Graphical Models and Computerized Adaptive Testing

Russell Almond and Robert J. Mislevy
Educational Testing Service

Abstract

This paper synthesizes ideas from the fields of graphical modelling and eductational testing, particularly Item Response Theory (IRT) applied to Computerized Adaptive Testing (CAT). Graphical modelling can offer IRT a language for describing multifaceted skills and knowledge and disentangling evidence from complex performances. IRT-CAT can offer graphical modellers several ways of treating sources of variability other than including more variables in the model. In particular, variables can enter into the modelling process at several levels: (1) in validity studies (but not in the ordinarily used model), (2) in task construction (in particular, in defining link parameters), (3) in test or model assembly (blocking and randomization constraints in selecting tasks or other model pieces), (4) in response characterization (i.e. as part of task models which characterize a response) or (5) in the main (student) model. The paper describes an implementation of these ideas in a fielded application: HYDRIVE, a tutor for hydraulics diagnosis.

1.0 Introduction

Computerized adaptive testing (CAT; Wainer et al., 1990) is perhaps the most significant advance 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, in particular, to applications grounded in the more complex perspective of cognitive psychology. These constraints are the limited scope of tasks which 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; Lauritzen, 1996; Almond, 1995; often called Bayesian Networks 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 simpistic nature and strong independence assumptions of the IRT-CAT model, its users have developed quite sophisticated techniques to ensure its success. Many variables seemingly ignored by the IRT model actually enter into the task creation and test assembly processes (often informally). As graphical modellers 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 et al., 1994) these techniques can be adapted to graphical modelling as well.

This paper synthesizes ideas from graphical modelling 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 (GMs; Almond, 1995; Pearl, 1988). The simplicity of IRT as a GM is deceiving. Section 4 describes classes of variables which do not appear in the IRT model, but are handled informally or implicitly in practical applications of IRT-CAT. 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 illustrates their use in HYDRIVE, an intelligent tutoring system for aircraft hydraulics troubleshooting (Mislevy & Gitomer, 1996; Steinberg & Gitomer, 1996). 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

IRT posits a collection of simple models for the examanee's propensity to make correct responses on a series of test items which are independent given an unobservable proficiency variable θ. A simple example is the Rasch model for n dichotomous test items:

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

where x_j is the response to Item j (1 for right, 0 for wrong), β_j is the 'difficulty parameter' of Item j, and $\mathrm{P}(x_1,\ldots,x_n | \theta,\beta_1,\ldots,\beta_n) = \prod_j \mathrm{P}(x_j | \theta,\beta_j)$. For selecting items and scoring examinees in typical applications, point estimates of the item parameters are based on very large samples of examinee responses and treated as known . Section 4.2 below will discuss modeling alternative sources of information, and remaining uncertainty, about $(\beta_1,\ldots,\beta_n)$, or $\mathbf{B}$ for short.

Equation (1) is interpreted as a likelihood function for θ, say $L(\theta|\mathbf{x},\mathbf{B})$, once a response vector $\mathbf{x} = (x_1,\ldots,x_n)$ is observed. 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 proceeds from the posterior distribution $p(\theta|\mathbf{x},\mathbf{B}) \propto L(\theta|\mathbf{x},\mathbf{B})p(\theta)$, which provides 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 θ, with greater precision when θ 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 θ in light of the responses to the first k items, say $\mathbf{x}^{(k)}$ (Wainer et al., 1990, Chap 5). One common approach evaluates $\hat{\theta}$ after each response, then selects the next item from the pool which provides a large value of Fisher information in the neighborhood of $\hat{\theta}$. A Bayesian approach determines the next item as the one which minimizes expected posterior variance, or $E_{x_j}\left[Var\left(\mathbf{x}^{(k)},x_j,\mathbf{B}^{(k)},\beta_j\right)\middle|\mathbf{x}^{(k)},\mathbf{B}^{(k)}\right]$ (Owen, 1975). Section 4.3 addresses additional constraints on item selection, such as item content and format. 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, or

$$p(X_1,\ldots,X_n) = \prod_{j=1}^{n} p\left(X_j\middle|\{\text{"parents" of } X_j\}\right), \tag{2}$$

where the "parents" of X_j is the subset of $\{X_{j-1},\ldots,X_1\}$ upon which X_j is directly dependent. In the 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. See Spiegelhalter et al. (1993) for a review of recent statistical developments.

Figure 1 shows the DAG that corresponds to IRT. The first panel suppresses the dependence on item parameters, while the second makes it explicit by indicating that the conditional probability distribution of each X_j given θ is a function of β_j. The posited conditional independence of item responses θ is the structure Spiegelhalter and Knill-Jones (1984) called an "idiot's Bayes" model. This depreciative term is undeserved in thoughtful implementations of IRT-CAT, because many variables that do not appear in the simple model have been handled behind the scenes, expressly to make sure that its simple structure will suffice for the task at hand.

One way of describing IRT-CAT from the perspective of graphical models is through the DAG with θ 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 θ node is $p(\theta)$; each item can be checked to find one that minimizes expected posterior variance; it is administered, and the process repeats after the response, now starting from $p\left(\theta\middle|x^{(1)}\right)$. The process is repeated with each successive $p\left(\theta\middle|\mathbf{x}^{(k)}\right)$ until testing is terminated. At each step, the observed value of the administered variable is conditioned upon, the distribution of θ is updated, and expectations for items as-yet-unadministered is revised for calculating expected variance if it were presented next.

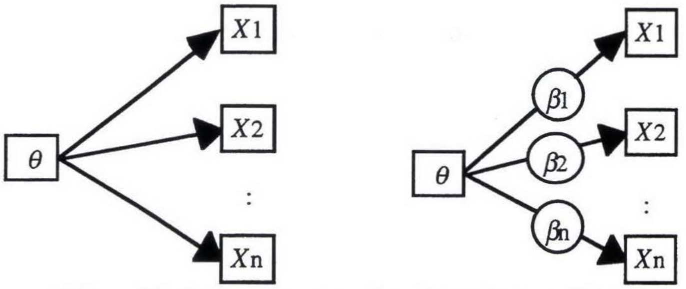

Figure 1: DAGs for an IRT model. Item parameters that determine conditional distributions of Xs given θ are implicit in the left panel and explicit in the right panel.

A second way of describing 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 θ and a library of nodes corresponding to test items, any of which can be "docked" with the θ node to produce a dyadic DAG as shown in the righthand panel of the figure. This small DAG is temporarily assembled to absorb evidence about θ from the response to a given Item *j*. It is disassembled after the response is observed and the distribution of θ updated accordingly. The new status of knowledge about θ either guides a search of the item library for the next item to administer or provides the grounds to terminate testing.

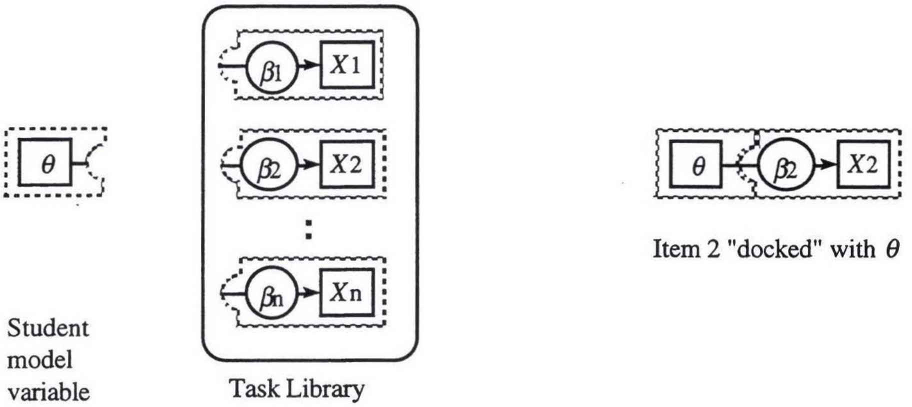

Figure 2: Left panel shows θ node and task-node library. Right panel shows Item 2 "docked" with θ to create a dyadic DAG.

4.0 Classes of Variables Involved 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 Speaking, Listening, and Writing gives the misleading impression that everything that's happening can be understood in terms of simple, one-variable, student model—the overall proficiency in a given scoring area. But many more variables are being managed behind the scenes; partly to create the variable being measured, and partly to ensure that the simple analytic model will adequately characterize the information being gathered. A quote from Pearl (1988, p. 44) captures the spirit: "[C]onditional independence is not a grace of nature for which we must wait passively, but rather a psychological necessity which we satisfy actively by organizing our knowledge in a specific way."

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 variation among examinees with respect to some features tends to accumulate over tasks, while variation with respect to other features doesn't tend to accumulate over tasks (Green, 1978). The variation that accumulates becomes "what the test measures", or the "construct" of interest. Other sources of variation introduce uncertainty about an examinee's standing on the construct.

What practices have evolved to guide testing practice under this perspective? This section discusses ways that the following classes of variables are addressed in IRT-CAT:

1. Variables addressed at the level of validity studies
2. Variables modeled at the level of task construction
3. Variables addressed at the level of test assembly
4. Variables that characterize responses
5. Variables included in the student model

Only the last class appears explicitly in the measurement model—in the case of IRT-CAT, the single domain-proficiency variable θ. We submit that θ should not be thought of as a pre-existing characteristic of examinees, but rather as a summary of evidence about a construct only brought into being through choices about, and manipulation of, many other "hidden" variables.

4.1 Variables Addressed at the Level of Validity Studies

Myriad aspects of learners' skills and knowledge effect their behavior in some area of interest, and not all of them can be addressed in any particular test. We must consider which aspects are most salient to the job at hand, and determine which of their facets to feature in the test and which to exclude. In TOEFL, for example, do we want to include scenarios that span all of college life, from doing the laundry to interacting with campus police, or shall we limit our attention to academic and classroom interactions? Should listening skills be assessed with closed-form items based on short tape-recorded segments, or in tasks that combine listening with speaking in a conversation with a human examiner? What do we lose if we trade off productive aspects of problem solving in return for the efficiencies of multiple-choice items?

As an example, multiple-choice items can economically provide indirect evidence about students' writing capabilities, since the tendencies to write effectively and to recognize effective choices among proffered alternatives are strongly associated—yet errors occur because some students are atypically better at one kind of task than the other. If the two kinds of tasks provide information useful enough to justify gathering data, but different enough to violate conditional independence, then the domain can be split into two separate pools of more nearly conditionally independent tasks. Type-of-task is crucial in defining these two θs, but imperceptible when testing with either.

A validity study gathers information across a broader range of scenarios than can be included in an operational test, along with additional information such as ratings of actual behavior in real contexts, to explore how test-specification decisions shape the information that a test provides, in light of its intended purposes and the available resources. When logistical constraints prohibit explicitly testing for certain kinds of tasks, external studies can examine the relationship between measured proficiency and performance on those tasks (Messick, 1989).

As an example, studies are used to ensure that test items are fair across gender and ethnic background. Differential item functioning (DIF) occurs when certain test content or format features prove relatively harder to members of different subpopulations for reasons unrelated to the skills and knowledge of central interest. Reading comprehension questions concerning baseball, for example, tend to be more difficult for girls than boys who would perform similarly on items in other topics. The DAG in Figure 3 depicts this unwelcome situation. Some potential causes of DIF can be avoided by defining variables that

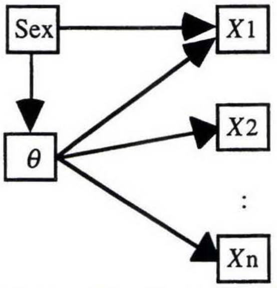

Figure 3: A DAG illustrating Differential Item Functioning (DIF). Response probabilities for Item 1 are dependent on sex, beyond associations "explained" by θ. Response probabilities of Items 2-n are conditionally independent of sex given θ.

identify problematic features of tasks, and excluding any tasks that posses these features from the domain. (Interestingly, 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.)

In sum, variables dealt with at the level of validity strongly influence what the test measures—the operational definition of the construct—yet they are not obvious once the test is set, because they effect decisions to constrain some aspects of items in the domain and preclude others entirely.

4.2 Variables Modeled at the Level of Task Construction

Individual tasks in a test can be described in terms of many variables. They concern such things as format, content, modality, situation, vocabulary load, grammatical structure, mathematical formulas required, cognitive processing requirements, 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, test developers 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.

Mislevy, Sheehan, and Wingersky (1993) have found that in unidimensional tests, these kinds of variables can be strong predictors of IRT item parameters. We noted above that item parameters are typically estimated well enough to be treated as known. For the more complex IRT model used in the GRE, for example, this requires pretesting each item on about 1000 examinees; i.e., administering it in a real testing context, but without using the responses for measuring those examinees. Mislevy et al. (op cit.) found that features of paragraph comprehension items that were related to difficulty but not included in test specifications provided enough information about item parameters to cut pretesting requirements to 250 examinees. In effect, they created the second-order DAG for modeling item parameters shown in Figure 4.

A second way to exploit 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 frameworks', for developing families of tasks with characteristics with 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 (see Hively et al., 1968, for a proposal along these lines before the days of IRT, and Embretsen, 1993, for a more recent investigation using more contemporary cognitive and measurement theory). Specific features of items within schemas affect the individual items' operating characteristics. They are residual variation of item parameters given values of variables that are monitored at this level, which can be reduced by item-specific information from pretest examinees.

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 β_j allows us to calculate the probability of a correct response from an student with any given θ. Conversely, we can give meaning to a value of θ 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 βs, then, we can describe the student's proficiency in terms of cognitively relevant skills: for example, "a student with θ=2 can usually solve document literacy problems requiring 3 feature matches in crossed lists on familiar topics, but have trouble with problems in unfamiliar contexts" (Sheehan & Mislevy, 1990). Section 5 discusses how these ideas extend to categorical student model variables defined by generalized descriptions of behavior.

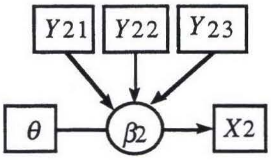

Figure 4: A 2-level DAG, with model for a model for the item parameter β which gives the conditional probabilities of the response to Item 2 given θ. Y_{21}, Y_{22}, and Y_{23} are coded features of Item 2.

4.3 Variables Addressed at the Level of 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 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 1 or 2 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 which 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 3-4-5 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 blocking constraints in place—its simplicity 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, but there are many topics within each genre and one is selected at random. Whether an examinee happens to be familiar or unfamiliar with a given topic undeniably affects her performance, but this interaction is not modeled; randomizing, the examiner leans on large sample theory to average over the effects.

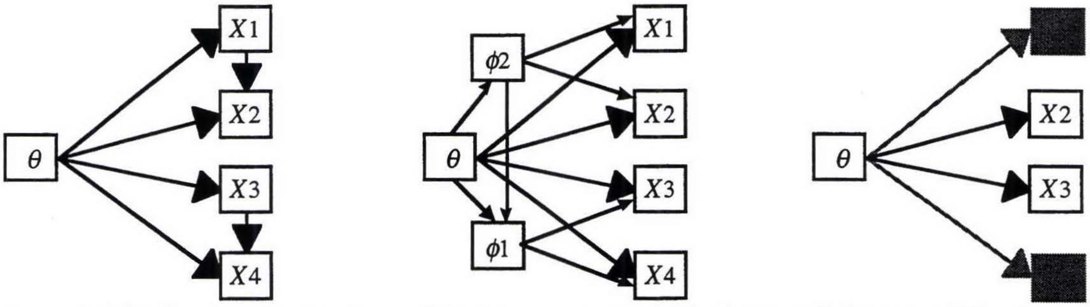

Figure 5: 3 DAGs related to overlap and blocking constraints. The first panel shows conditional dependencies among item sets. The second shows conditional independence achieved by adding student-model variables. The third shows conditional independence achieved within the IRT model by constraining what can be observed.

4.4 Variables that Characterize Responses

Characterizing student responses is straightforward with multiple-choice items in IRT-CAT: Did the student indicate the option prespecified as 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. Section 5 below expands on this theme for more general situations.

IRT models have been extended beyond dichotomous data to 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 θ. Dodd et al. (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 require a further step of evaluation in terms of abstracted properties of less constrained response behaviors.

4.5 Variables Included in the Student Model

Student model variables integrate information across distinct items 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 it should also be consistent with empirical response patterns and theories of performance in the domain.

The current TOEFL has three student-model variables—listening, reading, and writing, or L, R, and W—and each is evidenced by discrete tasks of its type only, with disjoint item domains and associated domain proficiency variables. θ_L, θ_R, and θ_W, each as depicted in Figure 1. These variables are used in for infrequent but consequential decisions such as the hiring of graduate teaching assistants from among non-native speakers of English. In contrast, an intelligent tutoring system (ITS) must define student-model variables at a finer grain-size, since it has 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). It is worth noting in passing an important tradeoff between accuracy and reliability: For a given number of observations, a more parsimonious model with fewer variables acquires more information for its variables than an ambitious model with more variables.

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 below discuss multidimensional student models further.

5.0 Graphical-Model Based Computerized Adaptive Testing

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 (1) complex tasks, in order to elicit evidence that draws upon multiple and interrelated aspects of skill and knowledge, and (2) 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 instead extending them. We can build on the same ideas of defining unobservable variables to "explain" patterns of observable responses, and "some patterns 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, noting how it builds on the issues discussed above that IRT-CAT has successfully confronted.

The basic idea is to express aspects of skill and knowledge in terms of unobservable student model variables and aspects of task performances as observable variables, then model probabilities of the performance variables conditional on the student model variables. Associations among student model variables express such relationships as prerequisition, empirical correlation, or logical connections including conjunction and disjunction. Associations among observable variables, beyond those induced by student model variables, express shared contexts or the kinds of incidental connections that overlap constraints would disallow in IRT-CAT.

Figure 6 offers a hypothetical example for a Graphical Model based CAT (GM-CAT) for more complex TOEFL-like items, in which the student produces multiple responses in a given situation, and different combinations of Reading, Writing, Speaking, and Listening can be involved in each subcomponent of the response string. The graphical model consists of two parts, the student model and the task model pool. The structure of the student model is constant across examinees. Each starts out with the same student model (based on population characteristics), but their performance in tasks updates the model differentially. The tasks can draw on the student model in complex patterns. For example, an item may require the examine to listen to an oral presentation, read instructions, and write a series of short sentences as a response. These items can be simple multiple choice or complex constructed response from which the computer or human graders extract multidimensional summaries of performance.

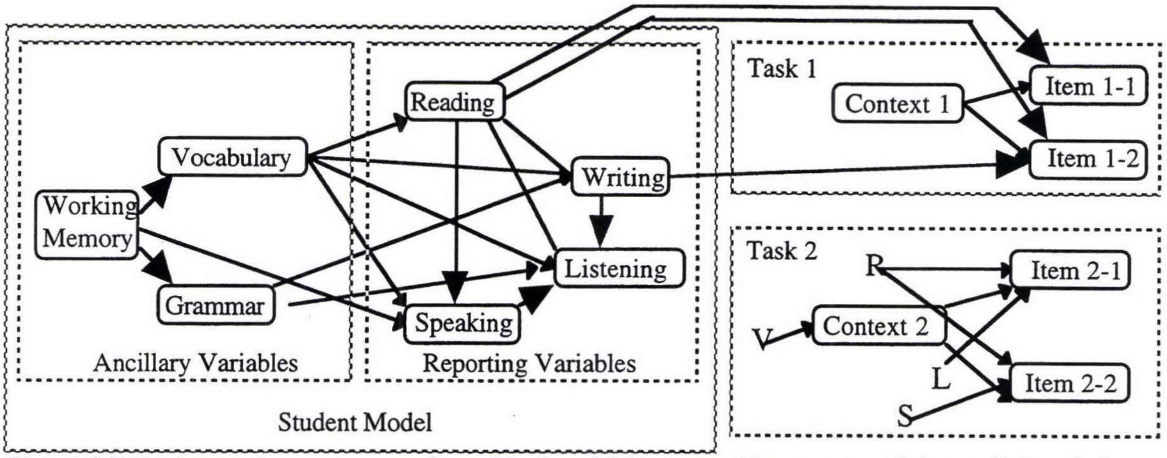

Figure 6. This picture shows a proposed graphical model for a test of language proficiency. It is a chain graph; the middle connected set of variable shows the four interrelated proficiencies which are the target of the test. The leftmost connected set shows some ancilliary proficiency variables. These two sets together make up the student model. The small graphical models attached on the right hand side are task models. The Graphical Model CAT algorithm would have its choice of tasks from the task pool.

The linkage of the task model to the student model play the role of the difficulty parameter in IRT, relating examinee proficiencies to conditional probabilities of making various patterns of responses. Task designers provide the structure of this linkage and 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. Empirical data can be used to refine the conditional probabilities for specific tasks, in light of their incidental features.

The nature of the student model variables can be extended to encompass categorical descriptions of developing competence in terms of qualities of actions in more complex and open-ended tasks, or "rubricized" student variables. An example of such a variable is "reading proficiency," as defined in the Guidelines of the American Council of Teachers of Foreign Language (ACTFL, 1989). The ACTFL description of a Low-Intermediate reader, for example, is as follows:

> Able to understand main ideas and/or some facts from the simplest connected texts dealing with basic personal and social needs. Such texts are linguistically noncomplex and have a clear underlying internal structure, for example chronological sequencing. They impart basic information about which the reader has to make only minimal suppositions or to which the reader brings personal interest and/or knowledge. Examples include messages with social purposes or information for the widest possible audience, such as public announcements and short, straightforward instructions for dealing with public life. Some misunderstandings will occur.

The ACTFL Guidelines are useful descriptors of language proficiencies, but they are underdetermined (Bachman, 1988); examinees' performances differ when they are assessed by different testing methods, and informed observers disagree about the level of a performance unless they have been intensively trained in a particular rating scheme. A significant step to imbuing meaning in such a scale would be to lay out a system of descriptors of tasks that present increasing challenges to students, then explicate, as conditional probabilities, expectations of response characteristics in tasks with given demand characteristics.

A CAT 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 and Middleton, 1993) and *weight of evidence* (Madigan & Almond, 1996) seem like promising candidates in this setting. The GM-CAT attaches the task model to the student model and absorbs the evidence provided by the examinees responses. The algorithm can then discard the task item, or maintain it in the model to measure dependence effects between tasks.

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.

As with the IRT CAT, variables can be used in GM-CAT to constrain the item selection algorithm by blocking or avoiding overlap. 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 HYDRIVE

The HYDRIVE intelligent tutoring system (Mislevy & Gitomer, 1996) illustrates graphical student modeling. HYDRIVE helps Air Force trainees develop troubleshooting skills for F-15 aircraft hydraulics systems. It simulates many of the important cognitive and contextual features of troubleshooting on the flightline. As the student performs simulated troubleshooting procedures, HYDRIVE's student model analyzes student's performance and recommends instruction and selects new tasks.

The grain-size and the nature of a student model in HYDRIVE are targeted to the instructional options available. The structure of the network, the variables that capture the progression from novice to expert hydraulics troubleshooter, and the conditional probabilities implemented in the network are based on in-depth analyses of experts and novices verbalizations of their problem-solving actions, and the observation of small numbers of students actually working through the problems in the HYDRIVE context. The ERGO computer program (Noetic Systems, 1991) is used to carry out the calculations involved in sequentially updating the student model variables.

Figure 7 is a simplified version of portions of the Bayes net through which the HYDRIVE student model is implemented and updated within a given problem. Four groups of variables appear: (1) The rightmost nodes are the "observable variables," actually the results of rule-driven analyses of student's actions in a given situation. (2) Their immediate parents are knowledge and strategy requirements for two prototypical situations addressed in this simplified diagram; the potential values of these variables, too many to depict in the limited space, are generalized "noisy-AND" combinations of system knowledge and strategies that are relevant in these situations, from the next group of variables: (3) The long column of variables in the middle concerns aspects of subsystem and strategic knowledge, corresponding to instructional options. (4) To their left are summary characterizations of more generally construed proficiencies.

"Observable" variables in HYDRIVE are not, strictly speaking, observable behaviors, but outcomes of analyses that characterize sequences of actions as "serial elimination," "redundant action," "irrelevant action," "remove-and-replace," or, in situations in which it is possible, "space-splitting"—all interpreted in light of the current state of the system and results of previous actions. HYDRIVE employs a relatively small number of interpretation rules (~25) to classify each troubleshooting action in these terms.

Potential observable variables cannot be predetermined and uniquely-defined in the manner of usual assessment items, since a student can follow countless paths through the problem. Rather than attempting to model all possible system states and specific possible actions within them, HYDRIVE posits equivalence classes of system-situation states, each of which could arise many times or not at all in a given student's work. Members of these equivalence classes are treated as conditionally independent, given the status of the requisite skill and knowledge requirements. Two such classes are illustrated in Figure 7: Canopy situations in which space-splitting is not possible, and landing gear situations in which space-splitting is possible.

Figure 7 depicts belief after observing, in three separate situations from the canopy/no-split class, one redundant and one irrelevant action (both ineffectual troubleshooting moves) and one remove-and-replace (serviceable but inefficient). Serial elimination would have been the best strategy in such situations, and is most likely only if the student has strong knowledge of this strategy and all relevant subsystems. Remove-and-replace is more likely when a student possesses some subsystem knowledge but lacks familiarity with serial elimination. Weak subsystem knowledge increases chances of irrelevant and redundant actions. All interpreted actions are possible from all combinations of student variable values; sometimes students with good understanding carry out redundant tests, for example, and sometimes students who lack understanding unwittingly make the same action an expert would. These possibilities are reflected in the conditional probabilities of actions, given the values of student model variables.

Subsystem and strategy variables serve to summarize tendencies in interpreted behaviors at a level addressed by instruction, and to disambiguate patterns of actions in light of the fact that inexpert actions can have several causes. As a result of the three inexpert canopy actions, Figure 7 shows belief shifted toward lower values for serial elimination, and for all subsystem variables directly involved in the situation: mechanical, hydraulic, and canopy knowledge. Any or all could be a problem, since all are required for high likelihoods for expert actions. Variables for subsystems not directly involved in these situations are also lower, because to varying extents, students familiar with one subsystem tend to be familiar with others, and, to a lesser extent, students familiar with subsystems tend to be familiar with troubleshooting strategies. These relationships are expressed by means of the more *generalized system and strategy knowledge* variables at the left of the figure. These variables exploit the indirect information about aspects

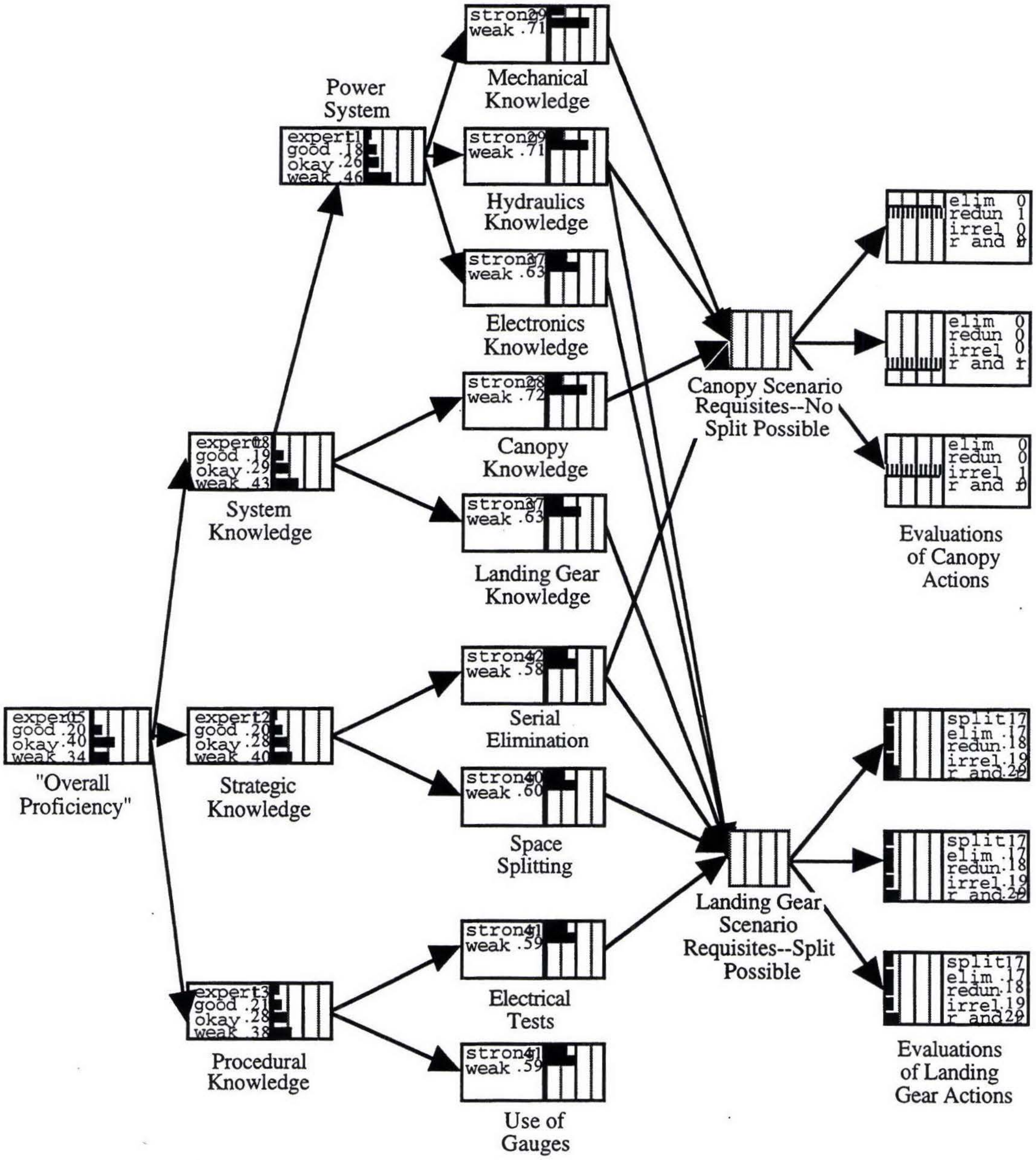

Figure 7: HYDRIVE DAG

of knowledge not directly tapped, and to summarize broadly construed aspects of proficiency for evaluation and problem-selection.

Perhaps the main lesson we take from HYDRIVE is the importance of cognitive grounding. In a field comparison against another ITS with a similar user interface and aircraft simulator, but without the cognitive student model, HYDRIVE trainees showed significantly greater increases in performance (Hall et al., 1996). The key requirements for probability-based student modeling are (1) understanding principles of the target domain and how people learn those principles, so as to structure the student model efficaciously, and (2) determining what one needs to observe, and how it depends on students' possible understandings, so as to structure observable variables and their relationship to student-model variables.

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 CAT. We have noted above the critical 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, Goldman, & Wellman, 1994) concerns the dynamic construction and manipulation of graphical models, adapting to changes in knowledge status but in importance of the questions being asked—i.e., 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 θ'—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 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.

Continuous variables in student models. The most common graphical model with both continuous and discrete variables is the Conditional Gaussian (CG) model (Lauritzen & Wermouth [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 which can fit the dynamic requirements of CAT.

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; e.g., Stout (1987) on assessing dimensionality in IRT, and Levine and Rubin (1979) on detecting aberrant response patterns.

Acknowledgments

This work was supported by the Test Of English a Foreign Language (TOEFL) and the National Center for Research on Evaluation, Standards, Student Testing (CRESST), Educational Research and Development Program, cooperative agreement number R117G10027 and CFDA catalog number 84.117G, as administered by the Office of Educational Research and Improvement, U.S. Department of Education. The HYDRIVE example is based on a project for Armstrong Laboratories of the United States Air Force, Drew Gitomer, Principle Investigator. We are grateful to Gary Buck, Drew Gitomer, Ed Herskovitz, David Madigan, and Linda Steinberg for discussions on various topics discussed here.

References

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

American Council on the Training of Foreign Languages. (1989). *ACTFL proficiency guidelines.* Yonkers, NY: Author.

Bachman, L. (1988). Problems in examining the validity of the ACTFL Oral Proficiency Interview. *Studies in Second Language Acquistition, 2,* 149-164.

Breese, J.S., R.P. Goldman, & M.P. Wellman (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.

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: Erlbaum.

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

Hall, E.P., Rowe, A.L., Pokorny, R.A., & Boyer, B.S. (1996, April). A field evaluation of two intelligent tutoring systems. Paper presented at the annual meeting of the American Educational Research Association, NY.

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.

Lauritzen, S.L. (1996) *Graphical Models.* Oxford Science Publications

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.

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

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.

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.

Noetic Systems, Inc. (1991). *ERGO* [computer program]. Baltimore, MD: Author.

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.

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

Segall, D. (1996). Multidimensional adaptive testing. *Psychometrika*, *61*(2) 331-354.

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