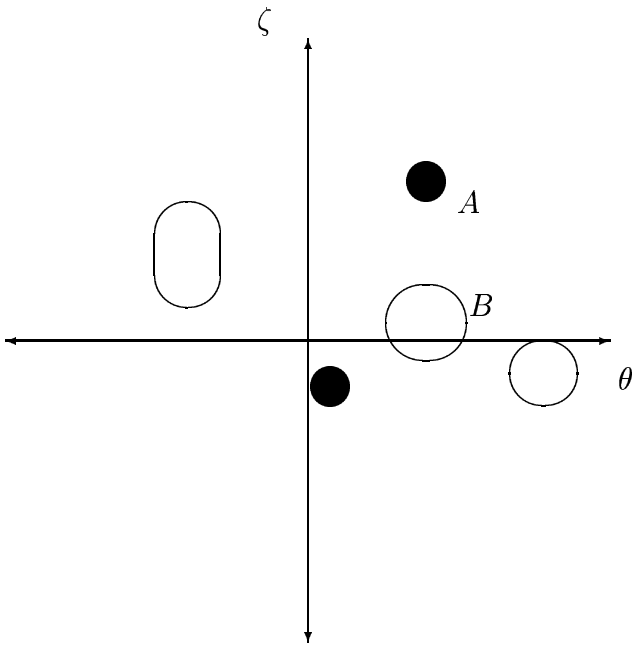


Figure 7: Clustering of examinees in Tatsuoka's (1990, 1995) rule space. θ is a unidimensional IRT-based student proficiency variable. ζ measures misfit of each student's pattern of rights and wrongs to the IRT model; ζ is sometimes called a "caution index". On the basis of a careful *a priori* analysis of student attributes needed to produce various answers on test items we can for example distinguish group A from group B (which are indistinguishable on the basis of θ alone) on the basis of differing correct and buggy rules possessed by students in each of these groups. See Tatsuoka (1995, Figure 14.5, p. 350) for a detailed description of various examinees' knowledge states based on this type of analysis applied to SAT-M items.

rule-space representation of Tatsuoka (1990, 1995). The first ingredient of this representation is a careful and complete analysis of correct and buggy rules, or attributes, underlying student performance on a set of items that are intended to be fitted with a unidimensional IRT model, including an analysis of dependencies between the rules. The second ingredient is a measure of the fit of each student's answer pattern of wrongs and rights to the IRT model, sometimes called a *caution index* in this literature; the closer to zero the caution index, the better the fit of the student's response pattern to the IRT model. Each student is thus assigned an estimated proficiency value θ and a caution index value ζ. All students' (θ, ζ) pairs are plotted, as in the schematic Figure 7; the $\theta \times \zeta$ plot is called a rule-space plot; clearly it is a kind of residual plot for the IRT model.

Usually the students cluster in some way in rule space; in the figure we can see that there are five clusters. Roughly speaking, we now examine the response patterns for the students in each cluster, to see what combination of learned and unlearned correct and buggy rules account for most of the answer patterns in each cluster. In Figure 7 all the clusters are already distinguished by their θ (proficiency) values, except for the clusters labelled A and B; the IRT model assigns a subset of the cluster B proficiencies to cluster A; the two clusters are initially only distinguished by their caution indices. It is also likely that a different combination of correct and buggy rules explain the response patterns of students in the two clusters; thus the rule-space plot shows how answer pattern residuals from a conventional unidimensional IRT model can exhibit considerable structure with respect to putative skills underlying the exam problems themselves.

The rule space plot can be converted into a linear display analogous to to figure Figure 6, except that only groups of student attributes and not tasks/items, are represented. Figure 14.5 of Tatsuoka (1995), for example, shows frequent knowledge states (groups of learned and unlearned attributes) corresponding to proficiency scores on a form of the Scholastic Aptitude Test (SAT) Mathematics Exam. As with Figure 6, both the rule space plot and the reduced plot showing only knowledge states corresponding to various proficiency levels are mappings of a high-dimensional space—the space of learned and unlearned discrete attributes—onto a one- or two-dimensional space. Many combinations of values in the high-dimensional space correspond to the same one- or two-dimensional space locations; hence no unique inference from the unidimensional space to the multidimensional space is possible. For this reason, neither Master's method of scaling attributes onto a unidimensional IRT scale, nor Tatsuoka's rule space method, allow direct model-based assessment of presence/absence of attributes within students. Such direct assessments again require us to go outside the unidimensional IRT framework.

The fundamental problem we have illustrated in this subsection is that there is a tradeoff to be made between fitting a model with too fine a granularity, degrading the quality of inferences; vs. starting with a model that is too coarse to allow formal inferences about student attributes of interest. The coarser models do allow a kind of informal, associational inferences to be made, and depending on the purposes of assessment this may be exactly what is required for reporting, or entirely inadequate.

3 Some Basic Statistical Assessment Models

In this section we provide a brief overview of the kinds of technical specifications that might underlie the models surveyed in Section 2. The models are treated in roughly the same order as Figures 1 through 4: In Section 3.1 we treat some simple deterministic and stochastic Bayesian networks for discrete variables; then in Section 3.2 we develop some basic item response theory (IRT) models. Finally in Section 3.3 we indicate how these models are really "the same", as alternate specifications of a two-way hierarchical modeling framework. Unless otherwise indicated, both task response variables and student attribute variables are binary (wrong/right task response; absent/present attribute). This simplifies the exposition; it doesn't represent a fundamental limitation of the models and methods described here.

It is convenient to think of the raw response data as an $N \times J$ matrix

$$\mathcal{X} = [X_{ij}] = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1J} \\ X_{21} & X_{22} & \cdots & X_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ X_{N1} & X_{N2} & \cdots & X_{NJ} \end{bmatrix}$$

In practice some of the responses X_{ij} may be missing, in all the usual ways that item response data can be missing in assessment problems; for ease of exposition we treat such missingness as irrelevant to our purpose; factors in models corresponding to missing data would simply be omitted from the relevant equations.

3.1 Two Simple Discrete-Attributes Models

A reasonable place to start the assessment modeling problem is with a list of J items, and a list of K student attributes, to which we attribute most of the variability in students' task performance. Depending on the nature of the assessment, the items may be whole tasks, or they may be subgoals nested within a collection of $M \geq J$ tasks. For simplicity of discussion however I will refer to them as tasks (so $M = J$).

3.1.1 A Deterministic Assessment Model

Many models that relate task performance to presence or absence of student attributes feature something like the *Q-matrix* (e.g. Tatsuoka, 1990, 1995), which we will write as a $J \times K$ matrix $Q = [q_{jk}]$ of 0's and 1's with entries

$$q_{jk} = \begin{cases} 1, & \text{if attribute } k \text{ is required by task } j \\ 0, & \text{if not} \end{cases} \tag{2}$$

(this is actually the transpose of the matrix Q that Tatsuoka uses, but it is more convenient for our exposition to write it like this). Recall from equation (1) in Section 2.2.2 that the Q matrix encodes the dependencies in a bipartite directed acyclic graph such as Figure 1.

While the Q-matrix does not capture all of the structure we may be interested in—Q treats the student attributes in a flat, non-time-ordered manner, and there may be both hierarchical and time-order structure in the attributes as they are applied to, or become relevant for the performance of, a task—it is a useful bookkeeping device. The basic Q-matrix idea—the incidence matrix of a directed acyclic graph—can be extended to accomodate layered models as in Figure 2, multi-strategy models as in Figure 5 and other complexities.

It is worth noting that the expression of the Q matrix is in terms of attributes relevant to the tasks, but for the purposes of assessment most of the attributes would be thought of as residing in the student, e.g. skills, beliefs and bits of knowledge acquirable by the student, and the purpose of the assessment is in fact to figure out which of these the student has acquired. One way to express this is to define another, possibly time-dependent matrix $A(t) = [\alpha_{ik}(t)]$ of dimension $N \times K$ (where N is the number of students), with elements

$$\alpha_{ik}(t) = \begin{cases} 1, & \text{if attribute } k \text{ is present for student } i \text{ at time } t \\ 0, & \text{if not} \end{cases} \tag{3}$$

Note that $q_j \stackrel{def}{=} [q_{j1}, \ldots, q_{jK}]$ lists all the attributes that are required for task j and $\alpha_i(t) \stackrel{def}{=} [\alpha_{i1}(t), \ldots, \alpha_{iK}(t)]$ lists all the attributes that student i posesses. If we now denote success or failure of task performance at time t as

$$X_{ij}(t) = \begin{cases} 1, & \text{if student } i \text{ performs task } j \text{ successfully at time } t \\ 0, & \text{if not} \end{cases}$$

a simple determinsitic model of task performance would posit

$$X_{ij}(t) = \begin{cases} 1, & \text{if } q_{jk} \leq \alpha_{ik}(t) \; \forall \; k \\ 0, & \text{if not} \end{cases} \tag{4}$$

that is, $X_{ij}(t) = 1$ when all the attributes needed for task j are present for student i at time t.

Most of the assessment models that we will discuss are one-time assessments and hence the time dependence will often be suppressed in what follows; it will really only arise in discussing the knowledge-tracing model of Corbett, Anderson and O'Brien (1995), in Section 5.1.

3.1.2 A Simple Stochastic Assessment Model

The model we present here is essentially identical to the conjunctive Bayesian network ("noisy AND-gate" model) investigated by VanLehn, Niu, Siler, and Gertner (1998), and discussed above in Section 2.3; it can also be considered as a one-time version of the Corbett, Anderson and O'Brien

(1995) model. This model can also be connected to multidimensional noncompensatory IRT models, since it is interpetable as a simplified version of Embretson's (1985) multicomponent latent trait (MLTM) model, which we alluded to in conjunction with Figure 4, p. 17, and which we discuss again below in Section 4.2. Define, exactly as in Section 3.1.1,

$$\begin{array}{rcll} X_{ij} & = & 1 \text{ or } 0 & \text{indicating whether or not student } i \text{ performed task } j \text{ correctly} \\ q_{jk} & = & 1 \text{ or } 0 & \text{indicating whether or not task } j \text{ requires attribute } k \\ \alpha_{ik} & = & 1 \text{ or } 0 & \text{indicating whether or not student } i \text{ posesses attribute } k \\ \xi_{ij} & = & \prod_{k:\ q_{jk}=1} \alpha_{ik} & \text{indicating whether or not student } i \text{ has the attributes needed for task } j \\ s_j & = & P[X_{ij} = 0 | \xi_{ij} = 1], & \text{a per-problem slip parameter} \\ g_j & = & P[X_{ij} = 1 | \xi_{ij} = 0], & \text{a per-problem guessing parameter} \end{array}$$

The basic response model is

$$\begin{aligned} P[X_{ij} = 1 | \underset{\sim}{\xi}, \underset{\sim}{\mathbf{s}}, \underset{\sim}{\mathbf{g}}] &= \xi_{ij}(1 - s_j) + (1 - \xi_{ij}) g_j \\ &= (1 - s_j)^{\xi_{ij}} g_j^{1-\xi_{ij}} \end{aligned} \tag{5}$$

and so for the entire examinees by tasks matrix $\mathcal{X} = [x_{ij}]$ of task responses, we have[19]

$$\begin{aligned} P[\mathcal{X} | \underset{\sim}{\xi}, \underset{\sim}{\mathbf{s}}, \underset{\sim}{\mathbf{g}}] &= \prod_i \prod_j \left[(1 - s_j)^{\xi_{ij}} g_j^{1-\xi_{ij}} \right]^{x_{ij}} \left[1 - (1 - s_j)^{\xi_{ij}} g_j^{1-\xi_{ij}} \right]^{1-x_{ij}} \\ &= \prod_i \prod_j \left[(1 - s_j)^{x_{ij}} s_j^{1-x_{ij}} \right]^{\xi_{ij}} \left[g_j^{x_{ij}} (1 - g_j)^{1-x_{ij}} \right]^{1-\xi_{ij}} \end{aligned} \tag{6}$$

Note that the ξ_{ij} in this model is defined in terms of the q_{ij}'s and α_{ik}'s exactly as X_{ij} was defined in the deterministic model (4) above, only now ξ_{ij} is a "latent response" (Maris, 1995): it is the function that combines attributes conjunctively for each observable response variable X_{ij}. The goal is to try to estimate the α_{ik}'s, or more precisely $P[\alpha_{ik} = 1]$, from the task performance data. A conceptually easy estimation method, and some notes on the sensitivity of estimated parameters to variation in the data, are recorded in Appendix B.6. Readers familiar with item response theory will also note a formal similarity betwen the likelihood (6) and the basic joint likelihood of an item response theory model, which we discuss next.

3.2 Item Response Theory

Item response theory (IRT; e.g. van der Linden and Hambleton, 1997) is not a "theory" in any scientific or mathematical sense, but rather a psychometric modeling tradition, much as discrete-node Bayes networks are a modeling tradition in AI; indeed, both IRT and discrete Bayes networks

[19] This formulation assumes that tasks are conditionally independent, as in Figure 1. More general dependence structures might replace Equation (6) a Bayesian network model, for example.

can be viewed as special cases of the hierarchical modeling framework described in Section 1. IRT typically employs a logistic response function[20]

$$P(X = 1|t) = \frac{1}{1 + \exp(-t)}$$

to relate task performance $X = 0$ or 1 as in Section 3 with a countinuous predictor t. The predictor t is usually decomposed into (at least) a fixed effect β_j for task j and a random effect θ_i for student i: $t = \theta_i - \beta_j$. Thus a student's "ability" θ_i competes with the "difficulty" of the task β_j to bias the probability that the task will be performed correctly. These are all fairly standard building blocks: a statistician looking at IRT will see mixed-effects logistic regression; a computer scientist might see a neural network with a sigmoidal response and a particularly simple topology.

The IRT model just described,

$$P[X_{ij} = 1|\theta_i, \beta_j] = \frac{1}{1 + \exp(-[\theta_i - \beta_j])} \tag{7}$$

is called the Rasch (1960) model or the one-parameter logistic (1PL) model. There is also a 2PL model

$$P[X_{ij} = 1|\theta_i, \alpha_j, \beta_j] = \frac{1}{1 + \exp(-\alpha_j[\theta_i - \beta_j])} \tag{8}$$

reflecting differential sensitivity of task performance to student "ability" through the so-called "discrimination" parameter α_j, and a 3PL model

$$P[X_{ij} = 1|\theta_i, g_j, \alpha_j, \beta_j] = g_j + (1 - g_j)\frac{1}{1 + \exp(-\alpha_j[\theta_i - \beta_j])} \tag{9}$$

incorporating a parameter for guessing behavior similar to that of our stochastic Q model, equation (5). These models also make a conditional independence assumption like equation (6):

$$P[\ \underset{\sim}{\mathbf{x}}\ |\ \underset{\sim}{\theta}, \ldots, \underset{\sim}{\beta}] = \prod_{i=1}^{N}\prod_{j=1}^{J} P[X_{ij} = x_{ij}|\theta_i, \ldots, \beta_j] \tag{10}$$

Many more details and variations can be found in the recent edited volumes by Fischer and Molenaar (1995) and van der Linden and Hambleton (1997).

Parametric IRT, as surveyed for example in the edited volumes of Fischer and Molenaar (1995) and van der Linden and Hambleton (1997), is a well-established, wildly successful statistical modeling enterprise. IRT models have greatly extended the data analytic reach of psychometricians, social scientists, and educational measurement specialists. Parametric IRT models, extended by hierarchical modeling and estimation strategies, make it possible in principle and in practice to incorporate auxiliary information in the form of covariates and other structure, to improve inferences about both items and students. Many violations of the basic local independence assumption of IRT models are in fact due to unmodeled heterogeneity of subjects and items, that can now be explicitly modeled using these methods.

[20] Though other sigmoidal functions are also used, the logistic is most common.

3.3 Two-Way Hierarchical Structure

As we saw in Section 2, IRT models and discrete attributes models are both instances of Bayesian networks. We shall see below that they are also both instance of the same hierarchical modeling framework; this common framework lets us think about modeling and estimation issues for both kinds of model; in fact it also lets us think about mixing and matching various model components to develop models that are well-tailored to a specific assessment situation (see e.g. Patz and Junker, 1999b, for an example of this).

Because of the two-way layout of the data in the raw response data matrix $\mathcal{X} = [X_{ij}]$—students i by assessment tasks j—the hierarchical framework for these models also has a two-way structure. This two-way structure was suppressed in the examples in Section 2.2.2 to focus attention on accumulating data for inferences about students based on assessment task performance, but it is already evident in the double-products in equations (6) and (10).

The general framework for statistical assessment models is a three-level, two-way hierarchical structure for N individuals and J response variables, which we depict initially in terms of notation that is similar to the IRT model notation; see Figure 8. The figure emphasizes that the model is built out of conditional probabilities (or conditional distributions), represented by edges in the graph, at three or more "levels", successively farther farther from the raw task performance data, represented by labelled nodes in the graph:

First Level: At level one, observable task response variables X_{ij} summarizing student i's response to task (or subtask) j follow distributions $p(X_{ij}|\underset{\sim}{\theta}_i, \underset{\sim}{\beta}_j)$ determined by student attribute parameters $\underset{\sim}{\theta}_i$ and task feature parameters β_j, for all students $i = 1, \ldots, N$ and all tasks $j = 1, \ldots, J$.

Second Level: At level two, student attribute parameters $\underset{\sim}{\theta}_i$ follow distributions $f_i(\underset{\sim}{\theta}_i|\lambda_f)$; task feature parameters $\underset{\sim}{\beta}_j$ follow distributions $g_j(\underset{\sim}{\beta}_j|\lambda_g)$.

These first two levels represent the formulation of a basic Bayesian model for the task performance data, with a likelihood

$$\prod_{i=1}^{N}\prod_{j=1}^{J} p(X_{ij}|\underset{\sim}{\theta}_i, \underset{\sim}{\beta}_j)$$

for the data from level one[21], and prior distributions

$$\prod_{i=1}^{N} f_i(\underset{\sim}{\theta}_i|\lambda_f) \text{ and } \prod_{j=1}^{J} g_g(\underset{\sim}{\beta}_j|\lambda_g)$$

from level two, to represent our uncertainty about the the likelihood parameters.

The "two way" structure of the model can be seen beginning at level two (and continues through level three and higher): there is a separate set of parameters $\underset{\sim}{\theta}_i$ for student attributes and $\underset{\sim}{\beta}_j$ for task features.

[21] Note the similarity to both of equations (6) and (10).

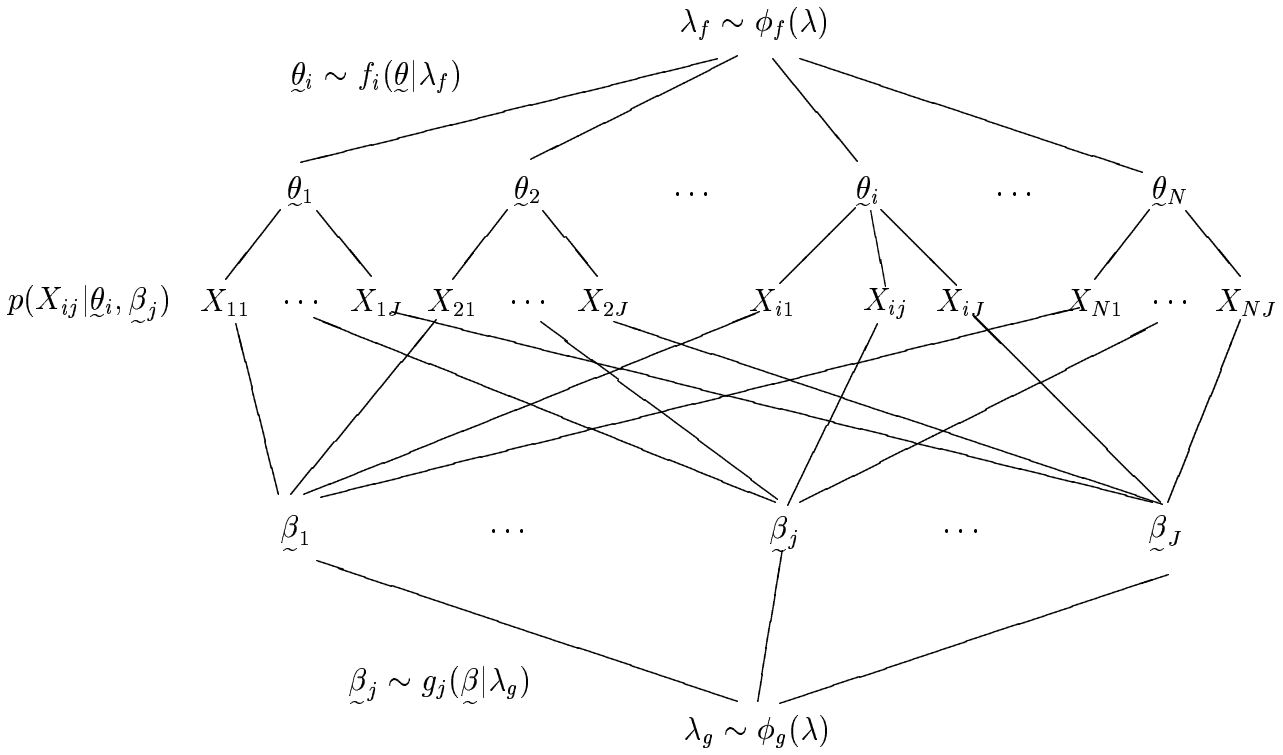


Figure 8: A three-level hierarchical Bayes view of many of the models in this report, described in equations (11). Both traditional IRT models and "discrete attributes" models of Figures 1 through 4 are represented; multiple strategy models such as Figure 5 can be accomodated as well. A Q matrix relating student attributes to tasks determines how coordinates of $\underset{\sim}{\theta}_i$ are related to tasks X_{ij}. A possibly different Q-matrix relating task features to tasks determines how coordinates of β_j are related to each task X_{ij}. See text in Section 3.3 for details. This figure is essentially identical to that of Mislevy (1994, p. 462).

Third Level (and higher): At level three (and possibly higher levels), alternate student strategies, differential performance among various groups of students, etc., are modeled with the parameters λ_f and λ_g which themselves follow distributions $\phi_f(\lambda)$, $\phi_g(\lambda)$.

Thus the hierarchical model is a kind of Bayesian network. The expression of the model in terms of conditional probability distributions is a way of expressing the fact that prediction of X from $\underset{\sim}{\theta}$ and β is uncertain, and also prediction of $\underset{\sim}{\theta}$ and β from the λ's is uncertain. Inference from observed data to likely values of $\underset{\sim}{\theta}$'s β's, etc. is via Bayes' rule. For ease of reference later, we record the hierarchical modeling structure compactly as follows:

$$\left.\begin{array}{lrcl} \text{First level:} & X_{ij} & \sim & p(X_{ij}|\underset{\sim}{\theta}_i, \underset{\sim}{\beta}_j), \\ & & & \qquad i = 1, \ldots, N;\ j = 1, \ldots, J \\ \text{Second level:} & \underset{\sim}{\theta}_i & \sim & f_i(\underset{\sim}{\theta}|\lambda_f), \quad \text{each } i \\ & \underset{\sim}{\beta}_j & \sim & g_j(\underset{\sim}{\beta}|\lambda_g), \quad \text{each } j \\ \text{Third level:} & \lambda_f & \sim & \phi_f(\lambda_f) \\ & \lambda_g & \sim & \phi_g(\lambda_g) \\ & \vdots & & \end{array}\right\} \tag{11}$$

Distributional assumptions at the first level, $p(X_{ij}|\underset{\sim}{\theta}_i, \underset{\sim}{\beta}_j)$, determine whether the model is an IRT model, a "discrete attributes" model, or something else. For example,

- In a traditional IRT model, the only student attribute $\underset{\sim}{\theta}_i$ is a continuous "general proficiency" variable, or possibly a small collection of continuous proficiency variables as in multidimensional IRT models, and task parameters $\underset{\sim}{\beta}_j$ typically code features such as task difficulty and discrimination (sensitivity to changes in $\underset{\sim}{\theta}$); see for example equation (8). As a description of IRT models, this diagram is equivalent to Figure 8 of Mislevy (1994, p. 462).

- In a "discrete attributes" model emphasizing the individual acquisition of many distinct attributes and pieces of knowledge, $\underset{\sim}{\theta}_i$ is replaced with the vector of binary (say) variables $(\alpha_{i1}, \ldots, \alpha_{iK})$ indicating the presence or absence of each of these distinct attributes in student i. $\underset{\sim}{\beta}_j$ might code task features such as guessing and slip parameters for the j^{th} item; see for example equation (5).

We see again, as pointed out in Section 2.2.2, that the principal differences between Bayes networks and general IRT models is in granularity of representation and particular distributional assumptions, nothing more.

The second and third (and higher) levels can be used to impose constraints on the first level parameters and latent variable, incoroporate other covariates into the model, etc. For example, in the "discrete attributes" model—and in the multidimensional IRT models discussed below—relationships between student attributes/proficiencies, which are coordinates of $\underset{\sim}{\theta}_j$, and tasks might be coded by a Q matrix.

Displaying assessment models as in Figure 8 makes the role, and demands, of the task parameters β_j as visible as the student attribute parameters $\underset{\sim}{\theta}_i$. We see immediately that values are required for the β_j in order for the model to work. Estimation methods that simulataneously obtain β_j and β_j are briefly described in Appendix B. It is immediately obvious that information obtained from the X_{ij} must always propagate up the branches of both trees; spreading data out like this leads to greater uncertainty in student atttribute estimation for example (since some information had to be "spent" on task parameter estimation). In many situations we "plug in" values for the task parameters β, based on guesswork or previous assessment data. This can artificially reduce uncertainty about proficiency estimates, since all the information from observing X_{ij} will then propogate up the attributes subtree.

4 Extensions of the Basic Models

In this section we survey a few ways in which the basic models described in Section 3 have been extended for use in cognitively richer assessment settings, both to give a feeling for the variety of model choices that exist and to suggest many areas of common ground between the models. A more complete survey of such models is provided by Roussos (1994).

4.1 The Linear Logistic Test Model

The Linear Logistic Test Model (LLTM; Scheiblechner, 1972; Fischer, 1973) places linear constraints on the task difficulty parameters β_j in the Rasch model:

$$P[X_{ij} = 1 \mid \theta_i, \beta_j] = \frac{1}{1 + \exp(-[\theta_i - \beta_j])}$$

where the vector $\underset{\sim}{\beta}$ is linearly constrained by

$$\underset{\sim}{\beta}_{J\times 1} = Q_{J\times K}\underset{\sim}{\psi}_{K\times 1} \tag{12}$$

and the entries q_{jk} of the "bookkeeping" matrix Q are

$$q_{jk} = \begin{cases} 1, & \text{if attribute } k \text{ is required by task } j \\ 0, & \text{if not} \end{cases} \tag{13}$$

The attribute can be a cognitive skill, a surface feature of the item, etc. (e.g. Fischer and Molenaar, 1995; Draney et al., 1995; Huguenard et al., 1997; Embretson, 1995a; 1995b; 1999). Thus task

difficulty parameters β_j have the linear structure

$$\begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{J-1} \\ \beta_J \end{bmatrix} = Q \begin{bmatrix} \psi_1 \\ \psi_2 \\ \vdots \\ \psi_K \end{bmatrix} \tag{14}$$

where Q is an appropriate design matrix of full column rank, to reflect dependencies between items due to common item features. Note that this use of the Q matrix is equivalent to specifying (linear) constraints on β_j through its prior distribution $g_j()$.

This model and its various generalizations (e.g. Glas and Verhelst, 1989; Patz and Junker, 1999b) continues to be used for psychological experiments with multiple outcomes per subject (e.g. Fischer and Molenaar, 1995 and the references therein) and for research in cognitively-motivated test design (Embretson, 1995a,b; 1999). Recently, Embretson (1999) has also been applying an LLTM-like decomposition to item discriminations in a 2PL model, to explore variation in item discrimination with increasing cognitive load, again with an eye toward test design. A conceptually straightforward method of estimation for the LLTM model (as well as for the basic 2PL IRT model) is considered briefly in Appendix B.4.

By itself, the LLTM cannot directly provide a vehicle for cognitive diagnosis or assessment, since the student parameter θ_i is unidimensional; and cognitive descriptions of performance are invariably multidimensional. Keeping this in mind, we now turn to some basic multidimensional IRT models.

4.2 Modeling Multidimensional Proficiencies

Research in multidimensional IRT models has concentrated on two extremes: "compensatory" models in which different proficiencies combine linearly, so that a lack in one proficiency may be made up with an excess in another proficiency, and "noncompensatory" models in which each observable response represents the conjunction of several subprocesses that individually follow unidimensional IRT models.

4.2.1 Multidimensional, Compensatory IRT Models

Compensatory models are obtained by replacing the proficiency variable θ in a unidimensional IRT model with an item-specific, known (e.g. Embretson, 1991; Stegelman, 1983; and Adams, Wilson and Wang, 1997) or unknown (e.g. Reckase, 1985; Wilson, Wood and Gibbons, 1983; Fraser and MacDonald, 1988; Muraki and Carlson, 1995) linear combination of components $a_{j1}\theta_1 + \cdots + a_{jd}\theta_d$ of a d-dimensional latent trait vector. For example in the dichotomous response case

$$P[X_j = 1|\theta_1, \ldots, \theta_d] = P(a_{j1}\theta_1 + \cdots + a_{jd}\theta_d - \beta_j),$$

where $P(t)$ might be the logistic function from Section 3.2, or a probit response function, etc. Béguin and Glas (1998) survey the area well and give an MCMC algorithm for estimating these models; Gibbons and Hedeker (1997) pursue related developments in biostatistical and psychiatric applications; see also Ackerman (1992, 1994). When the a_{jk} are known, this is a kind of additive or disjunctive Q-matrix decomposition of examinee proficiencies.

Estimating this model without firm constraints on the parameter vectors $\underset{\sim}{\alpha}_j = (a_{j1}, \ldots, a_{jd})$ is tantamount to performing an exploratory factor analysis on the logits of the success probabilities of the 0/1 data X_{ij}; exploratory factor analysis is already a difficult and unstable problem which is only exacerbated by the indirect connection between the data and the parameters through the logistic function $P(t)$. When the $\underset{\sim}{\alpha}_j$ are suitably constrained the model seems to be more stable to estimate, and several articles on estimating the more constrained version of this model using MCMC methods similar to those outlined in Appendix B.4 are making their way through the journal review process presently.

Another generalization that has some utility would be to add to the model in which the a_{jk} were known, a (perhaps different) Q-matrix decomposiion of the difficulty parameters β_j along the lines the LLTM, thus:

$$t = \theta_i - \beta_j = \underset{\sim}{\mathbf{a}}_j \cdot \underset{\sim}{\theta}_{\mathbf{i}} - q_j \cdot \underset{\sim}{\psi}$$

where now $\underset{\sim}{\theta}_i$ is a d-dimemsional vector of (say) attribute-specific continuous abilities, and $\underset{\sim}{\mathbf{a}}_j$ is a d-dimensional vector of known constants (usually 0's and 1's, analogously to the Q matrix) indicating which coordinates of $\underset{\sim}{\theta}_i$ are involved in performing task j. Adams, Wilson and Wang (1997) refer to this general model (or actually a polytomous generalization of it) as the multidimensional random coefficients multinomial logit model (MRCMLM), provide an E-M algorithm for estimating the model, illustrate some useful special cases of the model, and demonstrate several data analyses with it. A special case of this model, that adds dimensions to $\underset{\sim}{\theta}_{\mathbf{i}}$ to account for pre-test/post-test performance differences, has been proposed and used by Embretson (1991). A further elaboration of the MRCMLM model, called M^{2}RCMLM, allows for mixing over both LLTM-like item difficulty decompositions and over various compensatory compbinations of student proficiency variables; it is being developed by Pirolli and Wilson (1999) as a general stochastic model for cognitive assessment.

All three types of multidimensional IRT models represent real computational advances, and have applications in more traditional educational assessment and educational survey settings as outlined by Ackerman (1992, 1994) and Adams et al. (1997). As models for cognitively-relevant assessment, they assert a rather different relationship than the conjunctive discrete attributes model of Section 3.1.2 or the noncompensatory IRT models to be discussed next. The "cognitive attributes" parameters $\underset{\sim}{\theta}_i$ enter the model compensatorily: that is, a low value in one coordinate of $\underset{\sim}{\theta}_i$ can be compensated for by a high value in another, perhaps yielding a fairly high probability of successful task performance even though one attribute component was quite low. Whether this compensatory relationship is "right" for a given assessment depends is of course an empirical model fit issue.

The compensatory relationship just described has some of the flavor of "multiple strategies" where, if one strategy is absent (low value for that coordinate of $\underset{\sim}{\theta}_i$) another more fully-learned strategy (higher value for that coordinate of $\underset{\sim}{\theta}_i$) can compensate for it. But there is no attributes model underlying these "strategies"—just ordered values for coordinates of $\underset{\sim}{\theta}_i$. A more general approach to multiple strategies will be considered below.

4.2.2 Multidimensional, Non-Compensatory IRT Models

An alternative to the compensatory models of the previous section is suggested by inspecting the definition of ξ_{ij} in Section 3.1.2 above. Given an analysis of tasks into attributes which must be present *in conjunction* in order for the task to be correctly performed, we might posit variables

$$Y_{ijk} = \begin{cases} 1, & \text{if student } i \text{ executes a correct behavior when attribute } k \text{ is called for on item } j; \\ 0, & \text{if not} \end{cases}$$

which themselves follow a logistic IRT model with respect to a attribute specific trait θ_{ik},

$$P[Y_{ijk} = 1 \mid \theta_{ik}, \beta_{jk}] = \frac{1}{1 + \exp[-(\theta_{ik} - \beta_{jk})]},$$

say (of course other models besides the Rasch model are possible here). If the Y's are conditionally independent given the parameters, as is our usual assumption, then

$$P[X_{ij} = 1 \mid \underset{\sim}{\theta}_i, \underset{\sim}{\beta}_{\mathbf{j}}] = \prod_{k:\ q_{jk}=1} P[Y_{ijk} = 1 \mid \theta_{ik}, \beta_{jk}]$$

Embretson (1985) has proposed a multicomponent latent trait model (MLTM) that adds to this basic noncompensatory model slip and guessing parameters similar to those discussed in Section 3.1.2. In the MLTM,

$$P[X_{ij} = 1 \mid \underset{\sim}{\theta}_i, \underset{\sim}{\beta}_{\mathbf{j}}] = (1-s) \prod_{k:\ q_{jk}=1} P[Y_{ijk} = 1 \mid \theta_{ik}, \beta_{jk}] + g \left(1 - \prod_{k:\ q_{jk}=1} P[Y_{ijk} = 1 \mid \theta_{ik}, \beta_{jk}]\right)$$

where s and g are (universal) slip and guessing parameters satisfying

$$\begin{aligned} (1-s) &= P[X_{ij} = 1 | \prod_{k:\ q_{jk}=1} Y_{ijk} = 1] \\ g &= P[X_{ij} = 1 | \prod_{k:\ q_{jk}=1} Y_{ijk} = 0] \end{aligned}$$

If we also linearly decompose the difficulty parameters β_{ik} in the MLTM as in the LLTM, we obtain Embretson's general component latent trait model (GLTM). The MLTM has been applied

for example to synonym tasks by Janssen and De Boeck (1997), who also develop some heuristics for determining whether the MLTM is an appropriate model for the data at hand.

A fundamental issue with the MLTM is whether the Y_{ijk} need to be directly observable or not. In Embretson's early work on the model she assumed that the Y_{ijk} would be observable. As in the discussion of ways to improve the information for estimating student attributes in Section 2.3, this would be expected to lead to better parameter estimates on relatively little data. Later the requirement that Y_{ijk} be observable was relaxed. While it is computationally feasable for the Y_{ijk} to remain unobserved (a straightforward algorithm to estimate the model in this case would look somewhat like the algorithm for the discrete-attributes model described in Appendix B.6), this situation immediately leads us to the credit/blame and hiding issues outlined in our discussion of Figure 1 in Section 2.2.2.

4.3 Other Discrete-Attribute Approaches

4.3.1 Latent Class Models

Suppose students can be classified into W latent (unobservable) classes C_w, $w = 1, \ldots, W$, and let

$$\begin{aligned} \lambda_w &= P[\text{student } i \text{ is in class } C_w] \\ p_{wj} &= P[X_{ij} = 1 | \text{student } i \text{ is in class } C_w] \end{aligned}$$

As usual $X_{ij} = 1$ or 0 indicating correct performance of task j by student i. Since we do not know student i's latent class, the model for one student is

$$P[\underset{\sim}{\mathbf{X}}_i = \underset{\sim}{\mathbf{x}}_i \mid \underset{\sim}{\mathbf{p}}, \underset{\sim}{\lambda}] = \sum_{w=1}^{W} \lambda_w \prod_{j=1}^{J} p_{wj}^{x_{ij}} [1 - p_{wj}]^{1-x_{ij}}$$

and for an $N \times J$ matrix of response data $\mathcal{X}$, we obtain

$$P[\mathcal{X} \mid \underset{\sim}{\mathbf{p}}, \underset{\sim}{\lambda}] = \prod_{i=1}^{N} \left\{ \sum_{w=1}^{W} \lambda_w \prod_{j=1}^{J} p_{wj}^{x_{ij}} [1 - p_{wj}]^{1-x_{ij}} \right\} \tag{15}$$

This latent class model is somewhat different in form from the discrete attributes model (6) and the basic IRT model (10): the likelihood is not in double product form; there is a summation in the way. Most estimation methods become unwieldy with this model, as N, J and W grow large.

There is, however, a way to "decouple" this model and put it back in product form. The method goes by the name *data augmentation* in the statistical world (Tanner, 1996); it is also related to the latent response approach of Maris (1995). Let

$$z_{iw} = \begin{cases} 1, & \text{if student } i \text{ is in latent class } w \\ 0, & \text{if not} \end{cases}$$

Then to each students actual observed data $(x_{i1}, \ldots, x_{iJ})$ we will add the "data augmentation" variables $\underset{\sim}{\mathbf{z}}_i = (z_{i1}, \ldots, z_{iW}) = (0, 0, \ldots, 1, 0, 0, \ldots 0)$ with a 1 only in position w, student i's assigned latent class.

If we pretend that both x's and z's are observed, the model for the i^{th} student simplifies; it is now just

$$\begin{aligned} P[\underset{\sim}{\mathbf{x}}_i, \underset{\sim}{\mathbf{z}}_i | \underset{\sim}{\mathbf{p}}, \underset{\sim}{\lambda}] &= P[\underset{\sim}{\mathbf{z}}_i | \underset{\sim}{\lambda}] \cdot P[\underset{\sim}{\mathbf{x}}_i | \underset{\sim}{\mathbf{z}}_i, \underset{\sim}{\mathbf{p}}, \underset{\sim}{\lambda}] \\ &= \left\{ \prod_w \lambda_w^{z_{iw}} \right\} \cdot \prod_w \left\{ \prod_j p_{wj}^{x_{ij}} [1 - p_{wj}]^{1 - x_{ij}} \right\}^{z_{iw}} \end{aligned} \tag{16}$$

The joint probability model for $\mathcal{X}_{N \times J}$, $\mathcal{Z}_{N \times W}$, and the parameters, also simplifies; at least it is in product form again:

$$\begin{aligned} & P[\mathcal{X}, \mathcal{Z} | \underset{\sim}{\mathbf{p}}, \underset{\sim}{\lambda}] \pi(\underset{\sim}{\mathbf{p}}) \pi(\underset{\sim}{\lambda}) \\ &= \prod_i \prod_w \left\{ \lambda_w \cdot \prod_j p_{wj}^{x_{ij}} [1 - p_{wj}]^{1 - x_{ij}} \right\}^{z_{iw}} \pi(\underset{\sim}{\lambda}) \prod_j \pi_p(p_{wj}) \\ &= \left\{ \prod_w \lambda_w^{n_w} \right\} \pi(\underset{\sim}{\lambda}) \left\{ \prod_j p_{wj}^{c_{wj}} [1 - p_{wj}]^{n_w - c_{wj}} \pi_p(p_{wj}) \right\} \end{aligned} \tag{17}$$

where $n_w = \sum_i z_{iw}$, and $c_{wj} = \sum_i z_{iw} x_{ij}$ (and $\pi(\underset{\sim}{\lambda})$ and $\pi_p(p_{wj})$ are priors). The price we pay for this product structure is now every estimation algorithm will have to iterate between imputing values for $\mathcal{Z}$ and working with this new simpler joint likelihood. Alternation between imputing values for $\mathcal{Z}$ and maximizing the joint probability model above is at the heart of the E-M algorithm for example (see Appendix A). A sketch of a conceptually simple estimation method for this model based on Markov Chain Monte Carlo techniques instead is given in Appendix B.5.

Latent class models have been around for almost as long as psychometrics itself. For a history of their more modern treatement, see for example Bartholomew (1987). We have presented them here for two reasons: first, they will be helpful in describing Haertel and Wiley's approach to discrete attributes modeling, to which we turn next. Second, they provide a canonical example of the model structure needed to deal with multiple student strategies, to which we shall return in Section 6.

4.3.2 The Haertel/Wiley Restricted Latent Class Model

Haertel (1989) and Haertel and Wiley (1995) discuss a latent class model which represents a different way to try to stochasticize the deterministic model of Section 3.1.1. We begin again with a Q-matrix whose entries $q_{jk} = 1$ when attribute k is needed for response j, and $= 0$ when not.

We suppose that students can be classified into W latent classes, and students within the same latent class share the same knowledge state: they possess and lack exactly the same array of student attributes. Thus the attribute variables α_{ik} and the latent response variables ξ_{ij} used to conjunctively

combine attributes in the stochastic version of this model in Section 3.1.2, are assumed to be constant within each latent class:

$$\alpha_{wk} = \begin{cases} 1, & \text{if students in } C_w \text{ possess skill } k \\ 0, & \text{if not} \end{cases}$$

$k = 1, \ldots, K$; and for each task j we summarize the α's by

$$\xi_{wj} = \begin{cases} 1, & \text{if } \alpha_{wk} \geq q_{jk} \text{ for all } k = 1, \ldots, K \\ 0, & \text{if not} \end{cases}$$

i.e. $\xi_{wj} = 1$ indicates that all skills needed for problem j are present for students from latent class w. In the latent class model we set

$$p_{wj} = (1 - s_j)^{\xi_{wj}} g_j^{1-\xi_{wj}}$$

where s_j and g_j are per-task slip and guessing probabilities; the remainder of the model is identical to the latent class model as presented in equation (15) or (17). A discussion of a straightforward estimation method for this model is provided in Appendix B.5.

If the number of latent classes, W, is much smaller the number of rules K, the Haertel-Wiley model represents a great reduction in parametrization from the stochastic discrete attributes model of Section 3.1.2 and Figure 1; it is not neccessary to estimate the K parameters $P[\alpha_{ik} = 1 \mid \text{data}]$ for each attribute and student; rather we just estimate the W parameters $P[\text{Examinee in class } i \mid \text{data}]$. A serious challenge in actually using the model is figuring out what configurations of skills present and absent are really there in the student population.

A related assessment model is Polk, VanLehn and Kalp's (1995) ASPM2, which basically fits a very large version of the deterministic Q-model (4) by minimizing the Haming distance, that is, the number of mismatches, between the model's predicted performance on a set of tasks and a student's actual performance. Typically the fitting algorithm returns several different $\underset{\sim}{\alpha}_i$ vectors that predict the student's performance equally well.

Another way of "stochasticising" the deterministic model was developed independently, in another context, by Leenen, Van Mechelen, and Gelman (1999). Expressed in the assessment context, these authors basically treat the probability that a conjunctive deterministic model will match observed student performance on each task as independent Bernoulli events with a common match probability π, and develop a clever MCMC algorithm for selecting which sets of skills are most predictive of student performance in a given task domain.

Finally we note that the Haertel/Wiley model provides a way to operationalize Figure 5 on p. 19 of this report. Indeed, each configuration of skills in a latent class might be concieved of as a strategy, or level of competance, as in the figure. Then the Haertel-Wiley model might be able to directly estimate probabilities of membership in each strategy class.

4.3.3 The HYBRID Model

Yamamoto proposed a model (Gitomer and Yamamoto, 1991;, Yamamoto and Gitomer, 1993) that tries to integrate traditional IRT models with the Haertel-style cognitively-motivated latent class model, by using the coarser IRT modeling framework to account for lack-of-fit to the finer latent class assessment model.

A description of their HYBRID model can begin with a student variable

$$S_i = \begin{cases} 1 & \text{if student } i \text{ follows the latent class model} \\ 0 & \text{if not} \end{cases}$$

so that S_i codes whether or not student i "fits" the latent class models. Students that do fit are modeled as in the Haertel/Wiley framework, and students that don't are modeled using a coarser, unidimensional IRT model:

$$P[X_{ij} = 1 | S_i, \xi_{ij}, \underset{\sim}{\pi}, \{\text{IRT parameters}\}] = \begin{cases} \pi_{0j}^{1-\xi_{ij}} \pi_{1j}^{\xi_{ij}} & \text{if } S_i = 1 \\ P_j(\theta_i) & \text{if } S_i = 0 \end{cases}$$

where the probability model $\pi_{0j}^{1-\xi_{ij}} \pi_{1j}^{\xi_{ij}}$ for $S_i = 1$ is essentially the same as the Haertel/Wiley latent class model, and the model $P_j(\theta_i)$ for $S_i = 0$ is a standard IRT model such as the 1PL, 2PL or 3PL, depending on the relevant IRT parameters. Gitomer and Yamamoto (1991) provide an E-M algorithm for estimating a slightly different formulation of this model. An estimation algorithm similar to the Haertel/Wiley algorithm in Appendix B.5 could also be constructed.

Following the discussion of Figure 5 in Section 2.2.2, the point of including the IRT model is the hope that some students who don't fit the Haertel/Wiley model might still follow an IRT model in which increasing general proficiency leads to more items right. The IRT model thus provides an interpretation (albeit a cruder one) for at least some of the students who don't well-fit any of the structured latent classes.

4.3.4 DiBello and Stout's "Unified Model"

As noted in Section 2.2.3 above, DiBello, Stout and colleagues (DiBello, Stout, and Roussos, 1995; DiBello, Jiang, and Stout, 1999) have developed a multi-strategy model, which they call the "unified model" (UM), that generalizes several of the models we have considered here, and within which one can "play" with positivity, completness, multiple strategies and slips, by turning on and shutting off various parts of the model. Our description in terms of the stochastic Q-matrix model in Section 3.1.1 is mostly based on Roussos's (1994) account.

We begin with a model that is almost the UM; for easy reference we repeat the definitions of $\xi_{ij}^{(\ell)}$, and ϕ_{ij} from Section 3.1.2. If there are several strategies we will let $Q^{(\ell)}$ be the Q matrix for the ℓ^{th} strategy, and let α_{ik} indicate (0 or 1) whether skill k is present for student i. Then

$$\xi_{ij}^{(\ell)} = \begin{cases} 1, & \text{if } q_{jk}^{(\ell)} \leq \alpha_{ik} \; \forall \; k \\ 0, & \text{if not} \end{cases}$$

indicates whether all the skills are in place for student i for performing task j using strategy ℓ, i.e. $\xi_{ij}^{(\ell)}$ indicates whether strategy ℓ would be predicted to be successful if chosen. Let S_{ij} take values $\ell = 1, \ldots, \ell$ indicating which strategy student i chooses for task j; then probability that a student chooses a successful strategy is given by

$$\phi_{ij} = \sum_{\ell=1}^{\nu_j} \xi_{ij}^{(\ell)} P[S_{ij} = \ell]$$

Then a model similar to the UM can then be written as follows:

$$P[X_{ij}(t) = 1 | \alpha_{ik}(t), \ell = k, \ldots, K] = (1 - s_j)\phi_{ij}(t) = (1 - s_j) \left(\sum_{\ell=1}^{\nu_j} \xi_{ij}^{(\ell)} P[S_{ij} = \ell] \right)$$

To make this into the UM, we replace the $\xi_{ij}^{(\ell)}$ terms with a stochastic equivalent (see also the predictive equations (19) and (18) in the Corbett et al. model):

$$\xi_{ij}^{(\ell)} = \prod_{k: q_{jk}^{(\ell)} > \alpha_{ik}} \pi_{0ik\ell} \cdot \prod_{k: q_{jk}^{(\ell)} \leq \alpha_{ik}} \pi_{1ik\ell} \cdot P_{j\ell}(\underset{\sim}{\theta}_{i\ell})$$

where, analogously to the Haertel model,

$$\begin{aligned} \pi_{0ik\ell} &= P[\text{ student } i \text{ performs consistently with skill } k \text{ under strategy } \ell \mid \text{skill } k \text{ is missing}] \\ \pi_{1ik\ell} &= P[\text{ student } i \text{ performs consistently with skill } k \text{ under strategy } \ell \mid \text{skill } k \text{ is present}] \end{aligned}$$

and $P_{j\ell}(\underset{\sim}{\theta}_{i\ell})$ is a compensatory IRT response function.

The UM illustrates all four of the features that DiBello, Stout and Roussos (1995) discuss: *strategy* can be modeled item by item, using the ϕ_{ij}'s, *completeness* is indirectly represented in the model through the Q^{ℓ} matrices and their relationship to the underlying cognitive model; *postivity* is represented in the model parameters $\pi_{0ik\ell}$ and $\pi_{0ik\ell}$; and global slips are handled with the parameter s_j. The model can be adjusted to include a catchall strategy that is basically an IRT model, as in the HYBRID model discussed above, or to suppress IRT completely from the model.

As such the UM is rather complex and surely overparametrized. DiBello, Stout and Roussos (1995) discuss reasonable simplifying assumptions intended to make the model identifiable. Dibello, Jiang and Stout (1999) develop the model in detail, giving data analysis examples in which the model is fitted using an E-M algorithm in which the M-step, which involves optimizing over presence/absence of skills in the α vector, is accomplished using a genetic algorithm (e.g. Michalewicz et al., 1999).

5 Case Study: Two Approaches to Cognitive Assessment

To illustrate the differences between traditional IRT approaches and cognitively-motivated approaches to assessment, I briefly compare two published models intended to deal with essentially

the same data: task performance by students learning the LISP programming language using one of the computer based intelligent tutoring systems developed by John R. Anderson and his colleagues at Carnegie Mellon University (e.g. Anderson, Corbett, Koedinger and Pelletier, 1995). The first model is the assessment model actually embedded in the tutoring software, as described by Corbett, Anderson and O'Brien (1995); the second is an IRT-based model for essentially the same data, as presented by Draney, Pirolli and Wilson (1995).

5.1 The Corbett/Anderson/O'Brien Model

The first assessment model we will consider for this data is the "knowledge tracing model" embedded in the LISP tutor software, and described by Corbett, Anderson and O'Brien (1995). Using a notation similar to that of Section 3.1.2, we begin by defining

$X_{ij}(n)$	$=$	1 or 0	indicating whether or not student i performed task j correctly at time n
q_{jk}	$=$	1 or 0	indicating whether or not task j requires skill k
$\alpha_{ik}(n)$	$=$	1 or 0	indicating whether or not student i posesses skill k at time n
$\xi_{ij}(n)$	$=$	$\prod_{\{k:\ q_{jk}=1\}} \alpha_{ik}(n)$	indicating whether or not student i has the skills needed for task j at time n
s	$=$	$P[X_{ij}(n) = 0 \mid \xi_{ij}(n) = 1]$,	a universal slip parameter
g	$=$	$P[X_{ij}(n) = 1 \mid \xi_{ij}(n) = 0]$,	a universal guessing parameter

The model of task performance embodied by Corbett et al.'s (1995) knowledge tracing model is essentially

$$P[X_{ij}(n) = 1 \mid \alpha_{ik}(n), \ell = k, \ldots, K] = \xi_{ij}(n)(1 - s) + (1 - \xi_{ij}(n))g$$

This model is essentially identical to the model in equation (5), except that here s and g are universal rather than per-problem probabilities of slipping given that the strategy should be successful ($\xi_{ij} = 1$), and of guessing correctly, given that the strategy would not be successful ($\xi_{ij} = 0$). If we know the probabilities $P[\alpha_{ik}(n) \geq q_{jk}]$ that student i has skill k required for task j at time n, we can compute the probability $P[\xi_{ij}(n) = 1]$ that all the skills are in place for performing task j correctly at time n using the modeled strategy

$$P[X_{ij}(n) = 1] = P[\xi_{ij}(n) = 1](1 - s) + (1 - P[\xi_{ij}(n) = 1])g \tag{18}$$

where[22] $P[\xi_{ij}(n) = 1]$ follows a simple conjunctive model,

$$P[\xi_{ij}(n) = 1] = \prod_{k=1}^{K} P[\alpha_{ik}(n) \geq q_{jk}] \tag{19}$$

[22]The product formula in equation (19) assumes that the skills are present or absent independently of one another. Dependency among skills would be represented by replacing (19) with a more complex expression, perhaps coming from a simple Bayesian inference network (e.g. Mislevy, 1994) for the α's, for example.

It is worth noting that the $\xi_{ij}(n)$'s [or the $\alpha_{ik}(n)$'s] can be interpreted as playing the role of Maris's (1995) latent responses.

It is the probabilities $P[\alpha_{ik}(n) \geq q_{jk}]$, treating $\alpha_{ik}(n)$ as the unknown or random quantity, that are actually of primary interest in assessment based on this model of task performance. Given current estimates of $P[\alpha_{ik}(n) \geq q_{jk}]$, the tutor can both identify which skills need additional practice, and select items of suitable difficulty that exercise those skills, to assign to the student next.

Corbett, Anderson and O'Brien (1995) were particularly interested in how to gather evidence about $P[\alpha_{ik}(n) \geq q_{jk}]$ as the number of opportunities n to apply rule k increases—i.e. they are interested in modeling learning. They are able to observe responses (student performances of subtasks) involving each student attribute (production rule) separately within the ITS. Although there could in principle be dependency relations among the attributes, their model does not assume this; thus their model for collecting evidence about student attributes looks like Figure 9. In the figure, and in the text below, Attribute k is coded as present or not in student i at time n, according as $\alpha_{ik}(n) = 1$; otherwise $\alpha_{ik} = 0$, The α's cannot decrease as n increases. Similarly we will denote that student i performed an action correctly at time n when Attribute k was called for, by writing $a_{ik}(n) = 1$; otherwise $a_{ik}(n) = 0$. Because of guessing and slips, $a_{ik}(n)$ can increase or decrease as n increases. Also note that time n denotes the number of opportunities one has had to apply an attribute; thus as in the figure, n may have different values for different attributes (this is not reflected in the notation, since we will work with just one attribute at a time).

To account for the order in which correct and incorrect actions are observed when skill k is called for, they suggest treating $\alpha_{ik}(n)$ as a hidden Markov learning model with an absorbing state at "attribute learned". The transition matrix for the Markov chain is given below,

$$\begin{array}{rcl} P[\alpha_{ik}(n) \mid \overline{\alpha}_{ik}(n-1)] & = & T; \\ P[\alpha_{ik}(n) \mid \alpha_{ik}(n-1)] & = & 1; \\ P[\overline{\alpha}_{ik}(n) \mid \overline{\alpha}_{ik}(n-1)] & = & 1-T; \\ P[\overline{\alpha}_{ik}(n) \mid \alpha_{ik}(n-1)] & = & 0; \end{array} \tag{20}$$

where "$\alpha_{ik}(n)$" stands for "$\alpha_{ik}(n) = 1$" and "$\overline{\alpha}_{ik}(n)$" stands for "$\alpha_{ik}(n) = 0$".

The tutoring system is arranged to directly observe evidence $a_{ik}(n) = 0$ or 1 that the correct action was performed when skill k was called for, greatly simplifying the inferential task. Corbett, et al. (1995) posit the following relationships between the hidden state $\alpha_{ik}(n) = 0$ or 1 and the observable evidence $a_{ik}(n) = 0$ or 1:

$$\begin{array}{rcl} P[\alpha_{ik}(n) \mid a_{ik}(n)] & = & P[\alpha_{ik}(n-1) \mid a_{ik}(n)] + (1 - P[\alpha_{ik}(n-1) \mid a_{ik}(n)]) \cdot T \\ P[\alpha_{ik}(n) \mid \overline{a}_{ik}(n)] & = & P[\alpha_{ik}(n-1) \mid \overline{a}_{ik}(n)] + (1 - P[\alpha_{ik}(n-1) \mid \overline{a}_{ik}(n)]) \cdot T \\ P[\alpha_{ik}(n-1) \mid a_{ik}(n)] & = & \{(1-s)P[\alpha_{ik}(n-1)]\} \,/\, \{(1-s)P[\alpha_{ik}(n-1)] + gP[\overline{\alpha}_{ik}(n-1)]\} \\ P[\alpha_{ik}(n-1) \mid \overline{a}_{ik}(n)] & = & \{sP[\alpha_{ik}(n-1)]\} \,/\, \{sP[\alpha_{ik}(n-1)] + (1-g)P[\overline{\alpha}_{ik}(n-1)]\} \end{array} \tag{21}$$

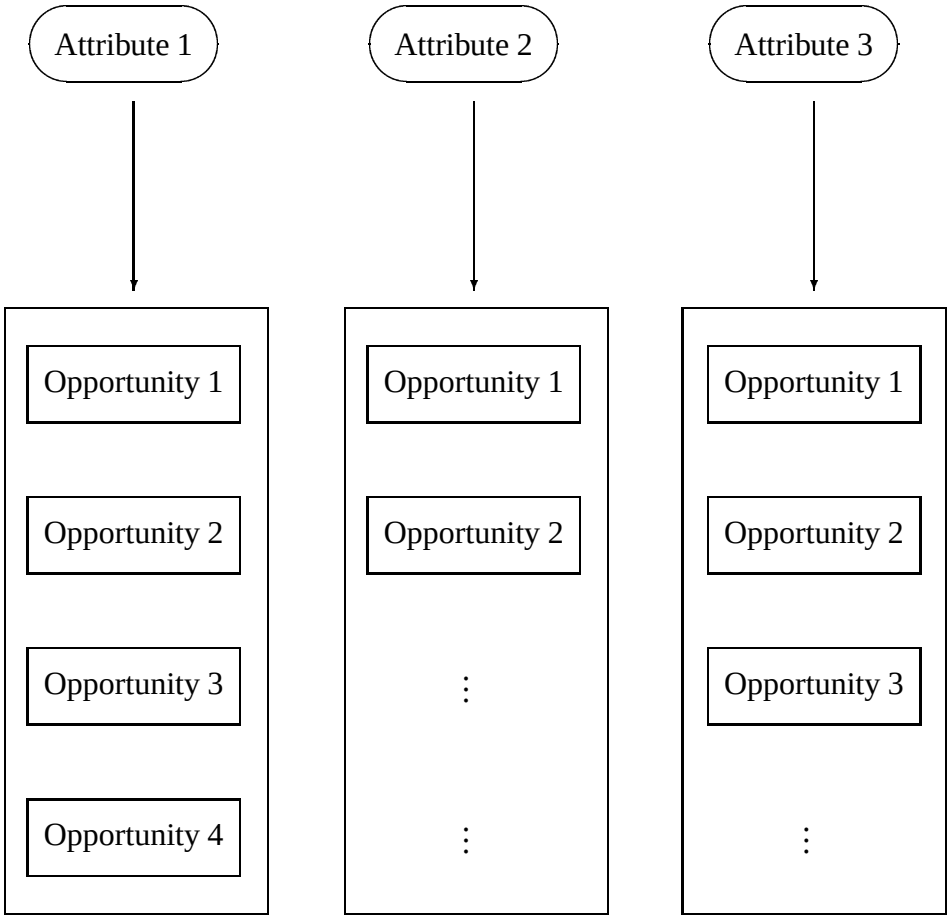


Figure 9: Arrangement of data collection for the Corbett, Anderson and O'Brien (1995) assessment model. Evidence for each attribute is collected as if independent of all other tasks and attributes, and is accumulated through separate hidden Markov learning models for each attribute, relating that attribute to repeated opportunities to perform the "skill" that the attribute names.

Given *a-priori* fixed values of T, s, g, and a probability p_0 that each skill is in the learned state before the tutoring begins, we may substitute p_0 for $P[\alpha_{ik}(n-1)]$ when $n = 1$, and the above formulas give an algorithm for recursively updating $P[\alpha_{ik}(n)]$ on the basis of the observed sequence of correct and incorrect actions in the first n opportunities to apply rule k. This is a particularly simple and fast computational method, capable of updating the tutor's model of the student's skills in real time as the student works with the tutor.

It is interesting to note that in order to produce a well-fitting model, Corbett, Anderson and O'Brien (1995) had to allow the probabilities T, s, g, and the probability p_0 that each skill was already in the learned state before the tutoring began, to be perturbed differently from overall population values[23] for each student (i); in statistical parlance we would say they allowed these parameters to become *random effects*. Thus, in addition to the individual differences in skills acquisition that the model had been designed to detect, there were substantial individual differences in starting knowledge of the students, in tendency to slip or guess, and in the rate of learning, under this model.

5.2 The Draney/Pirolli/Wilson Model

Draney, Pirolli and Wilson (1995) develop an LLTM-style model to analyze essentially the same data. The model they consider begins with an indicator $a_{ijk}(n) = 1$ if student i performs correctly when the n^{th} opportunity to use skill k occurs, under condition j; and $a_{ijk}(n) = 0$ otherwise. These $a_{ijk}(n)$ differ from Corbett et al.'s (1995) $a_{ik}(n)$ only in that the task that provides a context for performing the skill is allowed to affect the difficulty of correct skill performance. In the Draney et al. (1995) model, the "skill response functions" are given by

$$P[a_{ijk}(n) = 1 \mid \theta_i, \tau_j, \delta_k, \gamma] = \frac{1}{1 + \exp(-\theta_i + \tau_j + \delta_k - \gamma \log(n))}$$

where the logarithmic dependence on n is intended to follow the development of Anderson (1993, Appendix to Chapter 3). This model essentially decomposes the β parameter in the Rasch model according to a Q matrix, as in equations (14) and (2). For example, if there were eight observed behaviors, responding to the need for two different skills at two different times, performed within

[23] Except for computational details, Corbett et al. very nearly re-invented from scratch the hierarchical Bayes modeling framework (e.g. Gelman et al., 1995) to solve their model fitting problem here!

two different tasks, we might end up with the following decomposition:

$$\begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \\ \beta_7 \\ \beta_8 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & -\log(1) \\ 0 & 1 & 0 & -\log(2) \\ 0 & 0 & 1 & -\log(1) \\ 0 & 0 & 1 & -\log(2) \\ 1 & 1 & 0 & -\log(1) \\ 1 & 1 & 0 & -\log(2) \\ 1 & 0 & 1 & -\log(1) \\ 1 & 0 & 1 & -\log(2) \end{bmatrix} \cdot \begin{bmatrix} \tau \\ \delta_1 \\ \delta_2 \\ \gamma \end{bmatrix}$$

so that τ is a contrast for the tasks, δ_1 and δ_2 code the difficulties of the skills, and γ is the slope of the learning curve. If the decomposition of tasks into skills is complete, and the skills are of a suitable granularity, Anderson's (1993) ACT-R theory predicts that skill "performances" will be approximately independent[24] of one another, given the relevant difficulty and student parameters. This is essentially a statement of local independence, so that the "skill response functions" above may be multiplied together in the usual way to form an IRT likelihood.

It is interesting to compare the Corbett et al. and Draney et al. modeling approaches. For example, for a data set similar to that analyzed by Draney et al., the assessment model of Corbett, Anderson and O'Brien (1995, p. 32; see also Draney, Pirolli and Wilson, 1995, p. 115) would employ essentially four continuous latent variables, and 33 dichotomous latent attribute indicators, *per student tested*—in addition to 132 parameters to characterize features of the skills being assessed. Using their variation on standard IRT models, Draney et al. provided equivalent or better fit to learning curves employing *one* continuous latent variable per student tested (Draney, Pirolli and Wilson, 1995, p. 109), and 36 parameters for the skills being tested (Draney, Pirolli and Wilson, 1995, p. 115). Thus, if the goal is to model learning curves, clearly the more complex Corbett et al. model is not needed.

However, the uses to which the two models can be put, and the substantive interpretations of estimated parameters in the two models, are very different. The Corbett et al. model is immediately useful for diagnosing individual differences in student task performance behavior by relating it directly to specific skills in the task decomposition of their task domain, but can only indirectly assess the validity and reliability of the tasks in the assessment, through fit to learning curves data.

The Draney et al. model is not immediately useful for student diagnosis, since its student parameter is one-dimensional. After fitting the model, Draney et al. go back and rank students based on (empirical Bayes) estimated θ's[25], and compare estimated skill performance difficulties to the

[24] There is an important issue of granularity here that potentially affects both the Corbett and Pirolli models. While ACT-R predicts independent performance at a very small grain size, this grain size is probably smaller than any student attribute (production rule) in the LISP tutor (or any other Andersonian tutor). At larger grain sizes it may still be possible to obtain approximate independence among student attributes, but this is not guaranteed by the theory, and might not even be true for some students, depending on their prior experiences and expertise.

[25] In the context of learning curves analysis, the θ's essentially code students' initial facility in the task domain prior

students' aggregate θ distributon; indeed their Figure 5.1, p. 112, is very similar to our illustration in Figure 6 above. As we said before, such displays allow us to predict which skills a "typical" student with some fixed value of θ might be expected to perform correctly, and are very useful communication devices. However, detailed cognitive diagnosis of individual students on the basis of the θ's is not possible, without a post-hoc analysis of some sort. Indeed, for assessing whether individual students have learned particular skills, Draney et al. replace the IRT model with a Bayesian inference network that is focused on inferring the probability that an individual has learned one or more skills, using priors constructed from the fitted IRT model and new data from further attempts to perform the skill(s).

The utility of the Draney et al. model for identifying important task performance features should not be minimized however. For example, Koedinger, Lovett, Trottini, Junker, and Kass (in progress) are using essentially the same modeling framework to develop a semi-automatic stepwise variable construction/model selection procedure, with the goal of identifying skills that were either too narrowly or too broadly defined in a cognitive tutor (these result in stylized deviations, or "blips" from the theoretically predicted learning curves for the skills). In a similar vein, Huguenard, et al. (1997; see also Patz et al., 1996) applied a polytomous version of the LLTM to study the relationship between task features and working memory load, using IRT θ parameter to soak up residual between-subjects variation not modeled by the experimental and working memory factors.

6 Mixtures of Strategies

At several points in our discussion—Figure 5 of Section 2.2.2, data collection in Section 2.3, the latent class and related models of Section 4.3—we have encountered the problem of multiple strategies, and stumbled across some potential modeling solutions. We now wish to back up and consider the problem of accounting for multiple strategies in some detail. This section owes a lot in organization and point of view to the discussion of strategies in Mislevy, Steinberg and Almond (1997).

It is not difficult to believe that different students bring different problem-solving strategies to an assessment setting; sufficiently different curricular backgrounds provide a prima facie argument that this must happen; moreover comparative studies of experts and novices (e.g. Chi, Glaser and Farr, 1988) and theories of expertise (e.g. Glaser, 1991) suggest that, as in Figure 5, the strategies one uses to solve problems change as one's expertise grows. Kyllonen, Lohman and Snow (1984) show that strategy changes *within* a person between tasks also occur, and the evidence from ITS research is that it is not unusual for students to chance strategy within a task as well. We can distinguish then between at least four cases for modeling differential strategy use, in increasing order of difficulty for statistical modeling and analysis:

Case 0: No modeling of strategies.

to tutoring, much as the random effect version of p_0 does in the Corbett et al. model.

Case 1: Model strategy changes between persons.

Case 2: Model strategy changes between tasks, within persons.

Case 3: Model strategy changes within task, within persons.

Which case we take as our modeling task depends, as with all assessment modeling decisions, on tradeoffs between what students are actually doing and what the purpose of the assessment is. Returning to the science assessment of Baxter, Elder and Glaser (1996): we might decide to give the assessment specifically to identify which attributes of a high competance a particular student has, and then remediate positively toward the missing attributes without regard to what low competance attributes the child possesses; this could be an instance of Case 0. On the other hand if the goal was to identify the competancy level of the student—low, medium or high—and remediate accordingly, then a more complete between-persons model, as in Case 1, is called for. In addition, if the difficulty of the task depends strongly on the strategy used, we might be forced in to Case 1 or one of the other cases, to get an assessment model that fits the data well, even though the only valuable target of inference is the high-competance state.

Most models for Case 1—modeling strategy changes between students, but assuming that strategy is constant across assessment tasks—are variations on the latent class model of equation (15). As we have indicated earlier, the Haertel/Wiley latent class model, and Yamamoto's HYBRID model map well onto Figure 5; both include latent classes consisting of sets of attributes that are assumed to be present in all class members (and the Yamamoto model also contains an IRT model to help structure students who don't fit the latent class/discrete attributes part of the model.

If we replace the discrete-attributes models embedded within each strategy class in the Haertel/Wiley model with LLTM models, we obtain Mislevy and Verhelst's (1990) model for accounting for strategy effects on item difficulty in IRT: Let Q^w, $w = 1, \ldots, W$ be the Q matrices that relate student attributes to item difficulty under each of W different strategies; their basic approach is to postulate different Rasch models

$$P[X_{ij} = 1 | w, \underset{\sim}{\theta}, \underset{\sim}{\psi}] = \frac{1}{1 + \exp[\theta_{iw} - \beta_{wj}]}$$

where

$$\underset{\sim}{\beta}_w = Q^w \underset{\sim}{\psi}_w$$

for each latent strategy class. Note that the proficiency variable θ also depends on strategy. Mislevy and Verhelst (1990) provide an E-M algorithm for estimating this model (an MCMC algorithm along the lines of Appendix B.5 is also conceivable) and give an example from a spatial visualization task.

This approach basically exploits collateral information about the difficulties of the tasks under different strategies, to make inferences about what strategy is being used. Wilson's Saltus model (Wilson, 1989; Mislevy and Wilson, 1996) is quite similar, positing specific interactions on the θ

scale between items of a certain type and developmental stages of examineees. Either approach is likely to be succesful if the theory for positing differences between task difficulties unde different strategies produces some fairly large task difficulty differences across strategies; a similar point was made for the Haertel/Wiley model earlier.

The M^2RCML model of Pirolli and Wilson (1999) allows not only for mixing of strategy- or developmental-stage-based Q-matrices that drive item difficulty as in the Mislevy/Verhelst and Saltus models but also for mixing over various linear/compensatory combinations of student proficiency variables. Hence it too should be useful for Case 1 modeling of strategies.

Case 2, in which students change strategy from task to task, is more difficult. One example of a model intended to accomodate this is the "unified model" of Stout and DiBello. In fact, one can build a version of the Mislevy/Verhelst model that does much the same thing; we simply build the latent class model within task instead of between tasks. Let Q^{wj} be the Q-matrix (row-vector, really) for strategy w executed on item j, then

$$P[X_{ij} = 1|w, \underset{\sim}{\theta}, \underset{\sim}{\psi}_j] = \frac{1}{1 + \exp[\theta_{iw} - \beta_{wj}]}$$

where now

$$\beta_{wj} = Q^{wj}\underset{\sim}{\psi}_{wj}$$

The full model is not difficult to set up and it is not difficult to write down estimating equations for it. However it is very difficult to fit, because wrong/right, or even polytomously scored, responses, do not contain much information about the choice of strategy.

To make progress with Case 2, we must collect more data, as outlined in Section 2.3. Response latency in computerized tests; information about the performance of subtasks within a task; if that is informative about the strategy; asking students to answer strategy-related auxiliary questions, as did Baxter, Elder and Glaser (1996); asking students to explain the reasoning behind their answers; or even asking them directly what strategy they are using, can all be helpful. In the best case, this information-gathering reduces our assessment modeling problem to the case in which each student's strategy is known with certainty.

One might also make more progress with a stronger theory of task selection as well, but even here some additional information seems to be needed. For example if a developmental theory of progress from one stage to another exists for the domain, *and* if at least some tasks are known to be sensitive to boundaries between stages, then we can use these tasks to triangulate on a student's strategy.

Case 3, in which the student changes strategy within task, is impossible to model successfully without rich within-task data. Some ITS's systems try to do this, under the banner "model tracing" or "plan recognition. John Anderson's tutors generally do this by keeping students close to a modal solution path, but they have also experimented with directly asking students what strategy they are pursuing in cases of ambiguity (e.g. Anderson, Corbett, Koedinger and Pelletier, 1995). Others keep track of other environmental variables, to help disambiguate strategy choice within a

particular task performance (e.g. Hill and Johnson, 1995). Bayesian networks are commonly used for this purpose. The Andes tutor of mechanics problems in physics (e.g. Gertner, Conati and VanLehn, 1998) employs a Bayesian network to do model tracing. The student attributes are production rules; the observed responses are problem-solving actions; and strategy-use variables mediate the relationships between attributes and responses (inverting the relationship between "cluster" and "attribute" in Figure 2 for example). Various strategies have been proposed for controlling the combinatorial complexity as the number of possible strategies grows. Charniak and Goldman (1993) for example build a network sequentially, adding notes for new evidence with respect to plausible plans along the way.

7 Some Concluding Remarks

In recent years, as cognitive theories of learning and instruction have become richer, and computational methods to support assessment have become more powerful, there has been increasing pressure to make assessments truly criterion referenced, that is, to "report" on student achievement relative to theory-driven lists of examinee skills, beliefs and other cognitive features needed to perform tasks in a particular assessment domain. For example Baxter and Glaser (1998) and Nichols and Sugrue (1999) present compelling cases that assessing examinees' cognitive characteristics can and should be the focus of assessment design. In a similar vein, Resnick and Resnick (1992) advocate standards-referenced assessment closely tied to curriculum, as a way to inform instruction and enhance student learning.

Appropriate criterion-referenced testing can also be an effective teaching tool when embedded directly in teaching practice. Indeed there is substantial argument and evidence, as summarized for example by Bloom (1984), that part of what distinguishes higher student achievement in "mastery learning" and individualized tutoring settings as opposed to the conventional classroom, is the use of frequent and relatively unobtrusive formative tests coupled with feedback for the students and corrective interventions by the instructor, and followup tests to determine how much the interventions helped. This approach continues to be advocated as part of a natural and effective apprenticeship style of human instruction (e.g. Gardner, 1992), and it is the basis of many computer-based intelligent tutoring systems (ITS's, e.g. Anderson, 1993; and more broadly Shute and Psotka, 1996). Here too, a decomposition of assessment items into appropriate cognitive attributes is important: feedback and/or corrective action in a mastery class or from an ITS depends on knowing which cognitive attributes the examinee has mastered and which he or she has not.

Cognitive assessment models focused on these teaching and learning issues must generally deal with a more complex goal than linearly ordering examinees, or partially ordering them in a low-dimensional Euclidean space, which is what IRT has been designed and optimized to do. The goal of such assessments can be thought of producing, for each examinee, a sensibly organized checklist of skills or other cognitive attributes that the examinee may or may not possess, based on the evidence of tasks performed by the examinee.

Almost any assessment phenomenon—from between-examinee dependence due to institutional or sociological factors, to behavioral aspects of raters, to the analysis of item responses into requisite examinee attributes or item features—can be expressed in the hierarchical modeling framework, or the closely related Bayesian network framework, because of their great expressive power and conceptual simplicity. The first two sections of this report surveyed some important issues in developing a cognitive assessment system. The latter sections surveyed a wide variety of statistical models that might be used to make inferences about individual student attributes and proficiencies based on the assessment; even more models are out there (e.g. as surveyed by Roussos, 1994).

Recent advances in computation, and Markov chain Monte Carlo (MCMC) methods in particular, have made it possible to estimate a vastly wider variety of these models that would have been imaginable even ten years ago. However, the complexity of these models still puts us at the limits of our computing abilities. Speeding up the computations with approximations (including formal and informal applications of Laplace's method such as Rigdon and Tsutakawa, 1983 and Kass, Tierney and Kadane, 1990; blends of Monte Carlo and E-M approaches as surveyed in Tanner, 1996; and variational methods, e.g. Jaakkola and Jordan, 1999) continues to be an essential and fruitful avenue of research. On-the-fly assessment with these models is only possible if some parameters—the conditional probabilities in a Bayesian inference network, the difficulty and discrimination parameters of each item in an IRT-based computerized adaptive test—can be fixed in advance, either because they are so well estimated from past data that they can be considered known, or because the assessment outcomes are not particularly sensitive to their values within a certain range (VanLehn and Nui, 1999, illustrate such a sensitivity analysis applied to a complex assessment model embedded in an ITS).

Most of the illustrations in this report have been motivated by the information-processing style of cognitive psychology, and the related style of ITS's. The success of probabilistic reasoning in these domains is gratifying to me as an observer with an interest in statistics, but it is even more important as an illustration of the general way that the tools of probability-based reasoning can be applied to issues of uncertainty in any rational domain. The same kind of efforts that led to these advances in cognitive diagnosis can be brought to bear on assessment from any other psychological perspective as well. The nature of the student representation, the kinds of evidence, and the details statistical models and methods may be very different in their particulars. Guided by the substance and purpose of assessment in a new area, we frame our questions in terms of the the probability-based modeling framework presented here: what are the important features, what are the expected patterns, what sources of uncertainties might we expect, in what ways are these situations like these but unlike those? Fitting initial models based on these ideas to initial data will undoubtedly lead to improvements in the models, but it may also help refine our thinking about the substantive theory, think of better ways to make observations, and improved models.

8 References

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education, 7,* 255–278.

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29,* 67–91.

Adams, R. J., Wilson, M. and Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21,* 1–23.

Andersen, E. B. (1972). The numerical solution of a set of conditional estimation equations. *Journal of the Royal Statistical Society, Series B, 34,* 42–54.

Anderson, J. R. (ed.) (1993). *Rules of the mind.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Anderson, J. R., Corbett, A. T., Koedinger, K. R., and Pelletier, R. (1995). Cognitive tutors: lessons learned. *The Journal of the Learning Sciences, 4,* 167-207.

Baker, F. B. (1992). *Item response theory: Parameter estimation techniques.* New York: Marcel Dekker.

Bartholomew, D. J. (1987). *Latent variable models and factor analysis.* New York: Oxford University Press.

Baxter, G. P., Elder, A. D. and Glaser, R. (1996). Knowledge-based cognition and performance assessment in the science classroom. *Educational Psychologist, 31,* 133–140.

Baxter, G. P., and Glaser, R. (1998). Investigating the cognitive complexity of science assessments. *Educational Measurement: Issues and Practice, XX,* 37–45.

Béguin, A.A., Glas, C.A.W (1999). MCMC estimation of multidimensional IRT models. Research Report 98-14, Department of Educational Measurement and Data Analysis, University of Twente, the Netherlands. Available on the World Wide Web at `http://to-www.edte.utwente.nl/TO/omd/`.

Bloom, B. S. (1984). The 2-sigma problem: the search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher, 13,* 4–16.

Charniak, E. and Goldman, R. (1993). A Bayesian model of plan recognition. *Artificial Intelligence, 64,* 53-79.

Corbett, A. T., Anderson, J. R. and O'Brien, A. T. (1995). Student modeling in the ACT programming tutor. Chapter 2 in Nichols, P. D., Chipman, S. F. and Brennan, R. L. (eds.) (1995). *Cognitively diagnostic assessment.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Chi, M. T. H., Glaser, R. and Farr, M. (1988). *The nature of expertise.* Hilldate, NJ: Lawrence Erlbaum Associates.

Chib, S., and Greenberg, E. (1995). Understanding the Metropolis-Hastings Algorithm. *The American Statistician, 49,* 327–335.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12,* 671–684.

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

DiBello, L., Jiang, H. and Stout, W. F. (1999). A multidimensional IRT model for practical cognitive diagnosis. To appear, *Applied Psychological Methods.*

DiBello, L. V., Stout, W. F. and Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessement likelihoon-based classification techniques. Chapter 15 in Nichols, P. D., Chipman, S. F. and Brennan, R. L. (eds.) (1995). *Cognitively diagnostic assessment.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Dobra, A. (1999). *How big is the World Wide Web?* Advanced Data Analysis Project Report, Department of Statistics, Carnegie Mellon University, Pittsburgh PA.

Douglas, J. (1997). Joint consistency of nonparametric item characteristic curve and ability estimation. *Psychometrika, 62,* 7–28.

Draney, K. L., Pirolli, P. and Wilson, M. (1995). A measurement model for a complex cognitive skill. Chapter 5 in Nichols, P. D., Chipman, S. F. and Brennan, R. L. (eds.) (1995). *Cognitively diagnostic assessment.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Embretson, S. E. (1985). Multicomponent latent trait models for tests odesign. pp. 195–218 in Embretson, S. E. (ed.) (1985). *Test design: developments in psychology and psychometrics.* New York: Academic Press.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika, 56,* 495–515.

Embretson, S. E. (1995a). A measurement model for linking individual learning to process and knowledge: application to mathematical reasoning. *Journal of Educational Measurement, 32,* 277–294.

Embretson, S. E. (1995b). Developments toward a cognitive design system for psychological tests. Chapter 2 in Lubinski, D. and Dawis, R. V. (eds.) (1995). *Assessing individual differences in human behavior: new concepts, methods and findings.* Palo Alto CA: Davies-Black Publishing.

Embretson, S. E. (1999). *Generating items during testing: psychometric issues and problems.* Presidential Address to the 1999 European Meeting of the Psychometric Society, July 20, 1999, Seminaris Hotel, Lüneburg Germany.

Fienberg, S. E., Johnson, M. S. and Junker, B. W. (1999). Classical multilevel and Bayesian approaches to population size estimation using multiple lists. *Journal of the Royal Statistical Society, Series A, 162,* 383–405.

Fischer, G. H. (1973). Linear logistic test model as an instrument in educational research. *Acta Psychologica, 37,* 359–374.

Fischer, G. H., and Molenaar, I. W. (1995). *Rasch models: Foundations, recent developments, and applications.* New York: Springer-Verlag.

Fraser, C. and McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research 23,* 267–269.

Gardner, H. (1992). Assessment in context: the alternative to educational testing. pp. 77–119 in Gifford, B. R., and O'Connor, M. C. (eds.) (1992). *Changing assessments: alternative views of aptitude, achievement, and instruction.* Norwell, MA: Kluwer Academic Publishers.

Geman, S., and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6,* 721–741.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian data analysis.* New York: Chapman and Hall.

Gertner, A. S., Conati, C. and VanLehn, K. (1998). Procedural help in Andes: generating hints using a Bayesian network student model. pp. 106–111 in *Proceedings of the Fifteenth National Conference on Artificial Intelligence AAAI-98.* Cambridge MA: The MIT Press.

Gibbons, R. D. and Hedeker, D. R. (1997). Random effects probit and logistic regression models for three-level data. *Biometrics, 53,* 1527–1537.

Gitomer, D. H. and Yamamoto, K. (1991). Performance modeling that integrates latent trait and class theory. *Journal of Educational Measurement, 28,* 173–189.

Glas, C. A. W. and Verhelst, N. D. (1989). Extensions of the partial credit model. *Psychometrika, 54,* 635–659

Glaser, R. (1991). Expertise and assessment. pp. 17–30 in Wittrock, M. C. and Baker, E. L. (Eds.). (1991). *Testing and cognition.* Englewood Cliffs: Prentice-Hall.

Grayson, D. A. (1988). Two-group classification in latent trait theory: Scores with monotone likelihood ratio. *Psychometrika, 53,* 383–392.

Haberman, S. J. (1977). Maximum likelihood estimates in exponential response models. *Annals of Statistics, 5,* 815–841.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26,* 301–321.

Haertel, E. H. and Wiley, D. E. (1993). Representations of ability structures: implications for testing. Chapter 14 in Fredriksen, N. and Mislevy, R. J. (eds.) (1993). *Test theory for a new generation of tests.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Hambleton ,R. K. (1989). Principles and selected applications of item response theory. Chapter 4, pp. 147–200 in Linn, R. L. (Ed.) (1989). *Educational Measurement, Third Edition.* New York: American Council on Education and Macmillan Publishing Company.

Hastings, W. K. (1970). Monte Carlo simulation methods using Markov chains and their applications. *Biometrika, 57,* 97–109.

Hemker, B. T., Sijtsma K., Molenaar, I. W. and Junker, B. W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika, 62,* 331–347.

Hill, R. W. and Johnson, W. L. (1995). Situated plan attribution. *Journal of Artificial Intelligence in Education, 6,* 35–66.

Holland, P. W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika, 55,* 577–601,

Huguenard, B. R., Lerch, F. J., Junker, B. W., Patz, R. J. and Kass, R. E. (1997). Working memory failure in phone-based interaction. *ACM Transactions on Computer-Human Interaction, 4,* 67–102.

Hunt, E. (1995). Where and when to represent students this way and that way: an evaluation of approaches to diagnostic assessment. pp. 411–453, in Nichols, P. D., Chipman, S. F. and Brennan, R. L. (1995). *Cognitively diagnostic assessment.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Huynh, H. (1994). A new proof for monotone likelihood ratio for the sum of independent bernoulli random variables. *Psychometrika, 59,* 77–79.

Hrycek, T. (1990). Gibbs sampling in Bayesian networks. *Artificial Intelligence, 46,* 35–363.

Jaakkola, T. S., and Jordan, M. I. (1999). Bayesian parameter estimation via variational methods. In press, *Statistics and Computing.* Ms. obtained from the World Wide Web at address `http://www.cs.berkeley.edu/~jordan/`, September 1999.

Jameson, A. (1995). Numerical uncertainty management in user and student modeling: an overview of systems and issues. *User Modeling and User-Adapted Interaction, 5,* xxx-xxx.

Janssen, R. and de Boeck, P. (1997). Psychometric modeling of componentially designed synonym tasks. *Applied Psychological Measurement, 21,* 37–50.

Janssen, R. and de Boeck, P. (1996). The contribution of a response-production component to a free-response synonym task. *Journal of Educational Measurement, 33,* 417–432.

Johnson, E. G., Mislevy, R. J., and Thomas, N. (1994). Theoretical background and philosophy of NAEP scaling procedures. Chapter 8, pp. 133–146 in Johnson, E. G., Mazzeo, J. and Kline, D. L. (1994). *Technical Report of the NAEP 1992 Trial State Assessment Program in Reading.* Washington, DC: Office of Educational Research and Improvement, U.S. Department of Education.

Johnson, M. S., Cohen, W. and Junker, B. W. (1999). Measuring Appropriability in Research and Development with Item Response Models. CMU Statistics Department Technical Report #690. [WWW Document.] URL `http://www.stat.cmu.edu/cmu-stats/tr`.

Kass, R. E., Tierney, L. and Kadane, J. B. (1990). The validity of posterior expansions based on Laplace's method. In Geisser, S., Hodges, J. S., Press, S. J, and Zellner, A., ed's. (1990). *Bayesian and likelihood methods in statistics and econometrics: Essays in honor of George A. Barnard* (pp. 473–488). New York: North-Holland.

Koedinger, K., Lovett, M., Trottini, M., Junker, B. W., and Kass, R. E. (in progress). Searching for blips and hidden skills. Work in progress.

Kyllonen, P. C., Lohman, D. F. and Snow, R. E. (1984). Effects of aptitudes, strategy training, and test facets on spatial task performance. *Journal of Educational Psychology, 76,* 130–145.

Leenen, I., Van Mechelen, I., and Gelman, A. (1999). Bayesian probabilistic extensions of a deterministic classification model. Paper presented at an invited symposium on Bayesian Psychometrics (I. van Mechelen, H. Hoijtink and B. Junker, organizers), European Meeting of the Psychometric Society, July 1999, Lüneburg Germany.

Lesgold, A., Lajoie, S., Loga, D., and Eggan, G. (1990). Applying cognitive task analysis and research methods to assessment. Chapter 13 in Fredriksen, N., Glaser, R., Lesgold, A., and Shafto, M. G. (eds.) (1990). *Diagnostic monitoring of skill and knowledge acquisition.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Levidow, B., Hunt, E., and McKee, C. (1991). The DIAGNOSER: a HyperCard tol for building theoretically based tutorials. *Behavior Research Methods, Instruments and Computers, 23,* 249–252.

Liu, C. and Rubin, D. B. (1997). Maximum likelihood estimation of factor analysis using the ECME algorithm with complete and incomplete data. To appear, *Statistica Sinica*.

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

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

Maris, E. (1995). Psychometric latent response models. *Psychometrika, 60,* 523–547.

Martin, J. and VanLehn, K. (1995). A Bayesian approach to cognitive assessment. Chapter 7 in in Nichols, P. D., Chipman, S. F. and Brennan, R. L. (eds.) (1995). *Cognitively diagnostic assessment.* Hillsdale, NJ: Lawrence Erlbaum Associate.

Masters, G. N. and Evans, J. (1986). A sense of direction in criterion-referecned assessment. *Studies in Educational Evaluation, 12,* 257–265.

Masters G. N. and Forster, M. (1999). The Assessment Resource Kit. Camberwell, Victoria, Australia: Australian Council for Educational Research. Product description obtained from the World Wide Web at address `http://www.acer.edu.au/products/ed_resources/ark.html`, November, 1999.

Meng, X. L., and van Dyk, D. A. (1997). The EM algorithm—An old folk song sung to a fast new tune. *Journal of the Royal Statistical Society, Series B, 59,* 511–567.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equations of state space calculations by fast computing machines. *Journal of Chemical Physics, 21,* 1087–1091.

Michalewicz, Z., Esquivel, S., Gallard, R., Michalewicz, M., and Tau, G. (1999). *The spirit of evolutionary algorithms [some remarks on design of evolutionary algorithms]*. Paper presented at the 3rd On-line World Conference on Soft Computing in Engineering Design and Manufacturing (WSC3), Internet, June 21–30, 1998. Obtained from the World Wide Web at address `http://www.coe.uncc.edu/~zbyszek/`, September 1999.

Minstrell, J. (1998). Student thinking and related instruction: creating a facet-based learning environment. Working paper. Seattle WA: Assessment, Curriculum and Teaching Systems and Talaria Inc.

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika, 59,* 439–483.

Mislevy, R. J. and Bock, R. D. (1997). BILOG. Lincolnwood, IL: Scientific Software International. Product description obtained from the World Wide Web at address `http://ssi-central.com/product.htm`, September 1999.

Mislevy, R. J.; Sheehan, K. M. (1989). The role of collateral information about examinees in item parameter estimation. *Psychometrika, 54,* 661–679.

Mislevy, R. J., Steinberg, L., and Almond, R. G. (1997). On the design of complex assessments. Manuscript.

Mislevy, R. J. and Verhelst, H. (1990). Modeling item responses when different subjects emply different solution strategies. *Psychometrika, 55,* 195–215.

Mislevy, R. J. and Wilson, M. R. (1996). Marginal maximum likelihood estimation for a psychometric model of discontinuous development. *Psychometrika, 61,* 41–71.

Muraki, E. and Carlson, J. E. (1995). Full-information Factor Analysis for Polytomous Item Responses. *Applied Psychological Measurement, 19,* 73–90.

Nichols, P. and Sugrue, B. (1999). The lack of fidelity between cognitively complex constructs and conventional test development practice. *Educational Measurement: Issues and Practice, 18,* 18–29.

O'Connor, M. C. (1992). Overview: rethinking aptitude, achievement and instruction: cognitive science research and the framing of assessment policy. Pp. 9–35 in Gifford, B. R., and O'Connor, M. C. (eds.) (1992). *Changing assessments: alternative views of aptitude, achievement, and instruction.* Norwell, MA: Kluwer Academic Publishers.

Patz, R. J., Junker, B. W., Lerch, F. J. and Huguenard, B. R. (1996). Analyzing small psychological experiments with item response models. CMU Statistics Department technical report #644. [WWW Document.] URL `http://www.stat.cmu.edu/cmu-stats/tr`.

Patz, R. J. and Junker, B. W. (1999a). A straightforward approach to Markov Chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24,* 146–178.

Patz, R. J. and Junker, B. W. (1999b). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. In press, *Journal of Educational and Behavioral Statistics*. Also available as CMU Statistics Department Technical Report #670.

Patz, R. J., Junker, B. W. and Johnson, M. S. (1999). *The hierarchical rater model for rated test items and its application to large-scale educational assessment data.* Draft paper to accompany invited presentation April 23, 1999 at the Annual Meeting of the American Educational Research Association, Montreal Canada. [WWW document.] URL `http://www.stat.cmu.edu/~br`

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference.* San Mateo, CA: Morgan Kaufmann.

Pellegrino, J. W., Baxter, G. P., and Glaser, R. (1999). Addressing the "two disciplines" problem: linking theories of cognition and learning with assessment and instructional practice. To appear, *Review of Research in Education.*

Pirolli, P. and Wilson, M. R. (1999). A theory of the measurement of knowledge content, access, and learning. Working paper.

Polk, T. A., VanLehn, K. and Kalp, P. (1995). ASPM2: progress toward the analysis of symbolic parameter models. Chapter 6 in Nichols, P. D., Chipman, S. F. and Brennan, R. L. (eds.) (1995). *Cognitively diagnostic assessment.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Danish Institute for Educational Research.

Reckase, Mark D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9,* 401–412.

Resnick, L. B. and Resnick, D. P. (1992). Assessing the thinking curriculum: new tools for educational reform. pp 37–75 in Gifford, B. R., and O'Connor, M. C. (eds.) (1992). *Changing assessments: alternative views of aptitude, achievement, and instruction.* Norwell, MA: Kluwer Academic Publishers.

Rigdon, S. E. and Tsutakawa, R. K. (1983). Parameter estimation in latent trait models. *Psychometrika, 48,* 567–574.

Roskam, E. (1997). Models for Speed and Time-Limit Tests. Chapter 11 in van der Linden, W. J. and Hambleton, R. K. (eds.) (1997). *Handbook of modern item response theory.* New York: Springer Verlag.

Roudenbush, S. (1999). (Written materials supporting presentation to the Committee on the Foundations of Assessment, National Research Council, October 1, 1999).

Roussos, L. (1994). Summary and review of cognitive diagnosis models. Unpublished manuscript.

Sands, W. A., Waters, B. K., and McBride, J. R. (1997). *Computerized Adaptive Testing : From Inquiry to Operation.* Washington DC: American Psychological Association.

Scheiblechner, H. (1972). Das Lernen und Lösen komplexer Denkaufgaben. [The learning and solving of complex reasoning items.] *Zeitschrift für Experimentelle und Angewandte Psychologie, 3,* 456–506.

Shepard, L. P. (1992). Commentary: what policy makers should know about the new psychology of intellectual ability and learning. pp. 301–328 in Gifford, B. R., and O'Connor, M. C. (eds.) (1992). *Changing assessments: alternative views of aptitude, achievement, and instruction.* Norwell, MA: Kluwer Academic Publishers.

Shute, V. J., and Psotka, J. (1996). Intelligent tutoring systems: Past, Present and Future. Pp. 570–600 in D. Jonassen (Ed.), *Handbook of Research on Educational Communications and Technology*. Scholastic Publications.

Sijtsma, K. (1997). *Knowledge of solution strategies and IRT modeling of items for transitive reasoning.* Paper presented at the North American Meeting of the Psychometric Society, June 1997, Gatlinburg Tennessee.

Snow, R. E. and Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. Chapter 7, pp. 263–331 in Linn, R. L. (Ed.) (1989). *Educational Measurement, Third Edition.* New York: American Council on Education and Macmillan Publishing Company.

Spiegelhalter, D., Thomas, A., Best, N., and Gilks, W. (1996). *BUGS 0.5: Bayesian inference using Gibbs Sampling, Version ii*. Technical report of the MRC Biostatistics Unit, Institute of Public Health, Cambridge, UK. Available on the WWW at `http://www.mrc-bsu.cam.ac.uk/bugs`

Stegelmann, W. (1983). Expanding the Rasch model to a general model having more than one dimension. *Psychometrika, 48,* 259–267.

Tatsuoka, K. K. (1990). Toward an integration of item response theory and cognitive error diagnosis. Chapter 18 in Fredriksen, N., Glaser, R., Lesgold, A., and Shafto, M. G. (eds.) (1990). *Diagnostic monitoring of skill and knowledge acquisition.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: a statistical pattern recognition and classification approach. Chapter 14 in Nichols, P. D., Chipman, S. F. and Brennan, R. L. (eds.) (1995). *Cognitively diagnostic assessment.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Tanner, M. A. (1996). *Tools for statistical inference: methods for the exploration of posterior distriburtions and likelihood functions. 3rd Edition.* New York: Springer-Verlag.

Tanner, M. A., and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association, 82,* 528–550.

Thissen, D. M. (1976). Information in wrong responses to the Raven Progressive Matrices. *Journal of Educational Measurement, 13,* 201–214.

Thissen, D, (1997). MULTILOG. Lincolnwood, IL: Scientific Software International. Product description obtained from the World Wide Web at address `http://ssicentral.com/product.htm`, September 1999.

Thisted, R. A. (1988). *Elements of statistical computing: numerical computation.* NY: Chapman-Hall.

Tierney, L. (1994). Exploring posterior distributions with Markov Chains. *Annals of Statistics, 22,* 1701–1762.

Tsutakawa, R. K., and Soltys, M. J. (1988). Approximation for Bayesian ability estimation. *Journal of Educational Statistics, 13,* 117–130.

van der Linden, W. J. and Hambleton, R. K. (eds.) (1997). *Handbook of modern item response theory.* New York: Springer Verlag.

VanLehn, K. and Niu, Z. (1999). Bayesian student modeling, user interfaces and feedback: A sensitivity analysis. Submitted.

VanLehn, K., Niu, Z., Siler, S. and Gertner, A. (1998). Student modeling from conventional test data: a Bayesian approach without priors. pp. 434–443 in Goetl, B. et al. (Eds.) (1998). *Proceedings of the Intelligent Tutoring Systems Fourth International Conference, ITS 98.* Berlin, Hiedelberg: Springer-Verlag.

Wilson, D., Wood, R. L. and Gibbons, R. (1983). TESTFACT: test scoring and item factor analysis. [Computer program] Chicago: Scientific Software Inc. See also description at `http://ssicentral.com/product.htm`

Wu, M. L., Adams, R. J. and Wilson, M. R. (1997). *ConQuest: Generalized item response modeling software.* ACER.

Yamamoto, K. and Gitomer, D. H. (1993). Application of a HYBRID model to a test of cognitive skill representation. Chapter 11 in Fredriksen, N. and Mislevy, R. J. (eds.) (1993). *Test theory for a new generation of tests.* Hillsdale, NJ: Lawrence Erlbaum Associates.

A Some Estimation Methods for the LLTM and other IRT-like Models

In this appendix we will illustrate three currently popular ways of estimating modern IRT models, using the linear logistic test model (LLTM). The same general methods may be applied to any hierarchical Bayes model.

For the LLTM (and any other IRT model) these three approaches all begin with the product likelihood

$$L(\underset{\sim}{\mathbf{X}} \mid \underset{\sim}{\theta}, \underset{\sim}{\psi}) = \prod_i \prod_m P[X_m = 1 \mid \theta_i, \underset{\sim}{\psi}_m] \tag{22}$$

where $\underset{\sim}{\psi}_m$ is the subvector of parameters in $\underset{\sim}{\psi}$ relating to task m.

A.1 Conditional Maximum Likelihood

An approach called *conditional maximum likelihood* can be used; in this approach we calculate the conditional likelihood

$$L(\underset{\sim}{\mathbf{X}} \mid \underset{\sim}{\theta}, \underset{\sim}{\beta}, X_{1+}, X_{2+}, \ldots)$$

given the statistics $X_{i+} = \sum_m X_{im}$ which are sufficient statistics for θ_i when the β's are known. It turns out that this causes the parameters θ_i to drop out of the problem, and the remaining optimization over $\underset{\sim}{\beta}$ can be carried out using a combinatorial algorithm to compute the conditional likelihood in closed form and applying a variant of the Newton-Raphson or related methods (see e.g. Thisted, 1988, Chapter 4). See van der Linden and Hambleton (1997, Chapter 1), and especially the text by Fischer and Molenaar (1995) and the references therein, for details. This makes it possible for example to use the model to assess how good the task analysis was—are the ψ coefficients being estimated to be significantly different from zero, for example?

A.2 Marginal Maximum Likelihood and the E-M Algorithm

A second approach, called *marginal maximum likelihood* (MML)[26], has as its goal to produce maximum likelihood (or posterior mode) estimates of the parameters $\underset{\sim}{\beta}$, in the integrated, or marginal, likelihood,

$$\int L(\underset{\sim}{\mathbf{X}} \mid \underset{\sim}{\theta}, \underset{\sim}{\psi}) f(\underset{\sim}{\theta}) \mathrm{d}\underset{\sim}{\theta}$$

Up until recently the computational method of choice for MML was an expectation-maximization (E-M) algorithm (Dempster, Laird and Rubin, 1977; Tanner, 1996). In the most general setting, E-M consists of iterating the two steps

[26] Some authors have advocated the more accurate term *maximum marginal likelihood*, but putting "marginal" first is habitual in the literature, and the MML abbreviation is the same.

E-step Compute

$$G(\underset{\sim}{\psi}, \underset{\sim}{\psi}^{(h)}) = \int \log[L(\underset{\sim}{\mathbf{X}} \mid \underset{\sim}{\theta}, \underset{\sim}{\psi})] p(\underset{\sim}{\theta} \mid \underset{\sim}{\psi}^{(h)}, \underset{\sim}{\mathbf{X}}) d\underset{\sim}{\theta}$$

M-step Maximize $G(\underset{\sim}{\psi}, \underset{\sim}{\psi}^{(h)})$ in $\underset{\sim}{\psi}$ to obtain $\underset{\sim}{\psi}^{(h+1)}$

for $h = 1, 2, 3 \ldots$ times until $\|\underset{\sim}{\psi}^{(h+1)} - \underset{\sim}{\psi}^{(h)}\|$ or $G(\underset{\sim}{\psi}^{(h+1)}, \underset{\sim}{\psi}^{(h)})$ becomes sufficiently small. (See Tanner, 1996, pp. 70ff. for a straightforward demonstration that each iteration of E-M increases the marginal likelihood, and for some special simplifications of the algorithmx in the exponential family of probability models.) Because of the simple product form of $L(\underset{\sim}{\mathbf{X}} \mid \underset{\sim}{\theta}, \underset{\sim}{\psi})$, and because Q is usually a fairly sparse matrix, the M-step usually breaks down into a several uncoupled, low-dimensional maximizations. The integral in the E-step must usually be done numerically; this can be both a delicate and a slow feature of the algorithm.

The E-M method was used to estimate the model in Draney, Pirolli and Wilson (1995), and to estimate a model applied to a designed experiment on working memory capacity by Huguenard, Lerch, Junker, Patz, and Kass, (1997; see also Patz, Junker, Lerch and Huguenard, 1996). Variants of E-M exist for all the standard IRT models (e.g. Bartholomew, 1987; Baker, 1992), and it is currently the estimation method of choice for almost all small and large scale applications of IRT in education and psychology, largely due to the existence of computer programs like BILOG (Mislevy and Bock, 1997), MULTILOG (Thissen, 1997) and ConQuest (Wu, Adams, and Wilson, 1997).

Recently Embretson (1999) has started experimenting with a Q-matrix decomposition of the discrimination parameters in the 2PL model, in parallel with the Q matrix decomposition of the difficulty parameters β_j. In keeping with the interpretation of θ in these models as "nuisance" parameters that soak up variability not accounted for by the Q-matrix decomposition of task difficulty parameters, it is plausible for example that tasks for which the Q-matrix decomposition is incomplete will be more sensitive to variation in θ, and tasks for which the Q-matrix decomposition is more complete will be less sensitive to variation in θ. There does not exist a general CML algorithm for the 2PL model (nor indeed for most models outside the Rasch framework), so estimation must proceed via MML using E-M or some other method.

Holland (1990) surveys the literature comparing CML, MML and a less frequently recommended maximum likelihood approach called "joint maximum likelihood" (JML), which maximizes the likelihood (22) simultaneously in all θ and β parameters. This literature shows that, asymptotically as the number of students assessed grows, the CML and MML estimates of the β_j's will be indistinguishable. However, as Andersen (1972) and Haberman (1977) show, JML can and does lead to inconsistent parameter estimation unless the number of tasks and subjects both go to infinity at very carefully controlled rates (see also Douglas, 1997), so CML or MML are generally the preferred maximum likelihood approaches.

A.3 Markov Chain Monte Carlo

A third approach to estimating a model such as the LLTM is based on a very general computational method called *Markov Chain Monte Carlo* (MCMC, e.g. Gelman, Carlin, Stern, and Rubin, 1995). In our context, MCMC is tailored to compute approximations to the posterior distributions of model parameters, so instead of working with the likelihood (22) we work with a complete joint probability specification of the model

$$L(\underset{\sim}{\mathbf{X}} \mid \underset{\sim}{\theta}, \underset{\sim}{\psi}) \cdot \prod_i \pi(\theta_i) \cdot \prod_k \pi(\psi_k) \tag{23}$$

where the $\pi()$'s denote possibly different, but independent, prior densities for each unknown parameter. It is well known for example that this expression is proportional to the joint posterior distribution of the unknown parameters. For the purposes of describing the algorithm, let us rewrite the parameter vector as

$$\underset{\sim}{\eta} \stackrel{def}{=} (\underset{\sim}{\theta}, \underset{\sim}{\psi})$$

For any fixed partition $(\underset{\sim}{\eta}_1, \ldots, \underset{\sim}{\eta}_R)$ of the set of elements of the parameter vector η, MCMC proceeds by successively simulating random draws from the *complete conditional distributions* for each $\underset{\sim}{\eta}_r$: Thus after iteration h of the algorithm we might generate random draws

1. $\underset{\sim}{\eta}_1^{h+1} \sim p(\underset{\sim}{\eta}_1 | \underset{\sim}{\mathbf{X}}, \underset{\sim}{\eta}_2^{(h)}, \ldots, \underset{\sim}{\eta}_R^{(h)})$;
2. $\underset{\sim}{\eta}_2^{(h+1)} \sim p(\underset{\sim}{\eta}_2 | \underset{\sim}{\mathbf{X}}, \underset{\sim}{\eta}_1^{(h+1)}, \underset{\sim}{\eta}_3^{(h)}, \ldots, \underset{\sim}{\eta}_R^{(h)})$;
3. $\underset{\sim}{\eta}_3^{(h+1)} \sim p(\underset{\sim}{\eta}_3 | \underset{\sim}{\mathbf{X}}, \underset{\sim}{\eta}_1^{(h+1)}, \underset{\sim}{\eta}_2^{(h+1)}, \underset{\sim}{\eta}_3^{(h)}, \ldots, \underset{\sim}{\eta}_R^{(h)})$;

   $\vdots$

R. $\underset{\sim}{\eta}_R^{(h+1)} \sim p(\underset{\sim}{\eta}_1 | \underset{\sim}{\mathbf{X}}, \underset{\sim}{\eta}_1^{(h+1)}, \ldots, \underset{\sim}{\eta}_{R-1}^{(h+1)})$.

(Other orders, including randomly selecting steps from this list, are possible and can lead to more efficient simulations). The resulting partition $(\underset{\sim}{\eta}_1^{(h+1)}, \ldots, \underset{\sim}{\eta}_R^{(h+1)})$ is one step in a simulated Markov Chain whose stationary distribution is precisely the joint posterior distribution of $\underset{\sim}{\eta}$. Thus if we run the chain to stationarity, the succeding draws can be treated as (dependent) draws from this joint posterior, which is the object of our inference.

Standard names for common implementations of MCMC have come into use in the statistical community: If we implement an MCMC algorithm by simulating direcly from the complete conditional distributions $p(\underset{\sim}{\eta}_r \mid \underset{\sim}{\mathbf{X}}$, *coordinates of* η *not including* $\underset{\sim}{\eta}_r)$, the resulting Markov chain is called a *Gibbs Sampler*, following Geman and Geman (1984). If the complete conditionals are intractable for direct simulation, a clever modification of the classic Monte Carlo sampling scheme called "rejection sampling" can be used, leading to a Markov chain that is called a *Metropolis-Hastings Sampler* (Metropolis et al., 1953; Hastings, 1970). See Chib and Greenberg (1995) for a clear exposition of the methodology, with references for the underlying theory.

For each disjoint set of parameters $\underset{\sim}{\eta}_r$ in the partition $(\underset{\sim}{\eta}_1, \ldots, \underset{\sim}{\eta}_R)$, it is not difficult to see that the corresponding complete conditional distribution is proportional to the product of factors containing elements of $\underset{\sim}{\eta}_r$ in the joint probability model (23). For the same reason that the M-step in the E-M algorithm for IRT models usually breaks down into a set of low-dimensional maximizations—the likelihood is in product form and the Q matrix is usually sparse—it is usually possible to choose a partition of $\underset{\sim}{\eta}$ leading to low-dimensional complete conditionals, which usually eases the programming burden for simulation. For IRT models, the partition often involves singleton sets containing each θ_i and low-dimensional sets of β's or the underlying ψ's if the β have a Q-matrix decomposition.

For a large class of statistical models in and out of psychometrics and assessment, the process of selecting a partition and setting up a Gibbs sampler is so straightforward that a computer program (Spiegelhalter et al., 1996) endowed with some intelligent rules can set up and run a reasonable Gibbs sampler by directly inspecting a description of the joint probability model (23). A description of MCMC in general and Metropolis-Hastings in particular for IRT models is given in Patz and Junker (1999a); other applications of the methods to IRT-like data analysis problems are given in Patz and Junker (1999b), Patz, Junker and Johnson (1999) and Johnson, Cohen and Junker (1999). Patz and Junker (1999a) also briefly discuss the relationship between MCMC methods and the MML and JML formulations of the IRT estimation problem.

Generally speaking MCMC is more flexible than, but slower than, E-M, in IRT-like models. The greater flexibility comes from two places: (1) unlike E-M, neither differentiation nor explicit numerical integration are required to set up the algorithm; and (2) setting up the sampling from the complete conditionals is substantially more straightforward, and less dependent on regularity of the likelihood, than is maximization in E-M. For example, the manipulations involved in setting up an E-M algorithm are generally much easier if the factors in the product likelihood (22) are all from the same parametric family of functions. This is generally not a problem with MCMC, as illustrated by Patz and Junker (1999b). The lower speed comes primarily from the fact that MCMC is effectively estimating the entire posterior distribution by sampling from it (when the chain reaches stationarity), whereas E-M and other maximation methods are yielding only a posterior mode and perhaps a measure of posterior standard error. A variety of methods intermediate—in technique and in speed—between E-M and MCMC are surveyed in Tanner (1996; see also the variants of E-M surveyed by Liu and Rubin, 1997; and by Meng and van Dyk, 1997) as well. In models that look rather different from IRT models—for example Bayes Networks whose exact evaluation can be NP-hard—MCMC methods can be faster than deterministic methods for a given degree of accuracy of estimation (e.g. VanLehn and Niu, 1999).

B Talk given to the Committee, 1 October 1999

Some statistical models and computational methods that may be useful for cognitively-relevant assessment

Brian Junker[27]
Department of Statistics
Carnegie Mellon University
232 Baker Hall
Pittsburgh PA 15213
brian@stat.cmu.edu

Prepared for the Committee on the Foundations of Assessment, National Research Council.

B.1 Preface

- I want to do two things in this paper/presentation:

  1. Show a variety of statistical assessment models, especially arising out of the psychometric/statistical tradition.
  2. Indicate the flexibility of some current computational tools for estimating parameters and making inferences in these assessment models.

- There are a dizzying variety of models, and of computational burdens the models place on users.

> *Attend to the purposes of assessment. If your purpose doesn't require the additional complexity, don't build it into the assessment model.*

B.2 Outline of talk

1. *"Cliff's Notes" on Markov Chain Monte Carlo (MCMC)*
   - The only method I will discuss.
2. *Item response models*
   - Schematic MCMC for the 2-parameter logistic (2PL) model from item response theory (IRT).
   - Extension to the linear logistic test model (LLTM).
3. *Latent Class models*

[27] On leave at Learning Research and Development Center, 3939 O'Hara Street, University of Pittsburgh, Pittsburgh PA 15260.

- Schematic MCMC for latent class models
- Extension to the Haertel-Wiley restricted latent class model.

4. *Variant of the Corbett/Anderson/O'Brien model*
   - Simplification of the Corbett et al. model
   - Schematic MCMC algorithm
5. *Closing Thoughts:* Decoupling, credit/blame, complexity.

B.3 Bayes and MCMC: The Cliff's Notes

B.3.1 Data

: J tasks and N students generate an $N \times J$ matrix $\mathcal{X}$ of 0's and 1's, $X_{ij} = 1$ or 0 indicating correctness of response.

B.3.2 Parameters

: $\tau = (\underset{\sim}{\theta}, \underset{\sim}{\beta}, \underset{\sim}{\lambda})$, e.g. students, tasks, higher-order par's...
Basic identities:

- $p(X, \tau) \;=\; p(X|\tau) \cdot p(\tau)$
- $p(\tau|X) \;=\; \dfrac{p(X|\tau) \cdot p(\tau)}{\int p(X|t) \cdot p(t)\ dt} \;\propto\; p(X|\tau) \cdot p(\tau) \;=\; p(X, \tau)$

We always want to know something about $p(\tau|X)$:

- EAP: $argmax_\tau\, p(\tau|X)$
- MAP: $E[\tau|X] = \int \tau\, p(\tau|X)\, d\tau$
- CI: Find a set A such that $p(\tau \in A|X) = 0.95$
- Graph or "shape" of $p(\tau|X)$

The problem:

- Learn about $\pi(\tau) = p(\underset{\sim}{\theta}, \underset{\sim}{\beta}, \underset{\sim}{\lambda}|\mathcal{X})$, where $\tau = (\underset{\sim}{\theta}, \underset{\sim}{\beta}, \underset{\sim}{\lambda})$ is some high-dimensional set of variables (parameters).

The essential idea:

- Define a (stationary) Markov chain $\mathcal{M}_0, \mathcal{M}_1, \mathcal{M}_2, \ldots$ with states $\mathcal{M}_h = (\underset{\sim}{\theta}^h, \underset{\sim}{\beta}^h, \underset{\sim}{\lambda}^h)$; under regularity conditions (e.g., Tierney, 1994), $\mathcal{M}_h$ will converge in distribution to a *stationary distribution*, $\pi(\underset{\sim}{\theta}, \underset{\sim}{\beta}, \underset{\sim}{\lambda})$.

- Simulate the $\mathcal{M}_h$'s; sample statistics of these will approach sample statistics of $\pi(\underset{\sim}{\theta}, \underset{\sim}{\beta}, \underset{\sim}{\lambda})$.
- For Bayes, design chain so that $\pi(\underset{\sim}{\theta}, \underset{\sim}{\beta}, \underset{\sim}{\lambda})$ turns out to be the posterior distribution $p(\underset{\sim}{\theta}, \underset{\sim}{\beta}, \underset{\sim}{\lambda}|\mathcal{X})$.

The magic:

- Let $(\tau_1, \tau_2, \ldots, \tau_M)$ be a fixed, disjoint partition of $(\underset{\sim}{\theta}, \underset{\sim}{\beta}, \underset{\sim}{\lambda})$.
- General MCMC theory (e.g., Tierney, 1994; Chib and Greenberg, 1995): construct state $\mathcal{M}_h = (\tau_1^{(h)}, \ldots, \tau_M^{(h)})$, by sampling each τ_m from its "complete conditional" distribution:

To step from $\mathcal{M}_{h-1} = (\tau_1^{(h-1)}, \ldots, \tau_M^{(h-1)})$ *to* $\mathcal{M}_h = (\tau_1^{(h)}, \ldots, \tau_M^{(h)})$*:*

1. $\tau_1^{(h)} \sim p(\tau_1 | \tau_2^{(h-1)}, \ldots, \tau_M^{(h-1)}, X)$;
2. $\tau_2^{(h)} \sim p(\tau_2 | \tau_1^{(h)}, \tau_3^{(h-1)}, \ldots, \tau_M^{(h-1)}, X)$;
3. $\tau_3^{(h)} \sim p(\tau_3 | \tau_1^{(h)}, \tau_2^{(h)}, \tau_4^{(h-1)}, \ldots, \tau_M^{(h-1)}, X)$;

$\vdots \qquad \vdots \qquad \vdots \qquad \vdots$

M. $\tau_M^{(h)} \sim p(\tau_M | \tau_1^{(h)}, \tau_2^{(h)}, \tau_3^{(h)}, \ldots, \tau_{M-1}^{(h)}, X,)$;

- Gibbs, Metropolis-Hastings...

Suppose the θ's, β's and λ's have prior distributions $p(\theta|\lambda)$, $p(\beta)$ and $p(\lambda)$. Then the "complete conditional" for θ, for example, is

$$\begin{aligned} p(\theta|\textit{rest}) \;=\; p(\theta|X,\beta,\lambda) \;&=\; \frac{p(X,\theta,\beta,\lambda)}{\int p(X,t,\beta,\lambda)\,dt} \\ &=\; \frac{p(X|\theta,\beta)\,p(\theta|\lambda)\;p(\beta)\;p(\lambda)}{\int p(X|t,\beta)\,p(t|\lambda)\,p(\beta)\,p(\lambda)\,dt} \\ &\propto\; p(X|\theta,\beta)\,p(\theta|\lambda) \end{aligned}$$

The shape of $p(\theta|X,\beta,\lambda)$ *is determined by just the parts of the likelihood that depend explicitly on* θ*.*

B.4 Item Response Models

- Not by themselves cognitively useful; but simple enough to give flavor of MCMC quickly.
- A basic model is the two-parameter logistic (2PL):

$$P_j(\theta_i;\alpha_j,\beta_j) \equiv P[X_{ij}=1|\theta_i,\alpha_j,\beta_j] = \frac{1}{1+\exp(-\alpha_j[\theta_i-\beta_j])}$$

• Letting $\pi_\theta()$, $\pi_\alpha()$, $\pi_\beta()$ denote prior densities of parameters, the complete joint probability model is

$$\begin{aligned}
&P[\mathcal{X} \mid \underset{\sim}{\theta}, \underset{\sim}{\alpha}, \underset{\sim}{\beta}] \\
=&\prod_{i=1}^{N}\prod_{j=1}^{J} P_j(\theta_i;\alpha_j,\beta_j)^{x_{ij}}[1-P_j(\theta_i;\alpha_j,\beta_j)]^{1-x_{ij}} \cdot \prod_{i=1}^{N}\pi_\theta(\theta_i)\prod_{j=1}^{J}\pi_\alpha(\alpha_j)\pi_\beta(\beta_j) \\
=&\prod_{i=1}^{N}\left\{\prod_{j=1}^{J} P_j(\theta_i;\alpha_j,\beta_j)^{x_{ij}}[1-P_j(\theta_i;\alpha_j,\beta_j)]^{1-x_{ij}} \cdot \pi_\alpha(\alpha_j)\pi_\beta(\beta_j)\right\}\pi_\theta(\theta_i)
\end{aligned}$$

The complete conditional distributions from which we would construct an MCMC algorithm are:

$$p(\alpha_j|\textit{rest}) \propto \prod_{i=1}^{N} P_j(\theta_i;\alpha_j,\beta_j)^{x_{ij}}[1-P_j(\theta_i;\alpha_j,\beta_j)]^{1-x_{ij}} \cdot \pi_\alpha(\alpha_j) \tag{24}$$

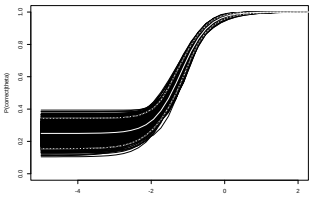
$$p(\beta_j|\textit{rest}) \propto \prod_{i=1}^{N} P_j(\theta_i;\alpha_j,\beta_j)^{x_{ij}}[1-P_j(\theta_i;\alpha_j,\beta_j)]^{1-x_{ij}} \cdot \pi_\beta(\beta_j) \tag{25}$$

$$p(\theta_i|\textit{rest}) \propto \left\{\prod_{j=1}^{J} P_j(\theta_i;\alpha_j,\beta_j)^{x_{ij}}[1-P_j(\theta_i;\alpha_j,\beta_j)]^{1-x_{ij}}\right\}\pi_\theta(\theta_i) \tag{26}$$

• Requires a standard rejection sampling "trick" called *Metropolis-Hastings within Gibbs*.

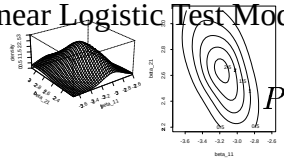
• Examples and illustrations in Patz and Junker (1999a,b).

(Data extract from NAEP 1992 Trial State Reading Assessment)



B.4.1 Item Response Models: LLTM

The Linear Logistic Test Model (LLTM; Scheiblechner, 1972; Fischer, 1973):



$$P[X_{ij}=1 \mid \theta_i, \beta_j] = \frac{1}{1+\exp(-[\theta_i-\beta_j])}$$

where the vector $\underset{\sim}{\beta}$ is linearly constrained,

$$\underset{\sim}{\beta}_{J\times 1} = q_{J\times K}\underset{\sim}{\psi}_{K\times 1} \tag{27}$$

and the entries q_{jk} of the "bookkeeping" matrix Q are

$$q_{jk} = \begin{cases} 1, & \text{if attribute } k \text{ is required by task } j \\ 0, & \text{if not} \end{cases} \tag{28}$$

where the attribute can be a cognitive skill, a surface feature of the item, etc. (e.g. Fischer and Molenaar, 1995; Draney et al., 1995; Huguenard et al., 1997; Embretson, 1995b; 1999).

The LLTM modifies the simpler IRT structure in two ways:

1. The α_j are taken to be identically equal to 1
2. The β_j have a linear structure, $\underset{\sim}{\beta} = Q\underset{\sim}{\psi}$

Modification #1 $\Rightarrow$ complete conditionals for α_j in equation (24) not needed; $\alpha_j \equiv 1$.

Modification #2 $\Rightarrow$ replace complete conditionals for β_j is equation (25) with complete conditionals for ψ_k:

Let $\mathcal{J}_k = \{$ j: β_j depends on $\psi_k\}$. Then

$$\begin{aligned} p(\psi_k|\textit{rest}) \quad \propto \quad & \prod_{i=1}^{N} \prod_{j \in \mathcal{J}_k} P_j(\theta_i; \beta_j^{(*,k)}(\psi_k))^{x_{ij}} [1 - P_j(\theta_i; \beta_j^{(*,k)}(\psi_k))]^{1-x_{ij}} \\ & \cdot \ \pi_\psi(\psi_k) \end{aligned} \tag{29}$$

where $\beta_j^{(*,k)}(\psi_k)$ is the j^{th} element of the vector

$$\underset{\sim}{\beta}^{(*,k)}(\psi_k) = Q \cdot \begin{bmatrix} \psi_1^* \\ \vdots \\ \psi_{k-1}^* \\ \psi_k \\ \psi_{k+1}^* \\ \vdots \\ \psi_K^* \end{bmatrix}.$$

where the starred ψ^*'s emphasize that all the ψ's except ψ_k are fixed. After all ψ_k's have been sampled, we compute $\underset{\sim}{\beta} = Q\underset{\sim}{\psi}$ for use in the other steps.

B.4.2 Item Response Models: Simple IRT vs. LLTM

There are three things to note about the MCMC schematic for simple IRT models and its modification for LLTM:

1. The simple IRT schematic is straightforward because factors in the likelihood are relatively "*decoupled*": parameters affect only one "dimension" (students or tasks) of the likelihood at a time.

2. The extension to LLTM is conceptually straightforward, and not much worse in practice. See for example the application of a similar model to multiple ratings of performance tasks in Patz and Junker (1999b).

3. The introduction of common parameters ψ_k that underlie several tasks *"recouples"* some factors in the likelihood: the terms in the complete conditionals (29) for ψ_k extend over both "dimensions" (students and tasks). *OK if Q is fairly sparse.*

B.5 Latent Class Models

- LLTM allows skills decomposition of *tasks*, latent class (LC) approach allows us to start assessing $P[\text{student has skill}]$, not just general θ.
- Schematic MCMC algorithm for LC illustrates a general technique for *"decoupling"* factors in a likelihood when they are coupled in a certain common way.
- Adaptation to the constrained latent class models of Haertel (1989) and Haertel and Wiley (1995) illustrates again how underlying skill parameters can *"recouple"* the likelihood factors.

B.5.1 Latent Class Models

Suppose students can be classified into W latent classes C_w, $w = 1, \ldots, W$, let

$$\begin{aligned} \lambda_w &= P[\text{student } i \text{ is in class } C_w] \\ p_{wj} &= P[X_{ij} = 1 | \text{student } i \text{ is in class } C_w] \end{aligned}$$

and as usual $X_{ij} = 1$ or 0 indicating correct performance of task j by student i. Since we do not know student i's latent class, the model for one student is

$$P[\underset{\sim}{\mathbf{X}}_i = \underset{\sim}{\mathbf{x}}_i \mid \underset{\sim}{\mathbf{p}}, \underset{\sim}{\lambda}] = \sum_{w=1}^{W} \lambda_w \prod_{j=1}^{J} p_{wj}^{x_{ij}} [1 - p_{wj}]^{1-x_{ij}}$$

and for an $N \times J$ matrix of response data $\mathcal{X}$, we obtain

$$P[\mathcal{X} \mid \underset{\sim}{\mathbf{p}}, \underset{\sim}{\lambda}] = \prod_{i=1}^{N} \left\{ \sum_{w=1}^{W} \lambda_w \prod_{j=1}^{J} p_{wj}^{x_{ij}} [1 - p_{wj}]^{1-x_{ij}} \right\} \tag{30}$$

- Not a product form for the likelihood
- *No useful decoupling at all!*
- Even conventional maximum likelihood and E-M methods become unwieldy as N, J and W grow large.

Decoupling technique: *data augmentation.*

Let

$$z_{iw} = \begin{cases} 1, & \text{if student } i \text{ is in latent class } w \\ 0, & \text{if not} \end{cases}$$

$\underset{\sim}{\mathbf{z}}_i = (z_{i1}, \ldots, z_{iW}) = (0, 0, \ldots, 1, 0, 0, \ldots 0)$ with a 1 only in position w, student i's assigned latent class.

So the model for the i^{th} student is now

$$\begin{aligned} P[\underset{\sim}{\mathbf{x}}_i, \underset{\sim}{\mathbf{z}}_i | \underset{\sim}{\mathbf{p}}, \underset{\sim}{\lambda}] &= P[\underset{\sim}{\mathbf{z}}_i | \underset{\sim}{\lambda}] \cdot P[\underset{\sim}{\mathbf{x}}_i | \underset{\sim}{\mathbf{z}}_i, \underset{\sim}{\mathbf{p}}, \underset{\sim}{\lambda}] \\ &= \left\{ \prod_w \lambda_w^{z_{iw}} \right\} \cdot \prod_w \left\{ \prod_j p_{wj}^{x_{ij}} [1 - p_{wj}]^{1 - x_{ij}} \right\}^{z_{iw}} \end{aligned}$$

so a joint probability model for $\mathcal{X}_{N \times J}$, $\mathcal{Z}_{N \times W}$, and the parameters, is

$$\begin{aligned} & P[\mathcal{X}, \mathcal{Z} | \underset{\sim}{\mathbf{p}}, \underset{\sim}{\lambda}] \pi(\underset{\sim}{\mathbf{p}}) \pi(\underset{\sim}{\lambda}) \\ &= \prod_i \prod_w \left\{ \lambda_w \cdot \prod_j p_{wj}^{x_{ij}} [1 - p_{wj}]^{1 - x_{ij}} \right\}^{z_{iw}} \pi(\underset{\sim}{\lambda}) \prod_j \pi_p(p_{wj}) \\ &= \left\{ \prod_w \lambda_w^{n_w} \right\} \pi(\underset{\sim}{\lambda}) \left\{ \prod_j p_{wj}^{c_{wj}} [1 - p_{wj}]^{n_w - c_{wj}} \pi_p(p_{wj}) \right\} \end{aligned}$$

where $n_w = \sum_i z_{iw}$, and $c_{wj} = \sum_i z_{iw} x_{ij}$ (and $\pi(\underset{\sim}{\lambda})$ and $\pi_p(p_{wj})$ are priors). *We have product structure and decoupling back!*

An MCMC algorithm *(even Gibbs!)* can be based on the following complete conditional distributions:

- From the second line of the complete probability model

$$p(p_{wj} | \textit{rest}) \propto p_{wj}^{c_{wj}} (1 - p_{wj})^{n_w - c_{wj}} \pi_p(p_{wj})$$

- From the second line again we can see that

$$p(\lambda_1, \ldots, \lambda_W | \textit{rest}) \propto \prod_w \lambda_w^{n_w} \pi(\underset{\sim}{\lambda})$$

- From the first line of the complete probability model

$$p(z_{i1}, \ldots, z_{iW} | \textit{rest}) \propto \prod_w (\lambda_{iw}^*)^{z_{iw}}$$

where

$$\lambda_{iw}^* = \lambda_w \prod_j p_{wj}^{x_{ij}} (1 - p_{wj})^{1 - x_{ij}}$$

B.5.2 Latent Class Models: Haertel and Wiley

Haertel (1989) and Haertel and Wiley (1995) discuss a latent class model in which the latent classes are characterized by skill variables

$$\alpha_{wk} = \begin{cases} 1, & \text{if students in } C_w \text{ possess skill } k \\ 0, & \text{if not} \end{cases}$$

$k = 1, \ldots, K$; and for each task j we summarize the α's by

$$\xi_{wj} = \begin{cases} 1, & \text{if } \alpha_{wk} \geq q_{jk} \text{ for all } k = 1, \ldots, K \\ 0, & \text{if not} \end{cases}$$

i.e. $\xi_{wj} = 1$ indicates that all skills needed for problem j are present for students from latent class w. In the latent class model we set

$$p_{wj} = (1 - s_j)^{\xi_{wj}} g_j^{1-\xi_{wj}}$$

where s_j and g_j are per-task slip and guessing probabilities.

To estimate the Haertel-Wiley model we replace the complete conditionals for p_{wj} in the MCMC schematic for LC with

$$p(g_j|\textit{rest}) \propto \prod_{w:\ \xi_{wj}=0} \left\{ g_j^{c_{wj}} (1 - g_j)^{n_w - c_{wj}} \right\} \pi_g(g_j)$$

and

$$p(s_j|\textit{rest}) \propto \prod_{w:\ \xi_{wj}=1} \left\{ (1 - s_j)^{c_{wj}} s_j^{n_w - c_{wj}} \right\} \pi_s(s_j)$$

where $\pi_s(s_j)$ and $\pi_g(g_j)$ are priors as usual.

- Introducing "skill structure" *re-coupled* factors of the LC likelihood.
- $C_w = (\alpha_{w1}, \ldots, \alpha_{wK})$ are fixed in advance, so the useful diagnostic parameter is $P[\text{student } i \text{ in class } C_w|\mathcal{X}]$ λ^*_{iw},
  from the complete conditionals for z_{iw} (see above).

B.6 Corbett/Anderson/O'Brien

To show how to apply what we have learned to a model that

- Is motivated entirely from a cognitive diagnosis point of view;
- Allows us to learn about individual skills that students may or may not have, as opposed to latent class "skill ensembles"

we consider a *simplified* version of the "knowledge tracing" assessment model in the LISP tutor described by Corbett, Anderson, and O'Brien (1995).

The simplifications are

- No hidden-Markov learning model; just a one-shot assessment;
- Behavior observed at the level of task, not skill, performance;
- Some probabilistic structures replaced with equivalent deterministic ones.

The basic model parameters are exactly the same as those in the Haertel-Wiley model:

$$\begin{array}{rcll} X_{ij} & = & 1 \text{ or } 0 & \text{for performance of task } j \text{ by student } i \\ q_{jk} & = & 1 \text{ or } 0 & \text{for dependence of task } j \text{ on skill } k \\ \alpha_{ik} & = & 1 \text{ or } 0 & \text{for possession of skill } k \text{ by student } i \\ \xi_{ij} & = & \prod_{k:\ q_{jk}=1} \alpha_{ik} & = 1, \text{ if student } i \text{ has skills for task } j \\ s_j & = & P[X_{ij} = 0 | \xi_{ij} = 1], & \text{per-task slip parameter} \\ g_j & = & P[X_{ij} = 1 | \xi_{ij} = 0], & \text{per-task guessing parameter} \end{array}$$

along with prior distributions $s_j \sim \pi_s(s_j)$, $g_j \sim \pi_g(g_j)$, $\alpha_{ik} \sim \pi_k^{\alpha_{ik}}(1-\pi_k)^{1-\alpha_{ik}}$, and perhaps $\pi_k \sim \pi(\pi_k)$.

The goal is to infer values *for each* α_{ik}, or more accurately, to estimate $P[\alpha_{ik} = 1 | \textit{the data}]$.

Just as in the Haertel-Wiley model

$$P[X_{ij} = 1 | \underset{\sim}{\xi}, \underset{\sim}{\mathbf{s}}, \underset{\sim}{\mathbf{g}}] = (1-s_j)^{\xi_{ij}} g_j^{1-\xi_{ij}}$$

and so

$$\begin{aligned} & P[\mathcal{X} | \underset{\sim}{\xi}, \underset{\sim}{\mathbf{s}}, \underset{\sim}{\mathbf{g}}] \\ & = \prod_i \prod_j \left[(1-s_j)^{\xi_{ij}} g_j^{1-\xi_{ij}}\right]^{x_{ij}} \left[1 - (1-s_j)^{\xi_{ij}} g_j^{1-\xi_{ij}}\right]^{1-x_{ij}} \\ & = \prod_i \prod_j \left[(1-s_j)^{x_{ij}} s_j^{1-x_{ij}}\right]^{\xi_{ij}} \left[g_j^{x_{ij}} (1-g_j)^{1-x_{ij}}\right]^{1-\xi_{ij}} \end{aligned}$$

For a schematic MCMC algorithm, the complete conditionals for the slip and guessing parameters look very much like those in the Haertel-Wiley model:

- For each s_j we obtain

$$p(s_j | \textit{rest}) \propto (1-s_j)^{\sum_i x_{ij}\xi_{ij}} s_j^{\sum_i (1-x_{ij})\xi_{ij}} \pi_s(s_j),$$

- For each g_j we obtain

$$p(g_j|\textit{rest}) \propto g_j^{\sum_i x_{ij}(1-\xi_{ij})}(1-g_j)^{\sum_i (1-x_{ij})(1-\xi_{ij})}\pi_g(g_j),$$

- If we *were* interested in estimating ξ_{ij} directly we'd also see

$$p(\xi_{ij}|\textit{rest}) \propto \left[(1-s_j)^{x_{ij}} s_j^{1-x_{ij}}\right]^{\xi_{ij}} \left[g_j^{x_{ij}}(1-g_j)^{1-x_{ij}}\right]^{1-\xi_{ij}} \pi(\xi_{ij})$$

Now we would like to replace the complete conditonals for ξ_{ij} with complete conditionals for α_{ik}, where

$$\xi_{ij} = \prod_{k:\ q_{jk}=1} \alpha_{ik}$$

Let $\xi_{ij}^{(-k)} = \prod_{\ell \neq k: q_{j\ell}=1} \alpha_{i\ell}$, which indicates presence of all skills needed for task j, <u>*except for skill k*</u>. Then the complete conditional distributions for α_{ik} are of the form:

$$\begin{aligned} &p(\alpha_{ik}|\textit{rest}) \\ &\quad\propto \prod_{j:\ q_{jk=1}} \left[(1-s_j)^{x_{ij}} s_j^{1-x_{ij}}\right]^{\alpha_{ik}\xi_{ij}^{(-k)}} \left[g_j^{x_{ij}}(1-g_j)^{1-x_{ij}}\right]^{1-\alpha_{ik}\xi_{ij}^{(-k)}} \\ &\qquad\qquad \times\ \pi_k^{\alpha_{ik}}(1-\pi_k)^{1-\alpha_{ik}} \end{aligned}$$

Comparing with the complete conditionals for ξ_{ij} we can see that again there is a kind of "recoupling" of the likelihood factors.

Note that

- When $\xi_{ij}^{(-k)} = 1$, the suggested model for α_{ij} is some sort of Bernoulli, which makes sense.
- When there are no tasks such that both $q_{jk} = 1$ and $\xi_{ij}^{(-k)} = 1$, then α_{ik} is drawn from the prior distribution $\pi_k^{\alpha_{ik}}(1-\pi_k)^{1-\alpha_{ik}}$: no learning from data occurs.
- This is really a version of the credit/blame problem: we can't infer whether α_{ik} was learned, if we are hypothesizing that another needed skill is still unlearned.

Finally if we want to estimate the skill base rates π_k we may include a fourth set of complete conditionals

$$p(\pi_k|\textit{rest}) \propto \pi_k^{c_k}(1-\pi_k)^{N-c_k}\pi(\pi_k)$$

where $c_k = \sum_i \alpha_{ik}$ is the number of students who are presently estimated to have skill k.

B.7 Closing Thoughts

- *Decoupling*. As I've informally used it here, "decoupling" involves two features:
  - The likelihood for the data is in product form

  - Each parameter affects a relatively small subset of likelihood factors.

  Decoupling facilitiates

  - Writing complete conditional distributions for implementing MCMC algorithms;
  - Reducing the dimensionality of maximization and integration problems in marginal/conditional/joint maximum-likelihood estimation.

  Localized dependence / sparse parametrization (Hrycek, 1990); data augmentation (Tanner, 1996); variational bounds (Jaakkola and Jordan, 1999).

- *Validity, Reliability, Credit/Blame*. DiBello, Stout and Rousos (1995) discuss some validity and reliability considerations in constructing cognitively diagnostic assessment models.

  - Validity: Incomplete Q matrix; multiple strategies.
    $*$ Assessing the completeness of Q may require modeling learning as well as inventorying skills that are in place.
    $*$ LC models provide a general approach to multiple strategy *when the strategy is constant across problems within person.* If the strategy can change from problem to problem I think data augmentation will produce decoupled models, but unless you ask the right questions there will be very little data on which to base strategy estimates.
  - Reliability: positivity (slips/guessing) for specific skills or for the task as a whole.
    $*$ s_j and g_j begin to get at task positivity at least;
    $*$ may be diagnostic of certain defects in Q.
  - Credit/Blame problems that arose in the simplified Corbett et al. model suggest that either one has to design the assessment so that individual skills are observed, or at least design the tasks, task scoring, and skill set, so that there are no "effectively hidden skills".

- Complexity. Even in this abbreviated survey there are many levels of complexity; and even one computational method varies greatly in its complexity in applications to models.