# Graphical Models and Computerized Adaptive Testing

CSE Technical Report 434

Robert J. Mislevy and Russell G. Almond
CRESST/Educational Testing Service

July 1997

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90024-6511
(310) 206-1532

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B60002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education and by the Test Of English as a Foreign Language (TOEFL).

The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education.

**Acknowledgments**

We thank Isaac Bejar, Dan Eignor, Drew Gitomer, Ed Herskovitz, David Madigan, Linda Steinberg, and Martha Stocking for comments and discussions on various topics discussed here, and the TOEFL Research Committee for their encouragement and support.

Several experts in communicative competence and language proficiency assessment contributed to this project by generously sharing with us their advice, wisdom, and reference materials: Lyle Bachman, Gary Buck, Frances Butler, Carol Chapelle, Dan Douglas, Joan Jamieson, Irwin Kirsch, Mary Schedl, and Carol Taylor. They are not responsible for our errors or misunderstandings, and their assistance should not be taken as agreement with our views or conclusions.

# GRAPHICAL MODELS AND COMPUTERIZED ADAPTIVE TESTING

**Robert J. Mislevy and Russell G. Almond**
**CRESST/Educational Testing Service**

## Abstract

This paper synthesizes ideas from the fields of graphical modeling and educational testing, particularly item response theory (IRT) applied to computerized adaptive testing (CAT). Graphical modeling can offer IRT a language for describing multifaceted skills and knowledge, and disentangling evidence from complex performances. IRT-CAT can offer graphical modelers several ways of treating sources of variability other than including more variables in the model. In particular, variables can enter into the modeling process at several levels: (a) in validity studies (but not in the ordinarily used model); (b) in task construction (in particular, in defining link parameters); (c) in test or model assembly (blocking and randomization constraints in selecting tasks or other model pieces); (d) in response characterization (i.e., as part of task models which characterize a response); or (e) in the main (student) model. The Graduate Record Examination (GRE) is used to illustrate ideas in the context of IRT-CAT, and extensions are discussed in the context of language proficiency testing.

## 1.0 Introduction

Computerized adaptive testing (CAT; Wainer et al., 1990) is one of the most significant practical advances in educational testing in the past two decades. Using the information in their unfolding patterns of responses to adaptively select items for examinees, CAT can improve motivation, cut testing time, and require fewer items per examinee, all without sacrificing the accuracy of measurement. The inferential underpinning of modern CAT is item response theory (IRT; Hambleton, 1989). Successful large-scale applications of IRT-CAT include the Graduate Record Examination (GRE) and the National Council Licensure Examination (NCLEX) for assessing nurses.

As useful as IRT-CAT has been, two constraints have blocked its extension to wider varieties of applications. These constraints are the limited scope of tasks that can be used without seriously violating IRT's conditional independence assumptions, and IRT's limited capabilities to deal jointly with multiple, interacting aspects of knowledge or skill. Graphical models (GMs; Almond, 1995, Lauritzen, 1996; they are often called Bayesian Inference Networks, or BINs, when used predictively; Pearl, 1988) provide a language for describing complex multivariate dependencies. A graphical modeling perspective extends the IRT-CAT inferential framework to accommodate richer tasks and more complex student models.

Despite the simplistic nature and strong independence assumptions of the IRT-CAT model, its users have developed sophisticated techniques to ensure its success in practical applications. Many variables seemingly ignored by the IRT model actually enter into the task creation and test assembly processes—often informally. These techniques could be adapted to other applications of graphical modeling as well, as graphical modelers move away from the idea of an all-encompassing model and toward collections of model fragments, which can be assembled on the fly to meet specific task demands (knowledge-based model construction; Breese, Goldman, & Wellman, 1994).

This paper synthesizes a number of ideas from graphical modeling and educational testing. To this end, Section 2 reviews the basic ideas of IRT and CAT, and Section 3 casts them as a special case of probability-based inference with graphical models. We then see that the simplicity of IRT as a GM is deceiving. Section 4 describes how many variables are handled informally or implicitly play crucial roles in practical applications of IRT-CAT, even though they do not appear in the IRT model. We sketch more complex GMs to reveal the significance of some of these hidden extra-measurement considerations. Section 5 outlines graphical-model-based assessment, adaptive if desired, with models that explicitly incorporate such considerations in order to handle more complex tasks or student models. Section 6 sketches two ways this approach might be employed in language proficiency assessments that employ complex, integrative tasks. (For an illustration of their use in a fielded application, see Mislevy & Gitomer, 1996, and Steinberg & Gitomer, 1996, on HYDRIVE, an intelligent tutoring system for learning to troubleshoot aircraft hydraulics

systems.) Section 7 lists some technical issues that must be explored in developing graphical-model-based assessment.

## 2.0 Item Response Theory and Computerized Adaptive Testing

An IRT model expresses an examinee's propensity to make correct responses or receive high ratings on a collection of test items in terms of an unobservable proficiency variable $\theta$. The responses are posited to be independent, conditional on $\theta$ and parameters that express characteristics of the items such as their difficulty or sensitivity to proficiency. A simple example is the Rasch model for $n$ dichotomous test items:

$$\mathrm{P}(x_1,\ldots,x_n|\theta,\beta_1,\ldots,\beta_n) = \prod_{j=1}^{n} \mathrm{P}(x_j|\theta,\beta_j), \tag{1}$$

where $x_j$ is the response to Item $j$ (1 for right, 0 for wrong), $\beta_j$ is the "difficulty parameter" of Item $j$, and $\mathrm{P}(x_j|\theta,\beta_j) = \exp\left[x_j(\theta-\beta_j)\right] / \left[1+\exp(\theta-\beta_j)\right]$. For selecting items and scoring examinees in typical applications, point estimates of the item parameters $(\beta_1,\ldots,\beta_n)$, or $\mathbf{B}$ for short, are obtained from large samples of examinee responses and treated as known. Section 4.2 below will discuss modeling alternative sources of information, and remaining uncertainty, about $\mathbf{B}$.

Once a response vector $\mathbf{x} = (x_1,\ldots,x_n)$ is observed, (1) is interpreted as a likelihood function for $\theta$, say $L(\theta|\mathbf{x},\mathbf{B})$. The MLE $\hat{\theta}$ maximizes $L(\theta|\mathbf{x},\mathbf{B})$; its asymptotic variance can be approximated by the reciprocal of the Fisher information function, or the expectation of second derivative of $-L(\theta|\mathbf{x},\mathbf{B})$, evaluated at $\hat{\theta}$. Bayesian inference is based on the posterior distribution $p(\theta|\mathbf{x},\mathbf{B}) \propto L(\theta|\mathbf{x},\mathbf{B})p(\theta)$, which can be summarized in terms of the posterior mean $\overline{\theta}$ and the posterior variance $Var(\theta|\mathbf{x},\mathbf{B})$.

Fixed test forms have differing accuracy for different values of $\theta$, with greater precision when $\theta$ lies in the neighborhood of the items' difficulties. CAT provides the opportunity to adjust the level of difficulty to each examinee. Testing proceeds sequentially, with each successive item $k$+1 selected to be informative about the examinee's $\theta$ in light of the responses to the first $k$ items, or $\mathbf{x}^{(k)}$ (Wainer et al., 1990, chapter 5). One common approach evaluates $\hat{\theta}$ after

each response, then selects the next item from the pool that provides a large value of Fisher information in the neighborhood of $\hat{\theta}$. A Bayesian approach determines the next item as the one that minimizes expected posterior variance, or $E_{x_j}\left[Var\left(\theta|\mathbf{x}^{(k)},x_j,\mathbf{B}^{(k)},\beta_j\right)\middle|\mathbf{x}^{(k)},\mathbf{B}^{(k)}\right]$ (Owen, 1975). Additional constraints on item selection, such as item content and format, are addressed below in Section 4.3. Testing ends when a desired measurement accuracy has been attained or a predetermined number of items has been presented.

## 3.0 IRT Computerized Adaptive Testing as a Graphical Model

Probability-based inference in complex networks of interdependent variables is an active topic in statistical research, spurred by such diverse applications as forecasting, pedigree analysis, troubleshooting, and medical diagnosis. The structure of the relationships among the variables can be depicted in an acyclic directed graph (commonly called a DAG), in which nodes represent variables and edges represent conditional dependence relationships. Corresponding to the DAG is a recursive representation of the joint distribution of the variables of interest, generically denoted $\{Z_1,\ldots,Z_m\}$:

$$p(Z_1,\ldots,Z_m)=\prod_{j=1}^{m}p\left(Z_j\middle|\{\text{"parents" of } Z_j\}\right), \tag{2}$$

where the $\{\text{"parents" of } Z_j\}$ is the subset of $\{Z_{j-1},\ldots,Z_1\}$ upon which $Z_j$ is directly dependent. In educational applications, for example, we posit unobservable variables that characterize aspects of students' knowledge and skill as parents of observable variables that characterize what they say and do in assessment situations. Spiegelhalter, David, Lauritzen, & Cowell (1993) review recent statistical developments in graphical modeling.

Figure 1 shows the DAG that corresponds to IRT. The generic $Z$ variables specialize to $\theta$ and the item responses $\{X_1,\ldots,X_n\}$. The first panel suppresses the dependence on item parameters, while the second makes the dependence explicit by indicating that the conditional probability distribution of each $X_j$ given $\theta$ is a function of $\beta_j$. Such a structure, which posits conditional independence of item responses given a single unobserved variable, is often

called a "naive Bayes" model since it rarely captures the subtle relationships found in real-world

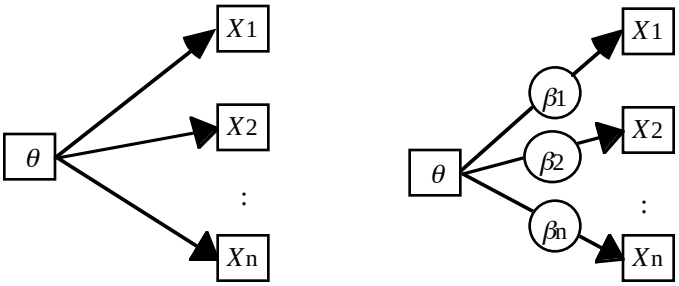


*Figure 1.* DAGs for an IRT model. Item parameters that determine conditional distributions of Xs given $\theta$ are implicit in the left panel and explicit in the right panel.

problems (Spiegelhalter & Knill-Jones, 1984). This depreciative term is undeserved in thoughtful implementations of IRT-CAT, however, because many variables that do not appear in the simple model have been handled behind the scenes, expressly to ensure that its simple structure will suffice for the task at hand.

One way to describe IRT-CAT from the perspective of graphical models is through the DAG with $\theta$ as the single parent of all items in the test pool, as in Figure 1. At the beginning of testing, the marginal distribution of the $\theta$ node is $p(\theta)$. Each item is checked to find one that minimizes expected posterior variance; it is administered, and the process repeats after the response, now starting from $p(\theta|x^{(1)})$. The process continues with each successive $p(\theta|\mathbf{x}^{(k)})$ until testing is terminated. At each step, the observed value of the administered variable is fixed, the distribution of $\theta$ is updated, and expectations for items as-yet-unadministered are revised for calculating the expected posterior variance of $\theta$ if each of the items were presented next.

A second way to describe IRT-CAT is statistically equivalent, but highlights the modularity of reasoning that can be achieved with graphical models. Figure 2 depicts the situation in terms of graphical model fragments: the student-model variable $\theta$ and a library of nodes corresponding to test items, any of which can be "docked" with the $\theta$ node to produce a dyadic DAG

as shown in the right-hand panel of the figure. This small DAG is temporarily assembled to absorb evidence about $\theta$ from the response to a given Item $j$. It is disassembled after the response is observed and the distribution of $\theta$ updated accordingly. The new status of knowledge about $\theta$ either guides a search of the item library for the next item to

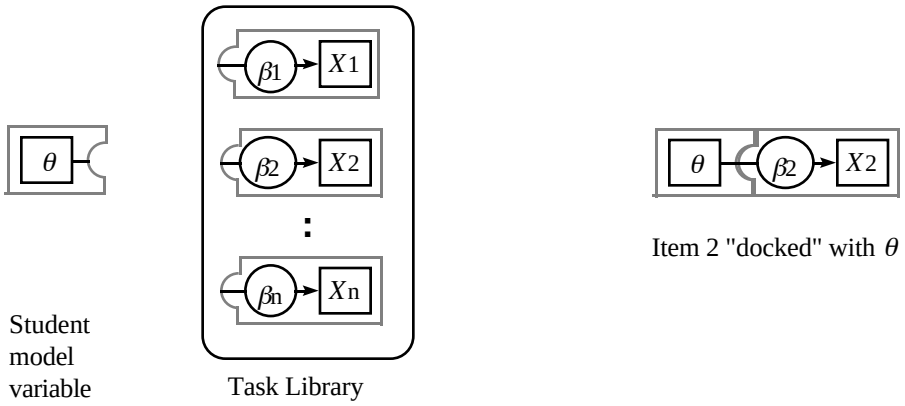


*Figure 2.* CAT as knowledge-based model construction. Left panel shows $\theta$ node and task-node library. Right panel shows Item 2 "docked" with $\theta$ to create a dyadic DAG.

administer or provides the grounds to terminate testing. This process is an example of knowledge-based model construction (Breese et al., 1994).

## 4.0 Roles of Variables in IRT-CAT

A first glance at the IRT models used in current tests such as the GRE's Verbal, Quantitative, and Analytic subtests or the Test of English as a Foreign Language (TOEFL) measures of Reading, Listening, and Structure gives the misleading impression that everything that is happening can be understood in terms of simple, one-variable student models—the overall proficiencies in each scoring area—and corresponding task pools. But many more variables are being managed behind the scenes, some to effectively define the variable being measured, others to ensure that the simple analytic model will adequately characterize the information being gathered.

Every real-world problem has its own unique mix of features and demands, and every person has a unique approach to its demands. This is true in particular of assessment tasks, and accordingly, examinees will vary in

their degree of success with each of them. Educational and psychological measurement, as it has evolved over the past century, defines domains of tasks so that differences among examinees with respect to some features tend to accumulate over tasks, while differences with respect to other features don't tend to accumulate (Green, 1978). The variance that accumulates becomes "what the test measures," or the operationally defined "construct." Other sources of variance constitute uncertainty about an examinee's standing on that construct.

What practices have evolved to guide testing practice under this perspective? This section discusses roles that variables serve to this end in IRT-CAT.

1. Variables can limit the scope of the assessment, and never appear in the analytic model.

2. Variables can describe task features, for constructing tasks and modeling item parameters.

3. Variables can control test assembly.

4. Variables can characterize responses (observables).

5. Variables can characterize aspects of proficiency (collectively, the student model).

A given variable can play different roles in different tests, according to the purposes and operational definitions of those tests. Only variables playing the last role in the list appear explicitly in the measurement model—in the case of IRT-CAT, $\theta$. $\theta$ is usefully thought of as a summary of evidence about a construct brought about through choices about, and manipulation of, many other "hidden" variables through the first four roles listed above.

### 4.1 Variables That Limit the Scope of the Assessment

This section shows how two kinds of studies usually thought of as validity analyses help ensure that the simple structure of IRT is adequate. In both cases, variables that might generate interactions among item responses beyond those accounted for by an overall proficiency variable are the focus of the study, and actions are taken so that these variables need not be included in the analytic model. Results in the first case lead one to constrain testing

contexts and methods, so that the operationally defined $\theta$ effectively conditions on specified values of these variables. Results in the second case can lead one to eliminate items that would engender strong interactions with unmodeled student characteristics, so that one can effectively marginalize over those characteristics.

**Delimiting the domain and the testing methods.** Myriad aspects of examinees' skills, knowledge, and experience affect their performance in any learning domain, not all of which can be, nor should be, encompassed in any particular test. We must consider which aspects of the universe of potential assessment tasks are salient to the job at hand and determine which of them to address in the test and which to exclude. In a test of academic language proficiency, for example, do we want to include scenarios that span all of college life in a test of English proficiency, from doing the laundry to interacting with campus police, or shall we limit attention to academic and classroom interactions? Should we assess listening skills with closed-form items based on taped segments, or with tasks that combine listening with speaking in a conversation with a human examiner? The way we elicit performance in language tests has a significant effect on performance; some examinees are relatively better at one kind of task than another, perform better in some settings than others, or are more familiar with some contexts than others. There will thus tend to be stronger associations among some tasks than others related to testing contexts and methods—interactions that invalidate the structure of the DAG in Figure 1. If we want to use IRT models, studying sources of variability in tasks (e.g., Bachman, Lynch, & Mason, 1995) helps us determine when we can ignore such interactions, and when they are so large we should consider scaling within more homogeneous subsets of tasks.

**Differential item functioning (DIF).** DIF occurs when, for reasons unrelated to the skills and knowledge of interest, certain task content or format features tend to be relatively harder for members of identifiable subpopulations, as defined for example by gender or ethnic background. Reading comprehension questions about baseball might be more difficult for girls than boys, who would perform similarly on items with the same language and use characteristics, but about other topics. The DAG in Figure 3 depicts this unwelcome situation. DIF

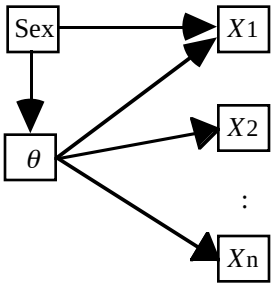


*Figure 3.* A DAG illustrating Differential Item Functioning (DIF). Response probabilities of Items 2-n are conditionally independent of sex given $\theta$. Response probabilities for Item 1 are dependent on sex as well as $\theta$.

analyses explore pretest data for its presence. Some potential causes of DIF can be avoided by defining variables that identify problematic features of tasks, and excluding any tasks that have these features from the domain. (In contrast, an instructional application might purposely seek out items for which personal interest is very high for certain students, in order to better motivate them to engage the underlying concepts.)

### 4.2 Variables That Describe Task Features

Individual tasks in a test can be described in terms of many variables. They concern such things as format, content, modality, situation, purpose, vocabulary load, grammatical structure, mathematical knowledge required, cognitive processing demands, and so on. Some of these variables appear formally in test specifications, but test developers employ far more when they create the tasks. Without formally naming or coding this information in terms of variables, writers of tasks draw upon such sources as past results with similar items, experience with how students learn the concepts, awareness of common misconceptions, and cognitive research about learning and problem solving in the domain. Studies have shown that these kinds of variables can be strong predictors of item difficulty (see, for example, Freedle & Kostin, 1993, on TOEFL listening comprehension tasks, and Chalifour & Powers, 1989, on GRE analytical reasoning tasks).

One way to use this collateral information about tasks is to supplement, perhaps supplant, data from pretest examinee samples as the source of information about the IRT item parameters **B** (Mislevy, Sheehan, &

Wingersky, 1993). In effect, one creates a second-order DAG for modeling item parameters (Figure 4).

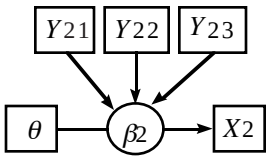


*Figure 4.* A portion of a two-level DAG, which posits a model for the item parameter $\beta$ that in turn gives the conditional probabilities of the response to Item 2 given q. $Y_{21}$-$Y_{23}$ are coded features of Item 2.

A second way to use the normally hidden variables that characterize test items is to erect a more principled framework for item construction. Such variables would be the basis of "item schemas" or "item shells," for developing families of tasks with characteristics and properties that are both fairly well understood and demonstrably grounded in a theoretical framework of the knowledge and skills the test is meant to elicit. Features of schemas and features of the elements that fill in schemas could then be used to model IRT parameters, as discussed above. The intimate connection between task construction and difficulty from a cognitive point of view is illustrated in Bejar (1990). See Hively, Patterson, and Page (1968) for a proposal along these lines before the days of IRT, and Embretson (1993) for a more recent investigation using contemporary cognitive and measurement theory.

A third way to use variables that characterize task requirements is to link values of student-model variables to expected observable behaviors. With the Rasch model, for example, knowing $\beta_j$ allows us to calculate the probability of a correct response from a student with any given $\theta$. Conversely, we can give meaning to a value of $\theta$ by describing the kinds of items a student at that level is likely to succeed with, and those he is not. To the extent that item features account for $\beta$s, then, we can describe the student's proficiency in terms of task characteristics and/or cognitively relevant skills (see Sheehan & Mislevy, 1990, for an example with document literacy tasks, and McNamara, 1996, chapter 7, for an example concerning Chinese language reading proficiency).

### 4.3 Variables That Control Test Assembly

Once a domain of items has been determined, test specifications constrain the mix of items that constitute a given examinee's test. We observe neither the whole of the task domain nor an uncontrolled sample, but a composite carefully assembled under prespecified rules for "blocking" and "overlap."

Blocking constraints ensure that even though different examinees are administered different items, generally of different difficulties in a CAT, they nevertheless get similar mixes of content, format, modalities, skill demands, and so on. Stocking and Swanson (1993) list 41 constraints used in a prototype for the GRE CAT, including, for example, the constraint that one or two aesthetic/philosophical topics be included in the Antonym subsection. Since it is not generally possible to satisfy all constraints simultaneously, these authors employed integer programming methods to optimize item selection, with item-variable blocking constraints in addition to IRT-based information-maximizing constraints.

Overlap constraints concern the innumerable idiosyncratic features of items that cannot be exhaustively coded and catalogued. Sets of items are specified that must not appear in the same test because they share incidental features, give away answers to each other, or test the same concept. Overlap constraints evolved through substantive rather than statistical lines, from the intuition that overlapping items reduce information about examinees. The graphical modeling formalism allows us to explicate why, how, and how much is lost. Each item is acceptable in its own right, but their joint appearance would introduce an unacceptably strong conditional dependence—"double counting" evidence (Schum, 1994, p. 129) under the simple conditional independence model.

Figure 5 illustrates the impact of test assembly constraints with a simple example. The item pool has just four items; Items 1 and 2 both use the unfamiliar word "ubiquitous," and Items 3 and 4 both concern right triangles. Overlap constraints would say a given examinee's test should not contain both Items 1 and 2, and not both Items 3 and 4. A blocking constraint would say that one item from each pair should appear in each examinee's test. The first and second panels in Figure 5 are alternative DAGs for the entire pool, one showing conditional dependencies among overlap sets and the other

introducing additional student-model variables. The third panel is the standard IRT-CAT DAG with overlap and

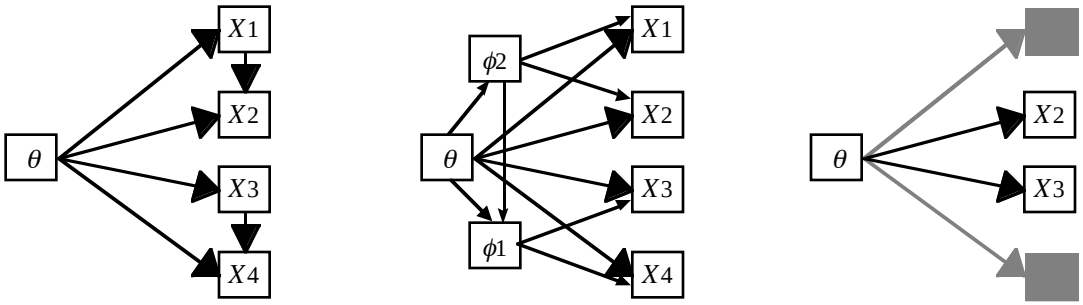


*Figure 5.* Three DAGs related to overlap and blocking constraints. The first panel shows conditional dependencies among item sets. The second shows conditional independence achieved by adding student-model variables. The third shows conditional independence achieved within the IRT model by constraining what can be observed.

blocking constraints in place—its simplicity is appropriate only because the inflow of evidence has been restricted so as to avoid some particularly egregious violations of its strong conditional independence structure.

Many other variables could be defined to characterize test items according to features not controlled by blocking or overlap constraints. These include the item-level variables discussed in Section 4.2 that can be used to model item parameters, as well as the many incidental and idiosyncratic features that make each item unique. These variables are dealt with by randomization; the particular values they take in any given examinee's test are a random sample from the pool, subject to blocking, overlap, and measurement constraints. The GRE Verbal CAT, for example, may require that each examinee receive one passage on a topic in science and another in literature. There are many topics within both areas, and one may be selected from each area in accordance with other constraints but ignoring the specific identification of topics within areas. Whether an examinee happens to be familiar or unfamiliar with a given topic undeniably affects her performance, but this interaction is not modeled; having randomized, the examiner leans on large sample theory to average over these effects.

### 4.4 Variables That Characterize Responses (Observables)

Characterizing student responses is straightforward with multiple-choice items in IRT-CAT: Did the student indicate the option prespecified to be

correct, or a different one? Open-ended responses can also be analyzed with dichotomous IRT models, but more judgment is required to distill "correctness" from unique performances. In these latter cases, variables can be defined to describe qualities of the products or performances students produce, and rules can be devised for mapping values of these variables into the correct/incorrect dichotomy.

More generally, salient characteristics of examinee responses can be coded in terms of fully or partially ordered rating categories. For example, Bachman and Palmer (1996, p. 214) offer a variable for coding "knowledge of syntax" as displayed in specific tasks by means of a five-point rating scale. The fourth point, "evidence of extensive knowledge of syntax," is marked by a large range with few limitations, and good accuracy with few errors. IRT models have been extended beyond dichotomous data to deal with these ordered response categories (see Thissen & Steinberg, 1986, for a taxonomy of models). In this case, $X_j$ is multinomial, and item parameters give the probabilities of response in the possible categories conditional on $\theta$. Dodd, De Ayala, and Koch (1995) describe IRT-CAT with such models. As with dichotomous models, the value of $X_j$ may either be immediate because of restrictions on possible response behavior, or it may require a further step of evaluation in terms of abstracted properties of less constrained response behaviors. When nontrivial differences may occur among qualified observers, IRT models that include effects for raters and diagnostic information for monitoring their work can be employed (e.g., Linacre, 1989; see McNamara, 1996, on the use of these models in language proficiency assessment).

### 4.5 Variables That Characterize Aspects of Proficiency (the Student Model)

Student-model variables integrate information across distinct pieces of evidence to support inference about examinees' skills and knowledge at a higher level of abstraction than the particulars of any of the specific tasks—a level consonant for instruction, documentation, or decision making, as the application demands. The nature of student-model variables should be driven by the purpose of the test, but also be consistent with empirical response

patterns and theories of performance in the domain. As further discussed in the following sections, it is neither possible nor desirable to include in the model variables for all conceivable aspects of proficiency. The choice is determined by utilitarian purposes, such as distinctions that will be important for reporting or decision making, as opposed to complete psychological and sociological explication of responses.

For example, the current TOEFL has three student-model variables—listening, reading, and grammatical structure, or L, R, and S—and each is evidenced by discrete tasks of its type only, with disjoint item domains and associated domain proficiency variables $\theta_L$, $\theta_R$, and $\theta_S$, each as depicted in Figure 1. These variables are used for infrequent but consequential decisions such as admitting non-native English speakers into undergraduate and graduate academic programs. In contrast, an intelligent tutoring system (ITS) must define student-model variables at a finer grain-size in order to provide instruction frequently and specifically. The guiding principle for ITSs is that student models should be specified at the level at which instructional decisions are made (Ohlsson, 1987).

Standard IRT-CAT is based on univariate student models. Multivariate student models become important when observations contain information about more than one aspect of proficiency, for which it is desirable to accumulate evidence. Segall (1996) describes CAT with multivariate normal student-model variables and logit-linear models linking their values to the probability of item responses. Sections 5 and 6 discuss multidimensional student models further, with some examples motivated by the TOEFL program's TOEFL 2000 project.

## 5.0 Graphical-Model-Based Computerized Adaptive Testing (GM-CAT)

Experts differ from novices, not merely by commanding more facts and concepts, but also by forging and exploiting richer interconnections among them (e.g., Chi, Feltovich, & Glaser, 1981). Direct assessment of increasing expertise, therefore, requires (a) complex tasks, in order to elicit evidence that draws upon multiple and interrelated aspects of skill and knowledge, and (b)

multivariate student models, in order to capture, integrate, and accumulate the import of students' performances across such tasks. The fact that standard IRT is not up to the task does not require abandoning its underlying inferential principles, but rather extending them. We can build on the same ideas of defining unobservable variables to "explain" patterns of observable responses, and "some sources of variation accumulating and others not"—and of using probability-based inference to manage accumulating knowledge and remaining uncertainty about student proficiency as assessment proceeds. This section sketches out an approach in general terms, noting how it addresses issues discussed above in the context of IRT-CAT. The following section illustrates the ideas with two examples from language proficiency assessment. Mislevy and Gitomer (1996) and Steinberg and Gitomer (1996) describe a simplified application of the approach in a fielded system, the HYDRIVE intelligent tutoring system for troubleshooting aircraft hydraulics systems.

Figure 6 illustrates one possible implementation of a GM-CAT. It is presented here to provide a visual reference for the discussion of the mathematical properties. Section 6.2 presents the language-testing motivation

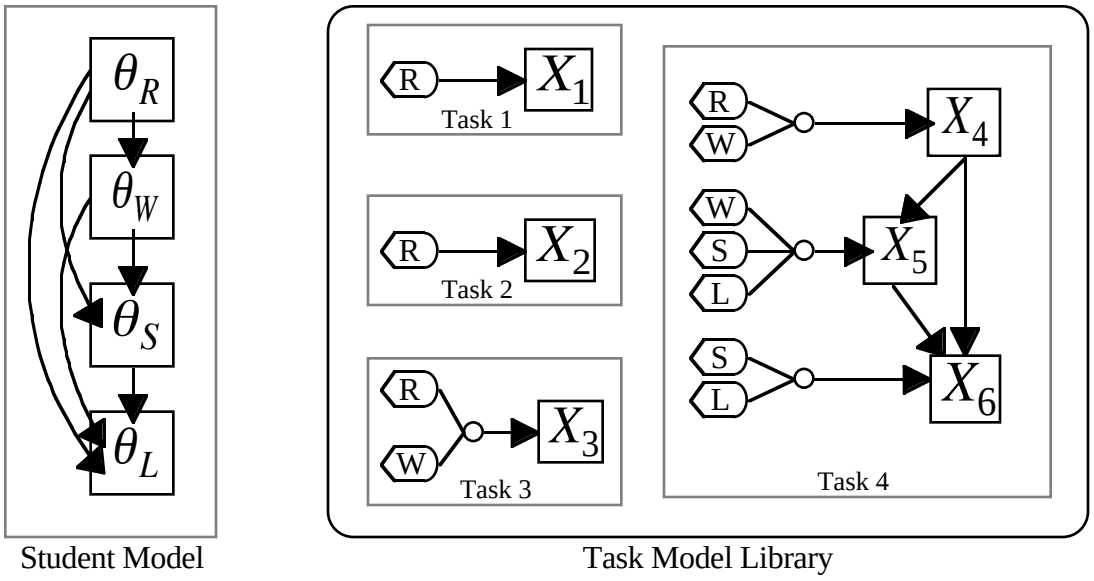


*Figure 6.* A "task-oriented" DAG. Information about examinee performance is accumulated in variables associated with the four traditional "skills." Conditional probabilities of task responses are modeled in terms of cognitively-relevant task

features (not depicted). Items 1 and 2 are conditionally independent, and each depends on only a single student-model variable. Item 3, a small "integrated" task, has two skill parents. Items 4, 5, and 6 are multiple aspects of response to a single complex task; each has multiple skills as parents, and conditional dependencies among items are further indicated to deal with context effects.

for this example, and Section 6.3 discusses an alternative approach. (It may be noted that some of the variables—in particular, $\theta_R$, $\theta_W$, $\theta_S$, and $\theta_L$—seem evocative of the concepts Reading, Writing, Speaking, and Listening. Whatever meaning was intended by placing those variables in the model, their operational meaning is an average over performance on tasks related to those modalities. Thus the true meaning of the variables in the model is controlled by variables that do not appear at all in Figure 6: that is, variables controlling the scope of the exam [Section 4.1] and the selection of tasks [Section 4.3].)

The model in the GM-CAT framework is spread among two sources. To the left is the student model, which is fixed across all administrations of the exams. To the right is a collection of task/evidence models, or DAG fragments, corresponding to a pool of tasks. A given examinee will see a subset of the tasks according to a *task selection algorithm*, which balances value of information considerations with content and overlap constraints. When an examinee is assigned a task, the evidence model associated with that task is attached to the student model (according to the pattern of stub variables in the evidence model). The evidence from the examinee's response to that task is then absorbed into the main student model, and the task/evidence model can be detached, leaving the updated student model ready for the next task. Thus, the GM-CAT framework is another application of knowledge-based model construction (Breese et al., 1994).

The nodes in the student model are unobservable variables related to examinee proficiency—a multivariate generalization of the role of IRT $\theta$. The student-model variables represent aspects of skill and knowledge and are included in the model either because they will be used to report students' performance, to accumulate supplementary patterns across task situations for diagnostic feedback, or to account for incidental dependencies across tasks. Their nature and number should be consistent with, but are not uniquely determined by, an understanding of performance in the domain. The final determination of the number and granularity of variables belonging in the

student model is governed by the requirements for reporting and diagnosis in the examination. Thus a pass/fail licensure exam will use a much coarser student model than an intelligent tutoring system.

The nodes in the task evidence models are observable variables that correspond to salient aspects of examinees' behaviors in specified task situations—a generalization of the IRT item responses. Generally, these will correspond to features of a task response. They could be as simple as "did the examinee give the correct response to a multiple-choice question" or as complex as dimensions of a multi-attribute rating produced by a human judge or by running a parser on a transcript of examinee actions in a simulator.

There are three kinds of associations among the student-model and observable nodes.

The first kind of association is the most important: Student-model variables are parents of observables. In this way, skills and knowledge "explain" patterns in observable behavior in the tasks at hand, and when responses are observed, belief about student-model variables is updated. The associations take the form of conditional probabilities of values of the observable variables, given the values of student-model variables—a generalization of IRT item parameters. When multiple aspects of skill and knowledge are posited as parents of a given observable, relationships such as conjunction, disjunction, and compensation may be proposed. Task designers indicate the structure of these associations (indicated by item stubs in Figure 6) and provide initial estimates of the conditional probabilities based on task-feature variables, response-feature variables, and expectations of the latter given the former at various levels of the student-model variables. These conditional probabilities may be further modeled as functions of task-characteristic variables, as a generalization of the IRT technique depicted in Figure 4.

A second kind of association is that among observables, over and above the associations induced by student-model variables. These occur when multiple aspects of a performance in the same task situation are captured as observables, and including them in the DAG is a way to model the effects of shared contexts, similarities in response methods, or incidental connections that overlap constraints would disallow in IRT-CAT. A task/evidence model

for a complex task would comprise multiple observables, perhaps with associations engendered by the commonalities induced by shared context, but probably with different student-model parents according to their particular demands. These associations are illustrated in Figure 6 by the arrows connecting observables $X_4$, $X_5$, and $X_6$.

A third kind of association is that among various student-model variables: that is, some student-model variables may appear as parents of other student-model variables in order to express such relationships as prerequisition, empirical correlation, or logical relationships such as conjunction and disjunction. These associations appear in Figure 6 as arrows connecting student-model variables to one another. In this way, direct evidence about one student-model variable can provide indirect evidence about another, thereby exploiting associations among skills or competences to improve the accuracy of reports.

Adaptive testing with a graphical model would use the current state of the student model as part of the item selection algorithm. Just as in the IRT-CAT, the GM-CAT selects tasks from a task pool to maximize some information metric. *Value of information* (Heckerman, Horvitz, & Middleton, 1993) and *weight of evidence* (Madigan & Almond, 1996) seem promising candidates. The GM-CAT attaches the task/evidence model to the student model and absorbs the evidence provided by the examinee's responses. The algorithm can then discard the task item, or maintain it in the model if it is needed to deal with dependence effects between tasks (i.e., overlap considerations addressed by modeling, as opposed to avoidance). The algorithm will still need to balance tasks' contexts, content, task types, and so on within examinees, since these specifications operationally define the student-model variables in the same sense that item pools and test assembly rules define $q$ in IRT.

The status of the student model is also used for reporting, or, in interactive applications, triggering feedback. If a single-number summary of performance is desired, one can project the current state of the student model onto a particular dimension such as expected performance on a market basket of typical tasks. Validity studies increase in importance, because validity internal to the model must now be monitored as well as relationships to variables outside the model.

## 6.0 Examples From Language Proficiency Assessment

This section illustrates the ideas of graphical-model-based assessment in the context of language proficiency testing. The TOEFL 2000 project and key language testing issues are introduced, then two approaches to modeling complex tasks are described.

### 6.1 Background

The current TOEFL described above is widely considered to be a discrete-point test built on the structuralist behaviorist model of language learning and testing. Both users and the language learning and testing communities have called for a new TOEFL test that more closely targets language use in the academic environment, as opposed to knowledge of vocabulary and surface linguistic features. The TOEFL 2000 project was thus initiated, with the goal of measuring communicative English-language competence that focuses on situations and tasks that reflect university life in North America. It is anticipated that the resulting assessment will (a) incorporate speaking and writing; (b) include more performance-based tasks; (c) include tasks that are integrated across modalities, such as writing based on listening to a conversation or speaking in response to a reading passage; and (d) provide reports that go beyond norm-referenced scores (Carol Taylor, personal communication, January 1997).

These aims reflect Hymes' (1972) "communicative competence" perspective. "[U]nlike the Chomskyan notion of linguistic competence, which is a property of the mind, communicative competence is a product of the psychological and social characteristics of situations on which language is used for communication" (Waters, 1996, p. 54). From this point of view, assessing communicative language proficiency requires both an analysis of the targeted language use situations and the kinds of knowledge that are needed to use language in those situations. McNamara (1996, chapter 3) provides an integrative review of recent models of communicative language proficiency, including Bachman's (1990) model comprising the components summarized in Figure 7.

TOEFL 2000 has made progress on several fronts. Integrative reviews, field surveys, and empirical research have addressed the issues of relevant

situations and language uses (e.g., Hudson, 1996; Waters, 1996), and linguistic, cognitive, and sociolinguistic features that influence language use task difficulty (e.g., Freedle & Kostin, 1993; Nissan, DeVicenzi, & Tang, 1996). The "Committee of Examiners Model" (Chapelle, Grabe, & Berns, in press) lays out considerations for task contexts, situations, and performances; it relates these task features to the processing required to negotiate them successfully; and it draws implications for task development and test validation. TOEFL test developers have created prototypes of integrated tasks that exhibit the integration of modalities and the context-embedding features that are called for (e.g., the "dinosaur task" mentioned below). And, as of this writing, a draft of a TOEFL 2000 test framework is circulating for comment and review. The framework takes steps to further specify the aspects of situations, materials, and uses of tasks that would constitute the assessment (Table 1) and begins to model relationships between these aspects and examinee performance. In sum, a number of relevant variables have been identified, which can be considered for various of the roles discussed above in Sections 4.1-4.4.

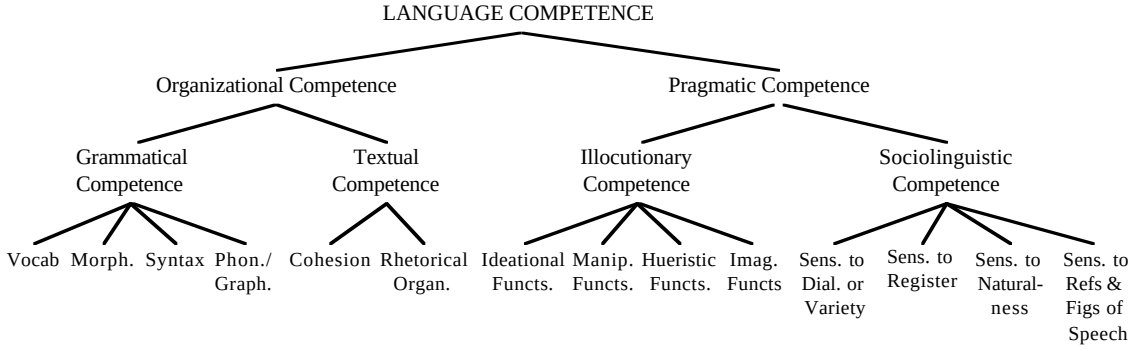


*Figure 7.* A schematic summary of Bachman's (1990) model. The components of "organizational competence," the foci of language tests with a structural perspective, are viewed as enabling skills that must be integrated with an understanding of situation and purpose for successful communication. The components of these latter capabilities are subsumed under "pragmatic competence."

Table 1

Further Breakdown of Aspects of Language Use Tasks

| Situation | Characteristics of input material | Types of questions |
|---|---|---|
| Setting | Grammatical features | Different types of questions |
| Participants and their roles | Pragmatic features | Type of information requested |
| Register | Discourse features | Type of match |
| Purpose | Text structure properties | Additional processing conditions |
| Content | Documents | Plausibility of distractors[a] |
| | Prose | |
| | Interactions | |

[a] In constructed responses and open-ended tasks, this term refers to the fineness of distinctions that must be made in order to negotiate the task successfully.

Less progress has been made in specifying a set of student-model variables (Section 4.5) and delineating evidentiary relationships between them and task performances. This charge has proved difficult for several reasons. There are vast numbers of plausible candidates for student-model variables. Richards (1983), for example, lists 33 "micro-skills" required for just for conversational listening and 18 for academic listening. Different authorities, writing from different theoretical perspectives or having different purposes in mind, offer proposals that are in some cases overlapping, in other cases orthogonal, and in still others, contradictory. It is generally acknowledged that skills and knowledge, however defined, always interact in use. Student-model variables cannot be decided upon in isolation, but are roles co-defined with all the other roles discussed above in light of the intended use of the assessment. Any descriptor of tasks, for example, can induce a student-model variable if multiple observations are made that share a feature while differing in other aspects.

This presentation is not intended to offer a definitive resolution to TOEFL 2000's student-model question. Its focus is rather to illustrate the concepts and tools that are available to carry out principled inference, regardless of which model is used. While the determination of the student model remains at issue, the kinds of tasks that are envisaged force us to deal with more complex

relationships among student-model variables and observable task performance variables. The following sections highlight inferential issues by illustrating how they arise under two rather different perspectives found in the language testing literature, namely, a task-centered view and a competence-centered view. The former can be viewed as extension of the inferential approach employed in the current TOEFL, to accommodate the reconception of language proficiency implied by integrative and contextualized tasks. The latter departs more radically from current procedures, incorporating student-model variables motivated by the Bachman model. It goes without saying that any of these approaches would need to be tested, criticized, and revised in light of empirical data before operational use.

### 6.2 Task-Centered Student Modeling

One approach to accumulating and reporting examinees' proficiencies in a TOEFL 2000 test would be to retain skill-based scores, but now for Reading, Writing, Speaking, and Listening (R, W, S, and L). There is a long tradition of reporting language proficiency in these terms, some of which evolved under the structuralist view of language competence (e.g., the current TOEFL), but some of which evolved to summarize performance in more authentic proficiency contexts that implicitly honor the tenets of communicative competence. Bachman and Palmer (1996, pp. 75 ff.) argue that "it is not useful to think in terms of 'skills,' but to think in terms of specific activities or tasks in which language is used purposefully. Thus rather than attempting to define 'speaking' as an abstract skill, we believe it is more useful to identify a specific language use task that involves the activity of speaking, and describe it in terms of its task characteristics and the areas of language ability it engages." This approach is taken directly in the following section. This section describes an indirect approach to the same end: A TOEFL 2000 assessment that reported R, W, S, and L scores would need to do so in a way that explicates the relationship between those scores and the behaviors observed (and expected) in specifiable language use situations.

An important step in this direction can be accomplished with tasks that focus on a modality, as in the document literacy scale of the Survey of Young Adult Literacy (SYAL; Kirsch & Jungeblut, 1986). After carefully delineating situations and uses to define a proficiency domain (re Sections 4.1 and 4.3),

cognitively relevant features that characterize tasks were used to describe expected outcomes of persons on a single proficiency variable (re Sections 4.2, 4.4, and 4.5). An examinee with an IRT $q$ of 1, for example, might be expected to manage unfamiliar tasks that require matching information across two organizing categories of a document, but have only even odds on tasks with requiring three matches (Sheehan & Mislevy, 1990; see McNamara, 1996, chapter 7, for further discussion and examples of exploring the meaning of IRT scales through task features).

To date, such applications have been limited to collections of tasks that tap a single student-model variable and are conditionally independent. Extension to the integrated and contextualized tasks proposed for TOEFL might be carried out in the manner depicted in Figure 6. Certain features are worth mentioning:

- The student model contains the four reporting variables $\theta_R$, $\theta_W$, $\theta_S$, and $\theta_L$. The relationships among them are empirical associations in the target population, specific to performance on tasks possessing the characteristics, and being assembled under the constraints, specified in the assessment design.

- The observables associated with tasks indicate their parents with "stubs" that represent where student-model and task-model BIN fragments must be connected when the task is administered.

- Some conditionally independent tasks addressing a single modality are included in the assessment to ground the definition of $\theta_R$, $\theta_W$, $\theta_S$, and $\theta_L$ (e.g., $X_1$ and $X_2$, associated with Tasks 1 and 2, both depend on $\theta_R$ only). As with the SYAL (also see Mosenthal & Kirsch, 1991), the conditional probabilities of response to these items, given their single $\theta$ parent, can be modeled in terms of selected cognitively-relevant features that influence difficulty, as in Figure 4 (the higher level DAGs are not shown in Figure 6 to save space). These features establish an interpretation of the $\theta$s beyond norm-referenced information. Other asks' features are used to control task selection, to balance content, situation, context, and other features of tasks across examinees.

- Some observables have multiple $\theta$s as parents (e.g., $X_3$, associated with Task 3, and $X_4$-$X_6$, associated with Task 4). Certain dinosaur items, for example, have a student read a passage about one theory for the extinction of dinosaurs, then ask her to write a response with specified features. Both $\theta_R$ and $\theta_W$ are parents of such an item; their relationship is conjunctive, and values of conditional probabilities

depend on both the reading-demand features and the writing-demand features, as they are defined and used for the single $\theta$ items.

- Some tasks generate multiple observable variables (e.g., observables $X_4$-$X_6$, all associated with Task 4). The dinosaur task requires several responses, with different mixes of parent $\theta$s and different values of variables that drive conditional probabilities, but all share the subject matter of dinosaurs.

With only four variables included in the student model, it is clear that many aspects of examinee skills and knowledge are confounded, and others are neglected. Some, such as general cognitive skills, grammatical competence, and aspects of illocutionary and sociolinguistic competence, will influence performance to some degree in all tasks; they account in part for the associations among $\theta$s. Others, such as motivation and affective response, are confounded with levels of performance; this model cannot distinguish low motivation or discomfort with the testing situation, for example, from lack of competence. Still others, such as examinees' differing profiles of skills and knowledge within the broadly-defined $\theta$s and their felicitous or debilitating interactions with particular contexts and task methods, will constitute sources of uncertainty about the $\theta$s so defined.

### 6.3 Competence-Centered Student Modeling

The approach illustrated in this section could use many of the same task variables and test assembly rules described in the preceding approach, but would accumulate evidence in terms of performance in variables motivated by Bachman's model of communicative competence. We should emphasize that competence variables could be defined at lower or higher levels of his model, or derived from a different or competing model. This choice is meant merely to illustrate inferential issues with some degree of complexity, without becoming notationally or graphically overwhelming. The diagramming conventions in Figure 8 are the same as those in Figure 6 above. The following points concern differences with respect to student-model variables:

- Student-model variables now appear for Grammatical Competence, $\theta_{GC}$; Sociolinguistic Competence, $\theta_{SC}$; and Conversational and Correspondence Competence, $\theta_{CVC}$ and $\theta_{CRC}$, which correspond to Discourse Competence in the Bachman model but distinguish between the forms and skills associated with Speaking/Listening and Reading/Writing (Bachman & Palmer, 1996, p. 128, attribute these

terms to Widdowson, 1978). These variables can serve as parents for observable variables that tap different modalities—$\theta_{GC}$ or $\theta_{SC}$ allowable for observables associated with any of the four traditional skills, to the degree they demand these competences, $\theta_{CVC}$ for observables involving speaking and/or listening, and $\theta_{CRC}$ for observables involving reading and/or writing. Conditional probabilities for observable variables with these parents will be functions of the degree and nature of demand on the given competence a task demands, as implied by task-feature variables again as in Figure 4.

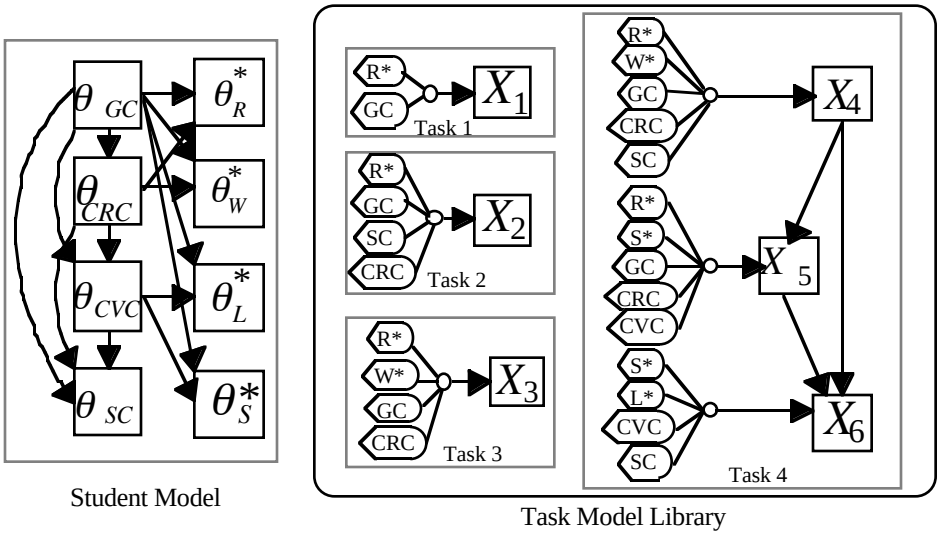


*Figure 8.* A "competence-oriented" DAG. Information about examinee performance is accumulated in variables associated with the four competences suggested by models of "communicative competence," with simple "selector" variables associated with skill modalities included to indicate the degree to which the examinee is able to exhibit those competences in performances that require functioning within the indicated modalities. Conditional probabilities of task responses are modeled in terms of cognitively-relevant task features (not depicted).

- Student-model variables also appear for Reading, Writing, Speaking, and Listening, but their operational definitions depart radically from the preceding example (the star notation emphasizes the distinction). These modality variables now serve primarily as "selector" variables, indicating which modalities are involved in a given observable. In this way they account for the common observation of examinees' differing profiles of strength in different modalities, above and beyond the cross-modality competencies discussed above. An observable would have a given $\theta^*$ as a parent if the modality was required in its negotiation. A given examinee's $\theta^*$ values would indicate the degree to which her cross-modality competences were either enabled or prohibited when carrying out tasks requiring that modality. The relationship among

these variables and the competences, for any given observable in which they were parents, would thus be conjunctive; e.g., a "fuzzy AND" gate.

In this approach, an examinee's performance across the balance of task types (specified as to characteristics of situation, materials, and use, as per Section 4.3) would be summarized in terms of cross-modality competencies and a profile of strengths and limitations associated with the modalities that must be employed to evidence those competences. For reporting purposes, projections could be made from these multivariate profiles to "scores" on designated sets of tasks of different types—one market basket of tasks related to, say, classroom interactions, and a different market basket representing interactions that teaching assistants have with students.

Figure 9 presents a simpler version of the student model, achieved through graphical-modeling approximation strategies. A "communicative competence" variable has been incorporated to model associations among the more narrowly defined cross-modality competences. No observables would have this variable as a parent, and its function is strictly utilitarian; its values might never be used for score reports or decision making, as market basket projections would give a better indication of students' communicative competence as it is currently construed. The anticipated strong associations between the Correspondence Competence variable and the Reading and Writing variables, and between Conversational Competence and Listening and Speaking variables, have been modeled explicitly. But other associations among the modality variables and the competences have been dropped, following the rationale in Patz and Mislevy (1995): With this simplification, one gains computing efficiency and retains consistent estimates of student-model variables, although trading away some precision in estimation.

## 7.0 Next Steps

A clear understanding of just what is involved in successful applications of IRT-CAT is a useful first step toward extending the approach to more complex settings. Probability-based inference with graphical models offers a framework for expressing, then confronting, the problems that will arise. Despite preliminary successes, there are still a large number of issues that must be addressed to develop a theory of graphical-model-based assessment, with fixed tests as well as

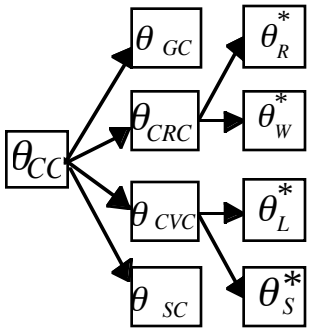


*Figure 9.* A simplified approximation for the "competence-oriented" DAG.

CAT. We have noted above the importance of the cognitive foundation of an application. Among the attendant technical challenges we have begun to address are the following.

**Knowledge-Based Model Construction (KBMC).** KBMC (Breese et al., 1994) concerns the dynamic construction and manipulation of graphical models, adapting to changes in knowledge status but in importance of the questions being asked; that is, revising models to reflect changing frames of discernment, to use Shafer's (1976) phrase, as well as changing states of knowledge and changing external situations. IRT-CAT adapts to changing knowledge states within a static frame of discernment—the question is always "What is $\theta$?"—and uses information formulas and task-based blocking and overlap constraints to select items. Generalizations of these rules are required for more complex models, in which different subparts of the model may shift into and out of attention.

**Task induced dependencies.** A task/evidence model could provide common descendants of two conditionally independent variables in the student model. Collapsing over tasks will produce new edges in the student model. The theory of GM-CAT will require both approximation techniques for determining when these edges can be observed and techniques for dynamic recompilation of the junction tree. Jaakkola and Jordon (in press) present a promising approach to this problem using variational techniques.

**Continuous variables in student models.** The most common graphical model with both continuous and discrete variables is the Conditional Gaussian (CG) model (Lauritzen & Wermuth, 1989). These models all have continuous (normal) variables conditioned on the discrete variables. In educational testing, however, it seems more natural to have the discrete item responses

conditioned on the continuous student proficiencies. Perhaps the multivariate IRT of Segall (1996) (a multivariate extension of the Rasch model) can be pressed into service here, but the lack of a closed form solution will require numerical solutions that can fit the dynamic requirements of CAT. The difficulty is that there are no closed form solutions when continuous variables are parents of discrete items; however, Jaakkola and Jordon (1997) present a possible approximation technique.

**Model fit.** More complex student models and task performance variables increase the analyst's burden in fitting, checking, and improving models. A particular advantage of using probability-based inference is that standard statistical techniques can be brought to bear on many of these questions, as Spiegelhalter et al. (1993) discuss in connection with the use of Bayes nets in expert systems more generally. In addition, more specialized diagnostics for models with unobservable variables can be adapted from the psychometric literature; see, for example, Stout (1987) on assessing dimensionality in IRT, and Levine and Rubin (1979) on detecting aberrant response patterns.

## REFERENCES


Almond, R. G. (1995). *Graphical belief modeling.* London: Chapman and Hall.

Bachman, L., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing, 12,* 238-257.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice.* Oxford: Oxford University Press.

Bejar, I. I. (1990). A generative analysis of a three-dimensional spatial task. *Applied Psychological Measurement, 14,* 237-245.

Breese, J. S., Goldman, R. P., & Wellman, M. P. (1994). Introduction to the Special Section on Knowledge-Based Construction of Probabilistic and Decision Models. *IEEE Transactions on Systems, Man, and Cybernetics, 24,* 1577-1579.

Chalifour, C., & Powers, D. E. (1989). The relationship of content characteristics of GRE analytical reasoning items to their difficulties and discriminations. *Journal of Educational Measurement, 26,* 120-132.

Chapelle, C., Grabe, W., & Berns, M. (in press). *Communicative language proficiency: Definition and implications for TOEFL 2000* (TOEFL Monograph Series). Princeton, NJ: Educational Testing Service.

Chi, M. T. H., Feltovich, P., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5,* 121-152.

Dodd, B. G., De Ayala, R. J., & Koch, W. R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement, 19,* 5-22.

Embretson, S. E. (1993). Psychometric models for learning and cognitive processes. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 125-150). Hillsdale, NJ: Lawrence Erlbaum Associates.

Freedle, R., & Kostin, I. (1993). The prediction of TOEFL reading item difficulty: Implications for construct validity. *Language Testing, 10,* 133-170.

Green, B. (1978). In defense of measurement. *American Psychologist, 33,* 664-670.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 147-200). New York: American Council on Education/Macmillan.

Heckerman, D., Horvitz, E., & Middleton, B. (1993). An approximate nonmyopic computation for value of information. *IEEE Transaction of Pattern Analysis and Machine Intelligence, 15*, 292-298.

Hively, W., Patterson, H. L., & Page, S. H. (1968). A "universe-defined" system of arithmetic achievement tests. *Journal of Educational Measurement, 5*, 275-290.

Hudson, T. (1996). *Assessing second language academic reading from a communicative competence perspective: Relevance for TOEFL 2000* (TOEFL Monograph MS-4). Princeton, NJ: Educational Testing Service.

Hymes, D. (1972). On communicative competence. In J. Pride & J. Holmes (Eds.), *Sociolinguistics* (pp. 269-293). Harmondsworth: Penguins Books.

Jaakkola, T., & Jordan, M. (in press). Recursive algorithms for approximating probabilities in graphical models. *Advances in Neural Information Processing Systems.*

Jaakkola, T., & Jordan, M. (1997). A variational approach to Bayesian logistic regression models and their extensions. *Preliminary papers of the Sixth International Workshop on Artificial Intelligence and Statistics* (pp. 283-294). January 4-7, 1997, Fort Lauderdale, FL.

Kirsch, I. S., & Jungeblut, A. (1986). *Literacy: Profiles of America's young adults.* Princeton, NJ: National Assessment of Educational Progress/Educational Testing Service.

Lauritzen, S. L. (1996). *Graphical models.* New York: Oxford University Press.

Lauritzen, S. L., & Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *Annals of Statistics, 17,* 31-57.

Levine, M., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics, 4*, 269-290.

Linacre, J. M. (1989). *Multi-faceted Rasch measurement.* Chicago: MESA Press.

Madigan, D., & Almond, R. G. (1996). Test selection strategies for belief networks. In D. Fisher & H. J. Lenz (Eds.), *Learning from data: AI and Statistics IV* (pp. 89-98). New York: Springer-Verlag.

McNamara, T. (1996). *Measuring second language performance.* London: Addison Wesley Longman.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education/Macmillan.

Mislevy, R. J., & Gitomer, D. H. (1996). The role of probability-based inference in an intelligent tutoring system. *User-Mediated and User-Adapted Interaction, 5,* 253-282.

Patz, R. J., & Mislevy, R. J. (1995). On ignoring certain conditional dependencies in cognitive diagnosis. *Proceedings of the Section on Bayesian Statistical Science of the American Statistical Association* (pp. 157-162). Washington, DC: American Statistical Association.

Mislevy, R. J., Sheehan, K. M., & Wingersky, M. S. (1993). How to equate tests with little or no data. *Journal of Educational Measurement, 30,* 55-78.

Mosenthal, P. B., & Kirsch, I. S. (1991). Toward an explanatory model of document literacy. *Discourse Processes, 14,* 147-180.

Nissan, S., DeVincenzi, F., & Tang, L. (1996). *An analysis of factors affecting the difficulty of dialogue items in TOEFL listening comprehension* (TOEFL Research Report #51). Princeton, NJ: Educational Testing Service.

Ohlsson, S. (1987). Some principles of intelligent tutoring. In R. W. Lawler & M. Yazdani (Eds.), *Artificial intelligence and education* (Vol. 1, pp. 203-237). Norwood, NJ: Ablex.

Owen, R. A. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70,* 351-356.

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

Richards, J. C. (1983). Listening comprehension: Approach, design, procedure. *TESOL Quarterly, 3,* 1-38.

Schum, D. A. (1994). *The evidential foundations of probabilistic reasoning.* New York: Wiley.

Segall, D. (1996). Multidimensional adaptive testing. *Psychometrika , 61,* 331-354.

Shafer, G. (1976). *A mathematical theory of evidence.* Princeton, NJ: Princeton University Press.

Sheehan, K. M., & Mislevy, R. J. (1990). Integrating cognitive and psychometric models in a measure of document literacy. *Journal of Educational Measurement, 27,* 255-272.

Spiegelhalter, D. J., Dawid, A. P., Lauritzen, S. L., & Cowell, R. G. (1993). Bayesian analysis in expert systems. *Statistical Science, 8,* 219-283.

Spiegelhalter, D. J., & Knill-Jones, R. P. (1984). Statistical and knowledge-based approaches to clinical decision-support systems, with an application in gastroenterology (with discussion). *Journal of the Royal Statistical Society, Series B, 147,* 35-77.

Steinberg, L. S., & Gitomer, D. G. (1996). Intelligent tutoring and assessment built on an understanding of a technical problem-solving task. *Instructional Science, 24,* 223-258.

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17,* 277-292.

Stout, W. (1987). A theory-based nonparametric approach for assessing latent trait multidimensionality in psychological testing. *Psychometrika, 52,* 589-617.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51,* 567-577.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (1990). *Computerized adaptive testing: A primer.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Waters, A. (1996). *A review of research into needs in English for academic purposes of relevance to the North American higher education context* (TOEFL Monograph Series Report No. 6). Princeton, NJ: Educational Testing Service.

Widdowson, H. G. (1978). *Teaching language as communication.* Oxford: Oxford University Press.