Journal of Educational Measurement
Winter 2007, Vol. 44, No. 4, pp. 341–359

Modeling Diagnostic Assessments with Bayesian Networks

Russell G. Almond
Educational Testing Service
Louis V. DiBello
University of Illinois at Chicago
Brad Moulder and **Juan-Diego Zapata-Rivera**
Educational Testing Service

This paper defines Bayesian network models and examines their applications to IRT-based cognitive diagnostic modeling. These models are especially suited to building inference engines designed to be synchronous with the finer grained student models that arise in skills diagnostic assessment. Aspects of the theory and use of Bayesian network models are reviewed, as they affect applications to diagnostic assessment. The paper discusses how Bayesian network models are set up with expert information, improved and calibrated from data, and deployed as evidence-based inference engines. Aimed at a general educational measurement audience, the paper illustrates the flexibility and capabilities of Bayesian networks through a series of concrete examples, and without extensive technical detail. Examples are provided of proficiency spaces with direct dependencies among proficiency nodes, and of customized evidence models for complex tasks. This paper is intended to motivate educational measurement practitioners to learn more about Bayesian networks from the research literature, to acquire readily available Bayesian network software, to perform studies with real and simulated data sets, and to look for opportunities in educational settings that may benefit from diagnostic assessment fueled by Bayesian network modeling.

Evidence-Centered Assessment Design and Bayesian Modeling

We situate our discussion within a general consideration of assessment systems, and draw from that a motivating argument for the use of Bayesian network modeling in cognitive diagnostic assessment. The National Research Council report *Knowing What Students Know* (2001) considered assessments as *evidentiary systems* consisting of three primary components, dubbed the assessment triangle: cognition, observation, and interpretation. The target of assessment, represented by the cognition vertex in the assessment triangle, is particular information about an examinee's knowledge, skills, and abilities, formulated with a certain express purpose in mind. In this paper we use the term *proficiency* to refer generally to any aspects of knowledge, skills, and abilities measured by assessments. Examinee proficiency levels are cognitive, reside within the examinee's mind, and are not directly observable. Observed data are collected from examinee performance on a set of tasks or items designed and selected for the purpose. Interpretations are made of the observed evidence, as inferences about examinees' latent proficiency levels. According to the assessment

purpose, the inferences are reported as assessment feedback to test consumers for applied uses such as supporting instruction, learning, decision making, and planning.

The Evidence Centered Design (ECD) paradigm (Mislevy, Steinberg & Almond, 2003) was developed to systematize the design and development of assessments as effective and valid evidentiary systems. The ECD paradigm begins by developing a clear understanding of the assessment purpose, and specific intended uses of assessment results, and constructs all aspects of the assessment around that core:

- identify the constructs to be measured and their expected interrelationships within the content domain being assessed, in light of the assessment purpose,
- identify observable behaviors to be captured to represent the set of constructs, and generate tasks that elicit those behaviors, and
- design the evidence and inference engines that operate on the observed data and generate inferences about student proficiencies.

The three components are animated by an underlying conceptual student model of how students represent and develop knowledge and competence within a domain. This conceptual student model is based on substantive knowledge and theory, and explicitly shaped and elaborated according to the assessment purpose. In addition to the high quality of each component, the proper functioning of an assessment requires the three components of the assessment triangle to be tightly coordinated with the underlying conceptual student model and with one another. Assessment validity depends upon the harmonious combination of these components. In particular, the validity of cognitive diagnostic assessment, based upon a necessarily complex and multidimensional substantive student model, demands psychometric models that reflect the substantive complexity to a sufficient degree.

This paper focuses on one approach—Bayesian network modeling—that is especially suited to building inference engines designed to be synchronous with the finer grained student models that arise, especially in cases of skills diagnostic testing. Across multiple contexts and circumstances, such student models demonstrate a broad range of characteristics and features. The need for flexibility and breadth forms a strong rationale for the selection of Bayesian network models as a psychometric modeling environment of choice for cognitive diagnostic assessment (Almond & Mislevy, 1999; Mislevy et al., 2003).

This paper reviews aspects of the theory and use of Bayesian network models as item response theory (IRT)-based cognitive diagnostic models. Here we intend IRT to be understood broadly to refer to item models—specification of the probability of response values as a function of examinee proficiencies and task characteristics. As we demonstrate below, Bayesian network models naturally function as item models within the probability modeling tradition of IRT, and can serve as the inference engines in the above ECD paradigm. We discuss how Bayesian network models are set up with expert information, improved and calibrated from data, and deployed to generate evidence-based inferences about proficiency levels for individuals and groups.

This paper, aimed at a general educational measurement audience, illustrates the flexibility and capabilities of Bayesian networks through a series of concrete examples, and without extensive technical detail. Section 2 offers a formal mathematical

definition of Bayesian networks, and briefly reviews Bayesian network concepts and notation. Section 3 examines two concrete examples of proficiency spaces and discusses issues related to proficiency modeling and the interrelationships among proficiencies. Section 4 describes the building of customized evidence models for complex constructed response tasks, within the context of a specific example. Section 5 considers Bayesian networks as scoring engines and remarks on several related topics, including reliability estimates and adaptive testing using weight of evidence. Section 6 presents summary and conclusions. We hope that educational measurement practitioners will be motivated to learn more about Bayesian networks from the research literature, to acquire readily available Bayesian network software, to perform studies with real and simulated data sets, and to look for opportunities in educational settings that may benefit from diagnostic assessment fueled by Bayesian network modeling.

A Brief Overview of Bayesian Networks

In this section we define Bayesian networks and discuss conditional dependencies and independencies in Bayesian network models. We identify model parameterizations and the natural division of Bayesian network models of assessment systems into two interacting sub-models: the proficiency and evidence models. (See Jensen, 1996, or Neapolitan, 2004, for good tutorial introductions to Bayesian networks.)

A Bayesian network is constructed as a pair consisting of a graphical network and a probability distribution. We define properties and terminology of graphs pertinent to Bayesian networks. Readers not interested in these formal definitions can safely skip to the next section. A *graph* $G = (\mathrm{V}, E)$ consists of a set V of *vertices or nodes*, along with a set $E \subseteq V \times V$ of ordered pairs of nodes, called the *edges* of G. An edge (v, w) is represented as an arrow pointing from node v to node w. Though we usually omit saying so, all graphs considered in this paper are *finite graphs*, i.e., the set of nodes V is finite.

For a pair of nodes w and v, if both (v, w) and (w, v) are edges of G, we say that the edge is *undirected* and picture it connecting the two nodes with no arrow on either end. A graph $G = (V, E)$ is *directed* if it has no undirected edges. A *directed path* from node v to node w in a directed graph $G = (V, E)$ is a sequence of nodes: $v = v_0, v_1, \ldots, v_n = w$, where for each i, (v_{i-1}, v_i) is a directed edge in G. A *directed cycle* in G is a directed path from v to itself. A directed graph G is *acyclic* if it has no directed cycles.

Bayesian networks are built on finite *acyclic directed graphs* (*ADG*; though commonly called directed acyclic graphs or *DAGs*—we use the more precise *ADG*).

If (v, w) is edge in G, we say that v is a *parent* of w, and w is a *child* of v. Every node v determines the set of all of its *parents pa* (v), and the set of all *children* of v, denoted *ch* (v) (either could be the empty set). The term *ancestors* of node v refers to the set of all of the parents of v together with the parents of the parents of v, etc. Similarly the set of *descendants* of node v consists of v's children, and its children's children, etc.

To prepare the way for a mathematical definition of Bayesian networks, we consider *ADG's* $G = (V, E)$ in which each of the nodes in V represents a random variable

with a finite number of possible values. For example node v could be dichotomous with states 0 and 1. Another node w could have three possible values: 0, 1, 2. At any point in time each of the nodes can take one of its finitely many possible values, or it can be unassigned. For example if X is a node representing a particular observable response to a task, then before the student responds to that task, that variable is unassigned. The set of all possible *states* of the graph $G = (V, E)$ are all possible assignments of values to all nodes. The set of all nodes together, $X_1, \ldots, X_n$, has a probability distribution called the joint distribution: $P(X_1 = x_1, \ldots, X_n = x_n)$.

A Bayesian network $B = (G, P)$ consists of a finite acyclic directed graph $G = (V, E)$ where each node X_i in $V = \{X_1, \ldots, X_n\}$ is a random variable with finitely many states, together with a joint distribution P that factors as follows:

$$P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i \mid pa(X_i)).$$

In this equation the factor $P(X_i = x_i \mid pa(X_i))$ is called the *local probability distribution* of variable X_i conditional only on the values of that node's parents $pa(X_i)$. In words the direct influence on each node X_i comes from $pa(X_i)$ the set of parents of X_i.

To appreciate the implication of this factorization, we recall here that for any set of random variables $V = \{X_1, \ldots, X_n\}$, any joint probability distribution can be factored according the multiplication rule as:

$$P(X_1 = x_1, \ldots, X_n = x_n) =$$

$$P(X_1 = x_1) * P(X_2 = x_2 \mid X_1 = x_1) * \cdots * P(X_n \mid X_1 = x_1, \ldots, X_{n-1} = x_{n-1}).$$

Further, we can reorder the variables X_i in any manner whatever, and the corresponding product is correct according to the multiplication rule. The specific factorization that comes with an *ADG* structure results in two potentially large advantages compared to factorization from the multiplication rule. First, the graphical structure, characterized by each node's set of parents, is meaningfully constructed from theory and knowledge. In an educational assessment, we may have two skills X_1 and X_2, with X_1 a logical or instructional prerequisite of X_2. We can build that relationship into the Bayesian network as a direct dependency arrow from X_1 to X_2, and thereby reflect that dependency in the local distribution at X_2. Second, the Bayesian network parent structure can be much sparser than the structure given by an arbitrary multiplication rule, with a corresponding reduction in the total number of parameters in the Bayesian network factorization, Fewer parameters results in quicker and more accurate model calibration, and a model that is easier to work with.

Deriving the set of *all* conditional independencies and dependencies satisfied by a given Bayesian network is nontrivial. Using notation $A \perp B \mid C$ to mean that A and B are conditionally independent given C, the following property is straightforward (Pearl, 1988; Heckerman, 1998):

Property 1. Let $B = (G, P)$ be a Bayes network on a set of random variables $X_1, \ldots, X_n$. For each variable X_i, define $MB(X_i) =$ the *Markov blanket of* X_i (Pearl, 1988) to be the set consisting of X_i itself, together with all parents of X_i, and all

children of X_i, and all parents of children of X_i. In symbols:

$$MB(X_i) \equiv \{X_i\} \cup pa(X_i) \cup ch(X_i) \cup (\cup_{Y \in ch(X_i)} pa(Y)),$$

where $ch(X_i)$ is the set of children of X_i and $pa(X_i)$ is the set of parents of X_i. Then X_i is conditionally independent of all other variables given its Markov blanket. Let V be the set of all variables X_i. This can be written:

$$X_i \perp (V - MB(X_i)) \mid MB(X_i).$$

A satisfactory accounting of the conditional dependencies and independencies for a given Bayesian network can be given in terms of graph separation properties, including *d-separation*; *I-map*; *D-map*, *Markov property* (Pearl, 1988). The details are outside the scope of the present paper.

Bayesian Networks for Assessment Systems

To specify a Bayesian network it is necessary to define a *structural part* that consists of an *ADG* $G = (V, E)$ and a *parametric part* that consists of all conditional probability tables required to define the local distributions $P(X_i = x_i \mid pa(X_i))$. The parameters determine the conditional probability tables for each local distribution, conditional on all possible configurations of values for the set $pa(X_i)$ of all the parents of X_i. When the local distributions are taken to be unrestricted conditional multinomial distributions, the number of parameters to be estimated can be quite large. A common procedure for reducing the number of parameters is to parameterize each node in a parsimonious way, consistent with substantive knowledge and theory. This can significantly reduce the number of parameters needed for a given node. (Almond et al., 2001; Almond, 2007)

When modeling assessment systems, the Bayesian networks naturally subdivide into two overlapping and interacting parts: (1) the *proficiency model* consisting of the nodes that represent student proficiencies to be measured, along with directed links that represent known or hypothesized relations between proficiencies, such as "prerequisite," "part-of," "is correlated with," "induces," and "inhibits;" and (2) the *evidence model* with nodes that correspond to task observables together with, for each observable, links from particular proficiencies to that observable—these are the proficiencies that are directly measured by that observable. In the case of Bayesian network models, we also have the capability to specify directed links between two observables within the same task, as needed by some applications and demonstrated in our examples below. In general, the mapping from proficiencies to task observables is many-to-many: a given task observable can measure multiple proficiencies, and, for measurement reliability, each proficiency typically will be measured by multiple task observables.

Let X_i be an observable in an assessment Bayesian network model, and for simplicity assume X_i has no observables as children. Then the Markov Blanket of X_i consists of X_i together with $pa\ (X_i)$. In this special case, Property 1 becomes: Conditional on its parents, X_i is independent of all other variables. In symbols:

$$X_i \perp (V - (\{X_i\} \cup pa(X_i))) \mid pa(X_i).$$

The probability modeling approach taken within item response theory, models the probability of an item response X_i as a function of student ability θ_n for student n and item parameters ξ_i for item i: $P(X_i = x_i \mid \theta_n, \xi_i)$. Either or both parameters θ_n and ξ_i can be multidimensional. The stochastic relation from θ_n and ξ_i to X_i is then reversed by using Bayes Theorem to derive information about ability θ_n and item characteristics ξ_i given one or more item responses X_i. We term the step of estimating the item parameters from a database of student item responses as the *model calibration* phase. The *scoring* of individual examinees means finding the posterior probability distribution of an examinee's ability parameter θ_n, allowing a proficiency inference, given the examinee's observed item responses, treating the estimated item parameters as known. A Bayesian network model for an assessment is set up in a similar manner: for each task observable X_i the local distribution $P(X_i = x_i \mid pa(X_i), \varphi_i)$ represents the probability of a particular value for response observable X_i as a function of the values of all the parents of X_i, and the conditional probability table parameters φ_i. The values of the proficiency parents $pa(X_i)$ are analogous to the ability θ_n and the conditional probability table values φ_i play the role of the item parameters ξ_i in traditional IRT.

In summary Bayesian networks provide a flexible mechanism for describing the joint probability distributions of discrete probability models using an *ADG*. They provide natural and convenient capabilities to model a proficiency variable space, including any necessary direct dependency links between pairs of proficiencies, along with, what we term in this paper, *task observable* variables, and direct links from certain proficiency nodes to each task observables. As a modeling bonus, we can represent any needed direct dependencies between pairs of observables within a given task, as we show later. The use of graphs and the corresponding factorization of the joint probability distribution has two significant benefits for complex models: (1) a graphical language for expressing direct dependencies among assessment variables, both proficiency and observable variables, and (2) some relative efficiencies of computational algorithms for both model calibration and student scoring by effectively reducing the number of parameters needed to define the joint distribution. That is the case when, for a given multidimensional proficiency space, the Bayesian network factorization of the joint distribution contains fewer parameters than would be implied by the general multiplication rule. The general multiplication rule corresponds to a fully connected *ADG* in which every pair of variables is linked.

Proficiency Models

We saw from the introduction that a valid assessment requires a proficiency model that reflects the most salient features derived from substantive knowledge and theory. Specifying a Bayesian network model requires modelers to identify proficiency nodes, specify the number of levels or other set of values for each proficiency, specify the parent link structure, and for each node specify the corresponding conditional probability tables—conditional on all configurations of values of the node's parents, specify for each task the task features that will be used as task observables, specify the links from proficiencies to observables, and specify the conditional probability tables for each observable conditioned on the set of all configurations of values of

its parents. The proficiency-to-observable link structure is similar to the information carried by the *Q*-matrix of other approaches (see Roussos, Templin, & Henson, this issue; Gierl, this issue) that specifies which skills are required by each item. Given the difficulties of learning Bayesian network structure from data in the presence of latent variables (discussed further below), most current assessment applications of Bayesian network models base the proficiency structure, as well as the links from proficiencies to task response observables, on expert information. The expert-derived structures can be locally modified based on data, but model calibration typically is focused on learning the conditional probability parameters for a given Bayesian network graphical structure; conditional probabilities for both proficiency and evidence models.

To illustrate the construction of a proficiency model, we discuss research that examined the use of a Bayesian network model as one of several possible scoring models for the Information Communication Technology (ICT) Literacy assessment (Almond, Yan, & Hemat, 2005; Yan, Almond, & Hemat, 2005).[1]

An initial content and cognitive framework for the assessment consisted of seven possible proficiency variables: *Access*, *Create*, *Communicate*, *Define*, *Evaluate*, *Manage*, and *Integrate* (Katz et al., 2004). Three levels were defined for each proficiency variable (high, medium, and low). These proficiency levels were defined concretely in terms of ECD *claims* about what an examinee at a given level actually is able to do. For example: "A person who is rated high on the *Communicate* proficiency is able to adapt a presentation for the target audience." An initial Bayesian network was constructed that posited a general *ICT Literacy* factor that was hierarchically related to the seven proficiency variables in such a way that the seven proficiency variables would be conditionally independent given this overall factor. The design team used this framework to create a task blueprint and to design and construct specific tasks for the assessment. In parallel the psychometric team investigated possible proficiency model structures.

Based on the modeling assumptions, members of an expert advisory committee provided conditional probability information for the seven proficiencies given in overall *ICT Literacy*, in the form of "pseudo-data." For example: "Given 100 students rated *high* on the *ICT Literacy* proficiency, how many do you expect to be *high*, *medium*, and *low* on the *Communicate* proficiency?" Expert pseudo-data were used to define preliminary conditional probability tables for the proficiency model.

A test blueprint was developed that specified types and numbers of tasks to be developed and how the tasks were linked to the proficiency variables. Based on expert pseudo-data and the test blueprint, a full Bayesian network was constructed, simulated data were generated, and a number of studies were made of model stability, estimated skill classification reliabilities, and model sensitivity. Given the relatively small numbers of tasks possible within constraints on testing time, the estimated reliability of each of the seven proficiencies was judged to be modest (Cohen's kappa's approximately 0.6). On that basis, the number of proficiencies to be reported was reduced to one overall ICT Literacy proficiency plus four component proficiencies (instead of one overall plus seven components) by combining each of three pairs of the component proficiencies. The resulting proficiency model is shown in Figure 1.

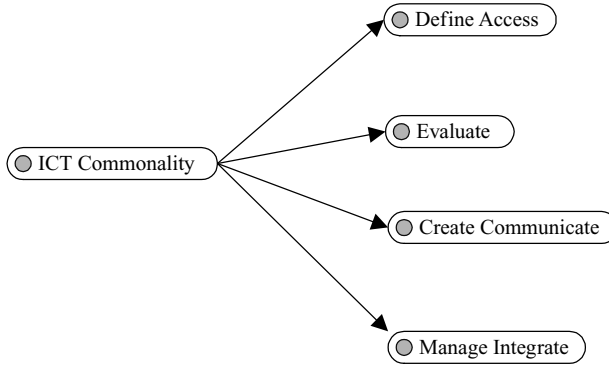

FIGURE 1. *ICT literacy candidate proficiency model.*

Up to this point, all preliminary model constructions and modifications had been accomplished prior to actual data being collected, through a systematic examination of expert-based design decisions along with creation and analysis of reasonable simulated data sets. Real data on nearly 5,000 examinees were collected in 2005 from a pilot administration of the ICT Literacy assessment. These data were used to calibrate three alternative models simultaneously: the five-proficiency Bayesian network model described above and in Figure 1 (ICT Commonality plus four component skills), and two continuous latent trait item response theory models: one unidimensional IRT Model that used only a single overall factor, with no component skills, and a second IRT model that used four independent unidimensional IRT scales, one for each of the four combined component proficiencies.

To succeed in getting a Markov chain Monte Carlo (MCMC) model calibration procedure to converge for the five-proficiency Bayesian network model, a hidden non-identifiability issue was identified and repaired. The problem was a version of the so-called label identification problem in unconstrained latent class models. If the labels were swapped for the *medium* and *low* states for the *ICT Literacy* variable, then swapping the second and third rows of the conditional probability tables for each of the other four variables would produce a mathematically equivalent model. Three MCMC chains were run to calibrate the model and two of the chains fell into the natural ordering and one into the swapped ordering. To solve the problem, the natural ordering was forced by adjusting the prior for one proficiency in a way that guaranteed that the rows would not switch. The modified model converged. A second problem arose with two of the observables measuring the *Evaluate* proficiency, that was traced to a field test form design problem. Overall the Bayesian network model offered a mechanism for separate reporting on four component proficiencies, with calibration stability comparable to the one factor model, and superior to the model with four continuous latent abilities. These findings represent some of the advantages of the Bayesian network model over the two continuous IRT models.

The proficiency model for ICT Literacy was relatively simple, though it was multidimensional as shown in Figure 1. We introduce here a more complex proficiency

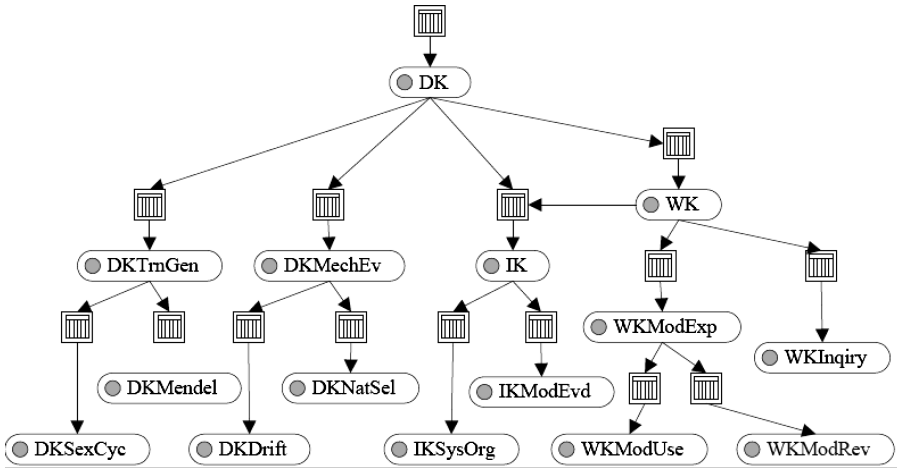

FIGURE 2. *Biomass proficiency model.*

model that was developed for a demonstration project called Biomass (Steinberg et al., 2003). The Biomass demonstration was a web-delivered prototype of an assessment system that could be used in two modes: (1) as an interactive, inquiry-based assessment of a segment of secondary school biology that served a formative purpose in supporting learning in that domain; and (2) as a culminating assessment that provided summative evidence of whether science learning standards were met, such as for college admissions or placement assessment. Consulting with a group of domain experts, a fifteen node proficiency model was developed as given in Figure 2.

The proficiency model consisted of three interconnected chunks of proficiencies: (1) Disciplinary Knowledge, consisting of the overall DK node at the top and all other proficiency nodes at the left whose names begin with DK—these represented declarative knowledge; (2) Working Knowledge, consisting of the overall WK node on the upper right and all WK nodes on the right—these proficiencies involved inquiry skills along with model explanation, usage and revision; and (3) Integrated Knowledge, consisting of the overall IK node located in the upper middle of the figure along with two other IK nodes—these nodes involved integration of domain and working knowledge, including thinking about and working with systems, models and evidence.

The complexity of this proficiency model resulted from its relatively larger size—the number of proficiency nodes was fifteen—as well as a more complex dependency structure: this proficiency model was not a tree structure. Acyclic graphs in which all nodes have at most one parent are called trees. Note that the IK (Integrated Knowledge) node had two parents: DK (Disciplinary Knowledge) and WK (Working Knowledge). Furthermore, from expert advice, the interaction between these two parents was not modeled as fully compensatory. A very low proficiency level on one of the parents would not be able to be fully compensated for with a correspondingly higher level of the other parent. Instead this interaction was modeled as *conjunctive*,

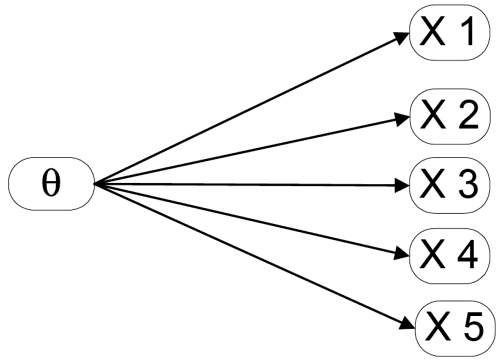

FIGURE 3. *Five task observables with 1 proficiency parent (modeled conditionally independent).*

in which a high level of IK proficiency required relatively high levels on *both* parent proficiencies DK and WK.

The two examples in this section demonstrate a capacity to build rich proficiency models according to specifications provided by experts, including "pseudo-data" elicited from experts, and analysis of model characteristics and behavior with simulated data that led to proficiency model modifications. The ICT Literacy field study resolved several calibration problems, and concluded with a potential advantage of Bayesian network models over the continuous latent trait models that were considered in that research. The Biomass proficiency model exemplifieds a more complex Bayesian network with fifteen proficiency nodes and one proficiency—IK (Integrated Knowledge)—that had two parents. We continue our discussion of Biomass in the next section.

Task-Specific Customization

For satisfying the synchrony requirement for IRT-based cognitive diagnostic assessment, modeling flexibility for evidence models is just as vital as for proficiency models. We demonstrate in this section complex evidence models developed for extended tasks that were designed to gather evidence of deeper knowledge and understanding. We describe several task model definitions created and analyzed for Biomass (Steinberg et al., 2003; Almond et al., 2001).

We begin with Figure 3 that shows one task, consisting of five multiple choice questions about implications of forms of genetic dominance. In this case the model was straightforward, and simple. The five questions were linked to parent Disciplinary Knowledge of Mendelian genetics (DKMendel, represented in this figure as θ). Each question produced a single observable and the five observables were modeled as conditionally independent given the parent proficiency level of DKMendel.

A goal of the Biomass project was to focus on higher level thinking skills and procedural knowledge, beyond simple declarative recall. Rich tasks were developed that

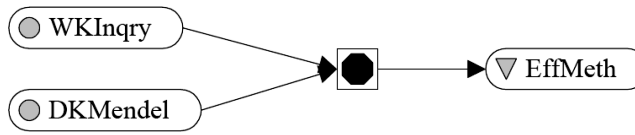

FIGURE 4. *Task observable with two proficiency parents (modeled as inhibitor).*

afforded examinees opportunities for higher level performances. Multiple observables were defined for each task that captured essential aspects of the higher level performances. We think of these observables as "*feature scores*" for the task that captured characteristics of the examinee work products relevant to the proficiencies being measured. Each task observable was linked to appropriate proficiency nodes, so the observable evidence could be propagated appropriately through the network to update proficiencies.

We note that computer presentation of the rich tasks was arranged in a way that supported the evidentiary needs of the tasks, and the computer scoring of examinee work products was demonstrated. These aspects of the Biomass project will not be covered in this paper. Instead we focus on psychometric modeling issues addressed within the Biomass demonstration. Specifically the modeling of multiple observables within a single task, modeling dependency relationships among multiple observables, and modeling the interactions within observable nodes of multiple proficiency parents.

For purposes of Biomass it was necessary to model various kinds of interactions among proficiency parents within a given observable node. For example, Figure 4 depicts the evidence model for a single observable called Effective Methodology, with two proficiency parents Working Knowledge of Scientific Inquiry (WKInquiry), and Disciplinary Knowledge of Mendelian genetics (DKMendel). The convention used in this and following figures is that nodes with a circle left of the name are proficiencies, and nodes with a triangle are observables. A menu of choices was offered for developers to be able to select types of parent interactions within an observable node. The menu of node-parent interaction types included: (1) *Compensatory* Relationship—the two parents can fully compensate for one another as in multiple factor analysis or multidimensional IRT; (2) *Conjunctive* Relationship—high performance on the observable requires high on both parents; as used in the Fusion Model (Roussos,Templin, & Henson, this issue); (3) *Disjunctive* Relationship—this corresponds to logic OR gates; probability of a high score on the observable requires *at least one* of the parents to be high; and (4) *Inhibitor* Relationship—an inhibitor parent must be at or above some minimum given level, before the other skill can operate. Each of these interaction types can be modeled by placing certain constraints on the conditional probability tables at the observable node. Each of the types of interactions among multiple proficiency parents of *observable* variables can be applied equally well to types of interaction among multiple proficiency parents of *proficiency* nodes. The details are beyond the scope of this paper; see Almond et al. (2001).

The evidence model shown in Figure 4 is for a Biomass task that asks the student what a hypothetical researcher should do next, after having formalized an hypothesis about the mode of inheritance of coat color based on a field population. At the

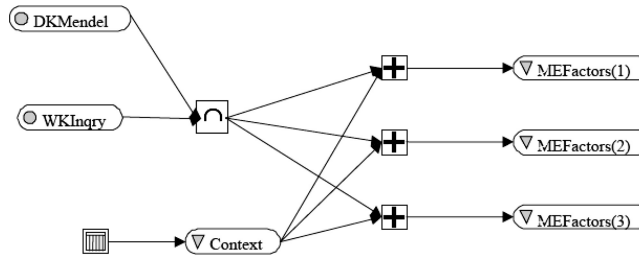

FIGURE 5. *Two parent proficiencies conjunctive plus common context.*

advice of domain experts the task was modeled with one observable and two parent proficiencies, with one proficiency acting as an *inhibitor*: the level of Disciplinary Knowledge of Mendelian genetics was required to be above Low to have any chance of a high quality response, and, when DKMendol was above low, the likelihood of a high quality response increased with WKInquiry level.

Figure 5 presents a more complex evidence model of a task with three different observables for one rich task. The observables for this task represented "feature scores" of a student's performance in filling out a table about implications of several statements about mode of inheritance. Three separate cases of inheritance data were given, and the student was asked whether a given statement can be confirmed or rejected based on data from various sources. Experts advised that each of the three task observables should be modeled to depend *conjunctively* on three proficiency parents: DKMendel, WKInquiry, and a common proficiency called Context effect. In the case of Figure 5, the three observables represented three independent actions the examinee had to take, and each was scored on a scale of 1–3. Each of the three actions was either taken or not, and when it was taken, it was of low, medium or high quality. Because this one task requires a substantial amount of time on one particular experimental setup, a *Context proficiency* was modeled to account for a common effect across all seven of these observables. The primary role of the context variable was to account for statistical dependencies among the observables stemming from the common experimental setup. The context variable had no other dependency connections with other variables, and was not intended for reporting. It functioned only while the student was working on this particular task. Once the examinee's observable evidence was propagated through the Bayesian network and the proficiency nodes updated, this context node was effectively thrown away before proceeding either to reporting assessment feedback or presenting a new task. (Almond, Mulder, Hemat & Yan, 2006 discuss other ways to model task observable dependencies.)

The model displayed in Figure 5 for the interactions within given observables is complex. The two DK and WK proficiencies are modeled as *conjunctive*, as noted above, and that conjunction is combined *compensatorily* with the Context proficiency. So the interaction structure here is a mixed or hybrid type.

One last example, illustrated in Figure 6, is a generic case of a multistep problem. There are three steps, and each can be scored separately, producing an observable for each step. Each of steps 2 and 3 depends on performance on the previous step. Consequently Observable 1 depends only on the required proficiency (which in general

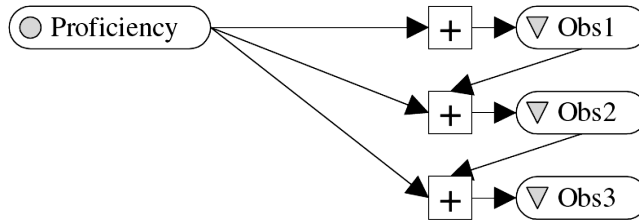

FIGURE 6. *Cascading model.*

could be multiple proficiencies). Observable 2 is modeled as a compensatory combination of the required proficiency plus good performance on Step 1, and Step 3 is modeled as a compensatory combination of the required proficiency and the result of Step 2. Note that in this example, Step 3 is independent of Step 1 given the required proficiency and the results of Step 2. Other cases might call for a different structure, and a different parent interaction model than compensatory—such as conjunctive or inhibitor.

In summary, these examples demonstrate evidence models involving different kinds of interactions between parents within an observable node, addition of a context effect to account for correlation that results from having a common setup for multiple observables, and examples of conjunctive, inhibitor and compensatory interaction modes.

Although significant amounts of research and development effort have gone into developing these modeling capabilities, and working with experts to elicit the necessary information, and with field data, much work remains to be done. It is not well understood how sensitive model performance is to various types of model variants; whether certain variations result in bias; when non-identifiability occurs and how to repair it, and other statistical questions. Preliminary results (Almond, Mulder et al., 2006) indicate that models of the "wrong type" can produce close approximations to the "true" probability table. More research is needed in the context of authentic educational settings, to study both the mathematical properties of the models and how easily practitioners can use them. The Biomass project represented useful research on both sides of this coin, and provided: (1) a rich set of examples of the powerful modeling flexibility available with Bayesian network models, and (2) a suggestive context of learning and assessment that potentially can derive great benefit from IRT based cognitive diagnostic assessment.

Scoring Engines and Adaptivity

Once an *ADG* graphical structure is fixed and model parameters estimated, ECD-based Bayesian network models can be used as scoring engines for making inferences about student proficiencies based on observable evidence. Each new set of evidence of student performance on a specific task, as values of corresponding observables, can be propagated through the model, updating the probability distributions of each proficiency variable in light of all the evidence incorporated so far. The updated local probability distribution at each proficiency node represents posterior conditional probabilities of each of the levels for a given proficiency, given the

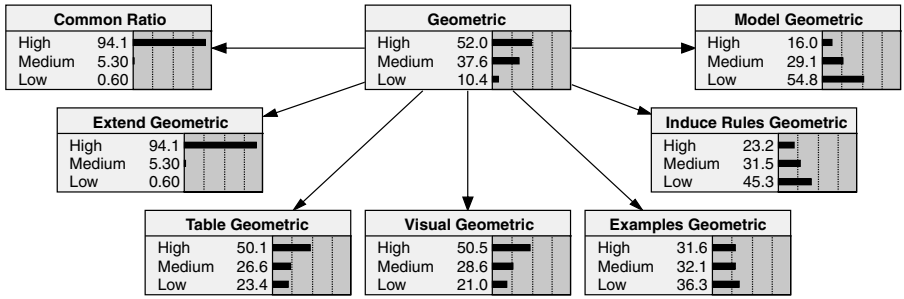

FIGURE 7. *ACED proficiency model before a new response evidence is read.*

evidence that has been propagated so far. The system then is ready to either report inferences formatted as a score report, or gather new evidence from new tasks and propagate as before (Almond, Shute, Underwood, & Zapata-Rivera, 2006).

The assumption of local independence of task observables across tasks given values of all proficiency variables is used to enable an efficient stepwise propagation algorithm that proceeds one task at a time (Almond & Mislevy, 1999). The same one-task-at-a-time procedure also can be employed as a "batch procedure" to score *all* tasks taken so far by an examinee. Testing constraints specific to the setting can indicate whether it is better to update proficiencies after each task or as a batch at the end.

We illustrate propagation with an example from the ACED (Adaptive Content for Evidence-Based Diagnosis) prototype assessment, a computer-based adaptive tool designed to provide evidence-based diagnosis to students in the domain of middle school mathematics on number sequences (Shute, Graf & Hansen, 2005; Shute, Hansen & Almond, 2007). The ACED Proficiency Model consisted of 42 nodes divided into three main branches: *arithmetic*, *geometric*, and *other recursive* sequences all under a main node called *sequences as patterns*. The lower level nodes then

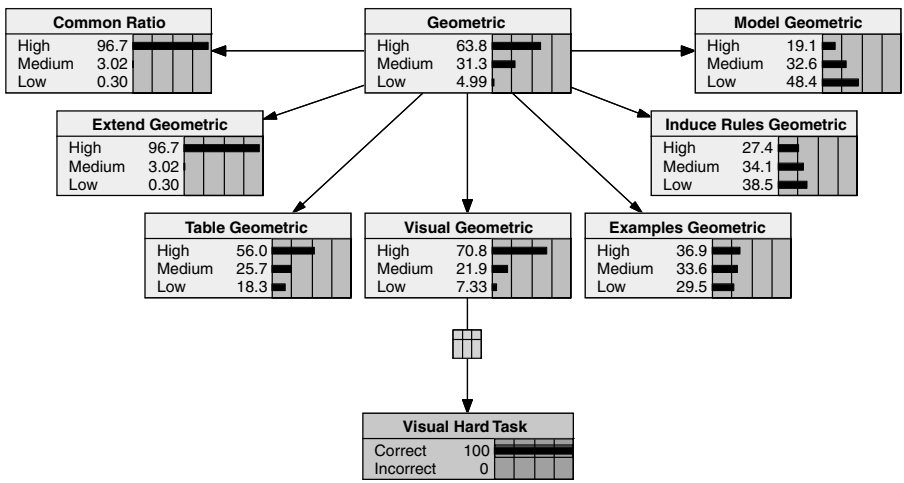

FIGURE 8. *ACED proficiency model after reading response to a difficult task.*

represented various ways sequences can be presented and manipulated. Figures 7 and 8 show a fragment of the *geometric* branch before and after gathering new evidence for a particular difficult task.

The one-task-at-a-time propagation and updating algorithm discussed above provides a procedure that can be used for incremental updating of a student's proficiency model in contexts such as an intelligent tutoring system (ITS) or computer adaptive test (CAT). In addition to incremental updating of the proficiencies, ITS and CAT systems require a *task selection mechanism*. A very general method has been developed for using expected weight of evidence in the context of Bayesian network models for task selection (see Madigan & Almond, 1995, for an approach called the *critiquing approach*, Miller, 1983; and Good, 1985, for expected weight of evidence).

ACED employed the following assessment-instruction cycle: (1) use an instructional heuristic to select an instructional area for testing, (2) identify all tasks suitable for the selected instructional area, (3) calculate the expected weight of evidence for each of the identified tasks, (4) select an identified task with highest expected weight of evidence, (5) administer the task, collect and score student response, (6) update the student's proficiency model, and (7) stop or return to the first step and iterate.

We end this discussion by noting that the ACED project reported high usefulness and validity of the ACED assessment. In particular, the Bayesian network overall score had high reliability (0.88) and provided incremental validity over the pretest score in predicting the posttest score. Further, in addition to valid measurement, ACED with the expected weight of evidence algorithm produced an increase in student learning (Shute, Hansen, & Almond, 2007).

Conclusions

We begin this section with a reflection on strengths and weaknesses of Bayesian network modeling, particularly for IRT based cognitive diagnostic assessment. As we have seen in the examples presented above, the Bayesian network modeling framework lends itself naturally and directly to item response modeling, with a portion of the network consisting of what we have called the proficiency model, with proficiency variables and direct links between them, and another portion called the evidence model that consists of one or more observable variables for each task, along with links from particular proficiency nodes to each observable.

As noted in the Introduction, a requisite for valid, high quality and effective assessment is harmony between the substantive theory that underlies the conceptual student model and the formal probability model supporting the assessment. Conceptual student models can take many forms, and excessive complexity quickly outstrips model tractability, under any probability modeling framework. Designing successful IRT-based cognitive diagnostic models requires systematic attention—part science, part art, part engineering craft—to finding a good balance of complexity that reflects just enough underlying substantive reality to ground the assessment's validity, and retains a practical capability to calibrate the models and generate inferences about examinees from data.

The Bayesian network modeling approach provides a powerful modeling flexibility, together with practical feasibility, including ready availability of theory, methods

and software for building and calibrating models and using them for scoring. That flexibility enables Bayesian network modeling to support a wide range of assessments, particularly the innovative and more complex assessments produced with the Evidence Centered Design (ECD) approach. Assessments discussed in this paper for example involved multidimensional proficiency models that included links between proficiencies, simulated performance environments, rich task types with multiple observables, and explicit dependency links among multiple observables within the same task. Concrete examples were discussed of different types of interactions of parent nodes within proficiency or observable nodes, including compensatory, conjunctive, disjunctive, and inhibitor types.

The modeling flexibility of Bayesian network models is especially pertinent to the design of assessments that attempt to measure deeper understanding than simple declarative recall. Bayesian networks can support rich tasks that are designed to gather multiple kinds of evidence necessary for deeper measurement.

The same significant advantages of broad Bayesian network modeling flexibility also entail a certain cost. The *capability* to do complex modeling implies the *necessity* to address complex modeling issues for both proficiency and evidence models. Depending on the setting, the building of proficiency and evidence models may require considerable effort. Substantive knowledge and theory can contribute to various aspects of design, including the dependency link structure and the selection of types of interactions within a node with multiple parents. But this requires that critical information be elicited from experts about model structure issues as well as prior probabilities. Quite a bit is known about the difficulties of eliciting accurate and consistent information from experts.

Issues of statistical identifiability are critical for these models, and can even cause problems in parameter calibration with a fixed link structure, as described for the ICT example above. Identifiability issues are more related to model assumptions than to the modeling approach, and must be addressed with any approach to complex modeling. As noted in the ICT example above, MCMC Bayesian network calibration, by generating an approximate posterior distribution, provides a strong capability for analyzing and understanding various types of nonidentifiability and determining minimal constraints that will repair the statistical identifiability without damaging the model's ability to capture salient aspects of the underlying cognitive reality.

As noted earlier, the most common procedure in assessment applications for learning Bayesian network models from data is to fix model structure and prior distributions based on expert information and to learn the conditional probability tables from data. Learning Bayesian network *structure* from data is an area of active research (Heckerman, 1998, and Neapolitan, 2004 provide overviews). Most existing literature assumes all variables are observed, whereas assessment applications invariably include latent proficiency variables. Learning structure for Bayesian networks with so-called "hidden nodes" is much more difficult (Eaton & Murphy, 2007; Elidan & Friedman, 2005). Input from domain experts can help constrain model search algorithms.

Complex models inevitably require heavier computing than simpler models, but details depend on the approach. MCMC methods have the advantages of allowing model changes to be implemented very quickly and produce an estimated full

posterior distribution rather than point estimates such as posterior mean or mode, but MCMC calibration is CPU intensive. CPU time is linear in the total number of parameters and also in the size of calibration sample, a possible conflict with the dependency of high calibration accuracy on larger sizes of calibration samples. Bayesian network parameter learning also can be accomplished successfully with Expectation and Maximization (EM) methods, as is done in commercial products such as HUGIN and Netica. EM methods in general are much less CPU intensive, but require a good deal of "set-up" work when a new model is built or adapted. Stringent steps are necessary to avoid "converging" to a local maximum.

As with any fully Bayesian approach, it is instructive to compare prior and posterior distributions. If priors are "mildly informative" and the posteriors are not much different from the priors, then we suspect the model is not capturing much information from data. If the priors are strongly informative and posteriors are significantly different from the priors, then we may suspect that the strong priors may not be appropriate. Coordinated interdisciplinary teams can design graphical structures and priors that represent the very best existing knowledge and theory. Comparing the posteriors and priors, as well as computing measures of model-data fit that are strongly related to the measurement purpose, simultaneously informs psychometrics as well as substantive areas with indicators of whether and to what degree substantive theories are confirmed or disconfirmed by these models and data.

In summary, this paper takes as a point of departure the argument that cognitive diagnostic assessment, considered from the point of view of evidentiary systems, demands a degree of alignment among multiple assessment components: finer grained, more ramified models of student cognition; richer tasks that afford opportunities to gather evidence of deeper knowledge and understandings; and psychometric models used as inference engines. Bayesian network modeling provides an essential flexibility for building applied diagnostic assessments that can be implemented and evaluated. The psychometric modeling capability currently provided by Bayesian network models as described in this paper—including supporting theory, available software, developed methods and procedures—can and should be applied now to build real and experimental IRT-based cognitive diagnostic assessment systems that support teaching and learning.

Note

[1]A different method was chosen to score the operation assessment, now named iSkills™.

References

Almond, R. G. (2007). "I can name that Bayesian network in two matrixes!" In K. B. Laskey, S. M. Mahoney, & J. A. Goldsmith (Eds.), *Proceedings of the 5th UAI Bayesian Application Workshop. CEUR Workshop Proceeding*. Available online at http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol. 268/.

Almond, R., DiBello, L., Jenkins, F., Mislevy, R., Senturk, D., Steinberg, L., & Yan, D. (2001). Models for conditional probability tables in educational assessment. In T. Jaakkola & T. Richardson (Eds.), *Artificial intelligence and statistics 2001 (*pp. 137–143). Morgan Kaufmann (http://www.mkp.com/).

Almond, R. G., & Mislevy, R. J. (1999). Graphical models and computerized adaptive testing. *Applied Psychological Measurement, 23,* 223–238.

Almond, R. G., Mulder, J., Hemat, L. A., & Yan, D. (2006). *Models for local dependence among observable outcome variables*. Technical report RR-06-36, Educational Testing Service. Available at: http://www.ets.org/research/researcher/RR-06-36.html

Almond, R. G., Shute, V. J., Underwood, J. S., & Zapata-Rivera, J. D. (2006). Bayesian networks: A teacher's view. Paper presented at the 4th Bayesian Modeling Application Workshop at the Conference on Uncertainty in Artificial Intelligence, Cambridge, MA.

Almond, R. G., Yan, D., & Hemat, L. A. (2005). Simulation studies with a four proficiency Bayesian network model (Draft), Draft Technical report, Version 1.12, Nov 11, 2005, Educational Testing Service. Available from ralmond@ets.org.

Eaton, D., & Murphy, K. (2007). Bayesian structure learning using dynamic programming and MCMC. In R. Parr & L. van der Gaag (Eds.), *Uncertainty in artificial intelligence: Proceedings of the twenty-third conference* (pp. 101–108). AUAI Press (http://www.quai.org/).

Elidan, G., & Friedman, N. (2005). Learning hidden variable networks: The information bottleneck approach. *The Journal of Machine Learning Research Archive, 6,* 81–127.

Gierl, M. J. Making diagnostic inferences about cognitive attributes using the rule-space model and attribute hierarchy method. *Journal of Educational Measurement*, this issue, 325–340.

Good, I. J. (1985). Weight of evidence: A brief survey. In J. Bernardo, M. DeGroot, D. Lindley, & A. Smith (Eds.), *Bayesian statistics 2* (pp. 249–269). North Holland (http://www.elsevier.com/).

Heckerman, D. (1998). A tutorial on learning with Bayesian networks. In M. I. Jordan (Eds.), *Learning in graphical models* (pp. 301–354). Kluwer Academic Publishers (http://www.springer.com/).

Jensen, F. (1996). *An introduction to Bayesian networks*. Springer-Verlag (http://www.springer.com/).

Katz, I. R., Williamson, D. M., Nadelman, H. L., Kirsch, I., Almond, R. G., Cooper, P. L., Redman, M. L., & Zapata, D. (2004). Assessing information and communications technology literacy for higher education. Paper presented at the 30th annual conference of the International Association for Educational Assessment, Philadelphia, PA.

Madigan, D., & Almond, R. (1995). Test selection strategies for belief networks. In D. Fisher & H. J. Lenz (Eds.), *Learning from data: AI and statistics V* (pp. 89–98). Springer-Verlag (http://www.springer.com/).

Miller, P. (1983). ATTENDING: Critiquing a physician's management plan. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 5,* 449–461.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessment (with discussion). *Measurement: Interdisciplinary research and perspective, 1*(1), 3–62.

National Research Council (2001). *Knowing what students know: The science and design of educational assessment*. Committee on the Foundations of Assessment. In J. Pellegrino, N. Chudowsky, & R. Glaser (Eds.), Board on Testing and Assessment, Center for Education. Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

Neapolitan, R. E. (2004). *Learning Bayesian networks*. Prentice Hall (http://www.prentice-hall.com/).

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

Roussos, L. A., Templin, J. L., & Henson, R. A. Skills diagnosis using IRT-based latent class models. *Journal of Educational Measurement*, this issue, 293–311.

Shute, V. J., Graf, E. A., & Hansen, E. G. (2005). Designing adaptive, diagnostic math assessments for individuals with and without visual disabilities. In L. M. Pytlikzillig, R. H. Bruning, & M. Bodvarsson (Eds.), *Technology-based education: Bringing researchers and practitioners together* (pp. 169–202). Greenwich, CT: Information Age Publishing.

Shute, V. J., Hansen, E. G., & Almond, R. G. (2007). An assessment for learning system called ACED: Designing for learning effectiveness and accessibility. ETS Research Report RR-07-27. Princeton, NJ: Educational Testing Service.

Steinberg, L. S., Almond, R. G., Baird, A. B., Cahallan, C., Chernick, H., DiBello, L. V., Kindfield, A. C. H., Mislevy, R. J., Senturk, D., & Yan, D. (2003). Introduction to the biomass project: An illustration of evidence-centered assessment design and delivery capability. Research Report 609, National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Available at: http://www.cse.ucla.edu/reports/R609.pdf.

Yan, D., Almond, R., & Hemat, L. (2005). *Bayesian network model for the ICT Literacy assessment*. Draft Technical report, Version 1.9 dated Dec 16, 2005. Educational Testing Service. Available from ralmond@ets.org.

Authors

RUSSELL G. ALMOND is a Senior. Research Scientist, M.S. 13-E, Educational Testing Service, Princeton, NJ 08541; ralmond@ets.org. His primary research interests include Bayesian networks, artificial intelligence, Bayesian statistics, knowledge and data engineering, and statistical software.

LOUIS V. DIBELLO is a Research Professor, Learning Sciences Research Institute, Mail Code 057, University of Illinois at Chicago, 1007 West Harrison Street, Chicago, IL 60607-7137; ldibello@uic.edu. His primary research interests include psychometrics, diagnostic assessment, and informative assessment.

BRAD MOULDER is a Psychometric Manager, MS. 06-P, Educational Testing Service, Princeton NJ 08541; bmouldor@ets.org. His primary research interests include psychometrics and diagnostic assessment.

JUAN-DIEGO ZAPATA-RIVERA is a Research Scientist, M.S. 13-E, Educational Testing Service, Princeton, NJ 08541; dzapata@ets.org. His primary research interests include Bayesian student modeling, inspectable Bayesian student models, assessment-based learning environments (ABLE), external representations, and virtual communities.