Check for updates

Article

Applied Psychological Measurement
2016, Vol. 40(1) 3–21
© The Author(s) 2015
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621615590401
apm.sagepub.com
SAGE

Bayesian Networks in Educational Assessment: The State of the Field

Michael J. Culbertson[1]

Abstract

Bayesian networks (BN) provide a convenient and intuitive framework for specifying complex joint probability distributions and are thus well suited for modeling content domains of educational assessments at a diagnostic level. BN have been used extensively in the artificial intelligence community as student models for intelligent tutoring systems (ITS) but have received less attention among psychometricians. This critical review outlines the existing research on BN in educational assessment, providing an introduction to the ITS literature for the psychometric community, and points out several promising research paths. The online appendix lists 40 assessment systems that serve as empirical examples of the use of BN for educational assessment in a variety of domains.

Keywords

Bayesian networks, graphical models, diagnostic testing, MIRT, computerized adaptive testing, intelligent tutoring systems

Introduction

Teachers are increasingly asked to cover more material in less time. Such instructional efficiency requires providing students with instructional material that is relevant to the educational content they have not yet mastered and not providing instructional material that is redundant with content they know. While teachers currently have some systems in place to gauge what their students know, they could potentially benefit in time and energy savings from detailed, relevant information provided by new diagnostic assessment tools (e.g., Almond, Shute, Underwood, & Zapata-Rivera, 2009). However, most of today's educational assessments, and particularly those already in place as a part of accountability systems, provide only high-level information about broad educational domains (such as mathematics, reading, science, etc.). If measurement models are going to be useful to teachers for rapid, targeted educational response, they will need to provide efficient analysis of student achievement on a variety of sub-domains closely aligned with their instructional context.

Several classes of models are common choices for providing sub-domain ability estimation, including Multi-Dimensional Item Response Theory (MIRT; Reckase, 2009) and diagnostic classification models (DCMs; for example, Rupp, Templin, & Henson, 2010). Most of these

[1]University of Illinois at Urbana–Champaign, USA

Corresponding Author:
Michael J. Culbertson, University of Illinois at Urbana–Champaign, 1310 S. Sixth St., Champaign, IL 61820, USA.
Email: culbert1@illinois.edu

models leave the structure of the underlying latent abilities unspecified (a notable exception being the Attribute Hierarchy Method; Leighton, Gierl, & Hunka, 2004), and it may be possible to achieve greater measurement precision or better model parsimony by modeling sub-domain relationships explicitly. Bayesian networks (BN) serve as one means for modeling these relationships. Although BN have received only modest attention in the psychometric community, they have been used extensively in the artificial intelligence community as student models for intelligent tutoring systems (ITS; for overviews from this perspective, see Conati, 2010; Desmarais & Baker, 2012; Jameson, 1995). Given their success in automated tutoring, BN have the potential to improve the complexity and sophistication of the measurement models available to human educators, as well.

The purpose of this article is to provide a thorough review of the literature on the use of BN in educational assessment. Here, the goal is not to provide an introduction to the use of BN, but rather to provide a quick reference or point of departure for accessing the relevant literature on methods, issues, and challenges of using BN specifically in educational assessment. The review has been organized as follows: Following a brief overview of BN, literature pertinent to the central process of graph development is presented. This includes a novel taxonomy for educational BN based on the level of detail of the representation of the given knowledge domain, as well as a discussion of hierarchical networks that accommodate multiple levels of granularity in this representation. This section also covers issues surrounding learning network structure, dynamic networks, and model criticism. Next, the review examines applications of BN to computerized adaptive testing (CAT), specifically in terms of item selection. Finally, current challenges and open research directions to the expanded use of BN in educational assessment are presented. Examples of existing BN-based assessments are presented throughout, and most of the currently published BN-based assessments can be found in the online appendix.

Bayesian Networks

BN (Pearl, 1988) provide a convenient and intuitive way to specify complex joint probability distributions, with both observed and latent variables. BN are not a type of model, per se; rather, BN represent a framework or approach to model building that capitalizes on relationships between a graphical representation and a complex joint probability distribution. A BN consists of an acyclic directed graph and a corresponding set of conditional probability distributions. In the graph, each node (i.e., a point in the graph) represents a variable, and the edges connecting the nodes represent relationships between variables (Figure 1). Pairs of connected nodes are referred to as parents and children, with directed edges flowing from parents to children. A conditional probability distribution is specified for each node given its parents.

The convenience of a BN comes from the conditional independencies implied by the graph and the corresponding simplification to the general multiplication rule. Each variable is conditionally independent of all other variables, given the variables that surround it (Almond, DiBello, et al., 2007; Pearl, 1988). The general multiplication rule gives a joint probability distribution as the product of successive conditional probability distributions, for example,

$$\begin{aligned} P(A,B,C,X_1,X_2,X_3,X_4) &= P(A)P(B|A)P(C|A,B)P(X_1|A,B,C)P(X_2|A,B,C,X_1) \\ &\times P(X_3|A,B,C,X_1,X_2)P(X_4|A,B,C,X_1,X_2,X_3). \end{aligned} \quad (1)$$

Although this recursive representation can be written for any order of variables, the order that matches the conditional probability distributions associated with the nodes of the BN permits considerable simplification due to the conditional independencies implied by the graph:

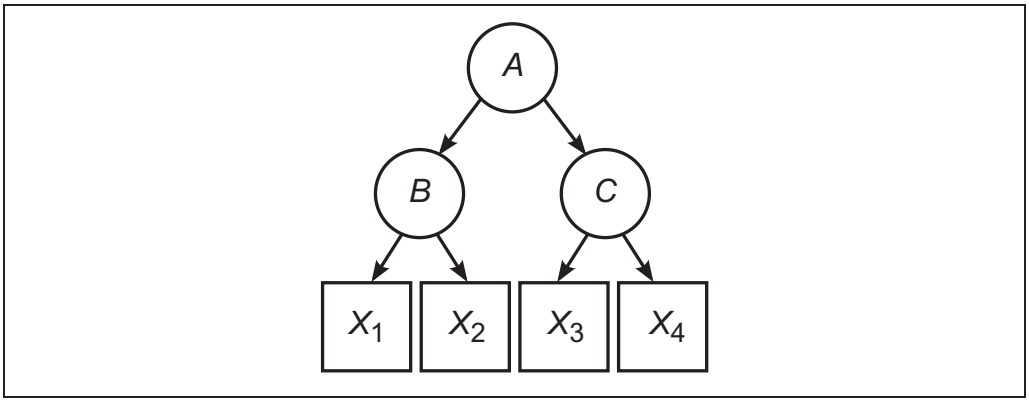

Figure 1. Sample Bayesian network.
Note. Circles represent latent variables, and boxes represent observed variables. Arrows represent conditional dependencies. Variables at the tail of an arrow are "parents," and variables at the head of the arrow are the "children." For example, *B* is the parent of X_1 and the child of *A*.

$$P(A,B,C,X_1,X_2,X_3,X_4) = P(A)P(B|A)P(C|A)P(X_1|B)P(X_2|B)P(X_3|C)P(X_4|C). \quad (2)$$

Moreover, marginal distributions can be specified simply by recursively taking the product of a node's conditional distribution and its parents' conditional distributions, for example,

$$P(X_3) = P(X_3|C)P(C|A)P(A). \quad (3)$$

Thus, complex relationships between many variables can be described through conditional relationships between much smaller subsets of variables, which may be read directly from the graph.

Although BN are not intrinsically linked with Bayesian estimation theory, their heavy use of conditional probability distributions often leads to efficient Bayesian estimation through application of Bayes Theorem, along with specialized estimation algorithms that capitalize on the graphical structure. When observed variables are specified as leaves (i.e., nodes with no children) and parents are taken to be prior distributions, Bayesian inference on latent variables and parameters can be accomplished by reversing the edge directions and applying Bayes Theorem. Some take the direction of edges in graphical models to represent ontological beliefs about the nature of latent variables (Anderson & Yu, 2007; Borsboom, Mellenbergh, & van Heerden, 2003): Realists draw arrows from latent variables to observed variables because they claim that latent variables exist in the world in a real sense and generate observations. Constructivists, however, draw arrows from observed variables to latent variables because they claim that latent variables are only constructed abstractions from observations. In a BN, however, arrows represent only the conditional specification of the probability distribution for each node, not an ontological claim about the nature of latent variables. Thus, reversing edge directions constitutes only a rearrangement of the parametric specification of the joint probability distribution through an application of Bayes Theorem, not a philosophical flip-flop.

In applications of BN to educational measurement, hidden nodes (i.e., latent variables) represent cognitive features of the content domain, and leaves generally represent observed variables (though observed variables may also be connected to one another to encode dependencies unrelated to the latent constructs, such as multi-part problems). Observed variables may include traditional test items (e.g., multiple choice) or other measurements based on examinee

performance, such as the output of acoustic processing routines in an oral word decoding task (e.g., Tepperman, Lee, & Alwan, 2011), interactions with a computer-based scientific inquiry environment (e.g., Ting & Phon-Amnuaisuk, 2012), or actions in instructionally relevant computer games (e.g., Shute, 2005). Items may be connected to one or more hidden nodes, similar to the Q-matrix in DCM (Almond, 2010; Tatsuoka, 1983). Edges connecting hidden nodes encode associations between cognitive states, knowledge, or abilities.

Although BN share many similarities with other graphical modeling techniques, such as structural equation models, the BN approach affords considerable flexibility in specifying joint probability distributions as a BN may use any probability distribution as a node's conditional distribution, and conditional distributions for different nodes are not required to belong to the same family. For educational assessments, the conditional distributions specified for observed variables will usually be familiar item response theory (IRT) models or DCMs. As such, traditional IRT and DCM can be viewed as special cases of BN with a single (potentially multidimensional) hidden node (Mislevy, Almond, Yan, & Steinberg, 2000). By explicitly modeling the relationships between latent variables, BN provide an additional advantage over unidimensional models, allowing prediction of students' achievement in un-assessed sub-domains based on results from related sub-domains that have already been assessed. Although this is generally possible with many multi-dimensional techniques, applications of MIRT and DCMs typically leave the structure of the relationships between dimensions unspecified. BN, on the contrary, allow for greater parsimony by constraining these relationships through an intuitive graphical modeling approach, reducing the number of parameters that must be estimated at the cost of potential model mis-specification. Moreover, BN readily lend themselves to modular model building in which fragments of BN may be developed separately and combined freely based on a wide variety of types of relationships, permitting flexible adaptation of psychometric models to particular assessment scenarios. This aspect of BN becomes particularly powerful when a well-defined theory can be used to guide the combination of model fragments, as illustrated by the Andes ITS below.

Due to ease of computation, most applications of BN have used discrete latent variables. However, there is no theoretical restriction to discrete variables; and advances in computing power and new computational algorithms, such as Markov chain Monte Carlo (MCMC; for example, Patz & Junker, 1999) and Metropolis–Hastings Robins–Monro (MH-RM; for example, Cai, 2010), render BN with continuous latent variables more feasible. Structural equation models, for example, can be expressed as a BN with normally distributed continuous latent variables.

Graph Development

The crux of a BN is the graph that specifies the conditional relationships between variables. Obtaining this graph is a non-trivial process that involves a decomposition of the target knowledge domain and specifying a priori or learning empirically the structural relationships between variables.

Decomposition of the Knowledge Domain

The first step toward developing the graph for a BN is to enumerate the latent variables and their corresponding psychometric constructs. This involves, in essence, a decomposition of the target knowledge or skill domain into constituent parts. Ontologies may be helpful in decomposing the knowledge domain (e.g., Chung, Niemi, & Bewley, 2003; Koenig, Lee, Iseli, & Wainess, 2010). In addition to ''correct'' knowledge, the constituent parts may include misconceptions or ''bugs''

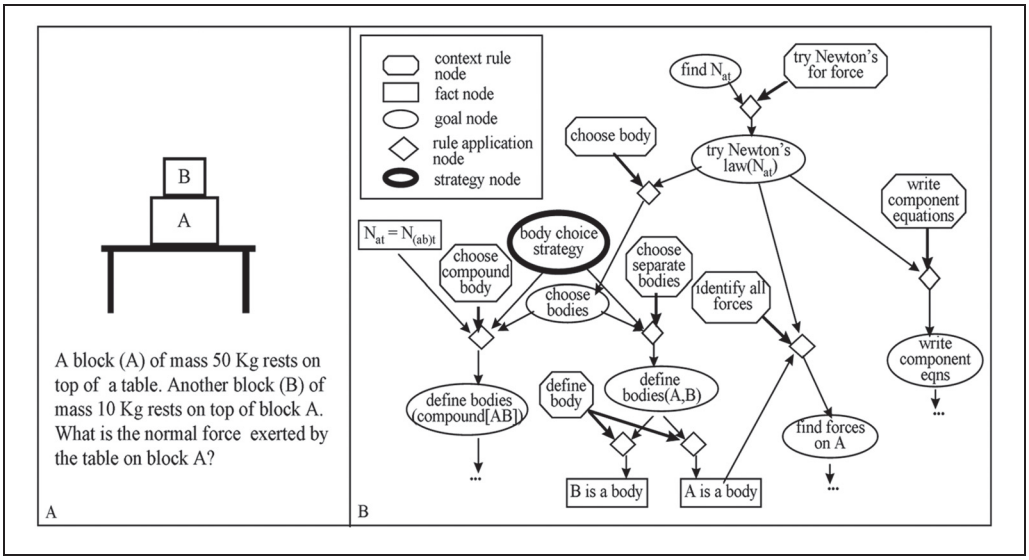

Figure 2. Fragment of the Andes Bayesian network for a sample problem.
Source. Reprinted from Conati, Gertner, VanLehn, and Druzdzel (1997, Figure 1) with kind permission from Springer Science+Business Media.

(e.g., Goguadze, Sosnovsky, Isotani, & McLaren, 2011). BN for student knowledge can be classified roughly by the grain size (level of detail) of the constituent parts.

Low-level networks. When a target domain is limited and very well defined, such as mathematics or physics problem solving, modelers may elect to use a fine-grain, low-level network to obtain highly detailed information about students' cognitive processes. In these networks, the constituent parts of the domain are taken to be singular properties, relationships, rules, or misconceptions of the domain or of the specific problem context, as well as particular problem-solving steps that combine these beliefs to create new beliefs through logical deduction. For example, the Andes ITS (Conati, Gertner, & VanLehn, 2002; Conati, Gertner, VanLehn, & Druzdzel, 1997) decomposes Newtonian physics problem solving into rule nodes, context nodes, strategy nodes, and goal nodes (Figure 2). Rule nodes describe all of the mathematical relationships in Newtonian physics; context nodes indicate mathematical propositions about the given problem; and goal nodes represent physical quantities the student is working toward. Domain-based rules and problem-based context propositions are combined with ''rule application'' nodes as parents for new beliefs about the problem. In these low-level models, the problem graph is constructed by an automated problem solver that applies a set of production rules representing all mathematical relationships in physics to the initial set of given conditions, enumerating all of the possible (logically valid) solution paths for the problem. This results in a large and incredibly complex network with the power to diagnose exactly where in the problem solving the student is likely to be at any point in time, which allows the ITS to detect when students are stuck and offer an appropriate hint (Gertner, Conati, & VanLehn, 1998). Andes illustrates modular model building with BN by the principled combination of theory-driven model fragments, which can be expanded dynamically as students interact with the system.

When domains consist of well-defined facts and a relatively small set of deterministic procedures, low-level networks such as those in Andes can provide highly detailed information about students' cognitive processes during problem-solving exercises due to the highly specific and

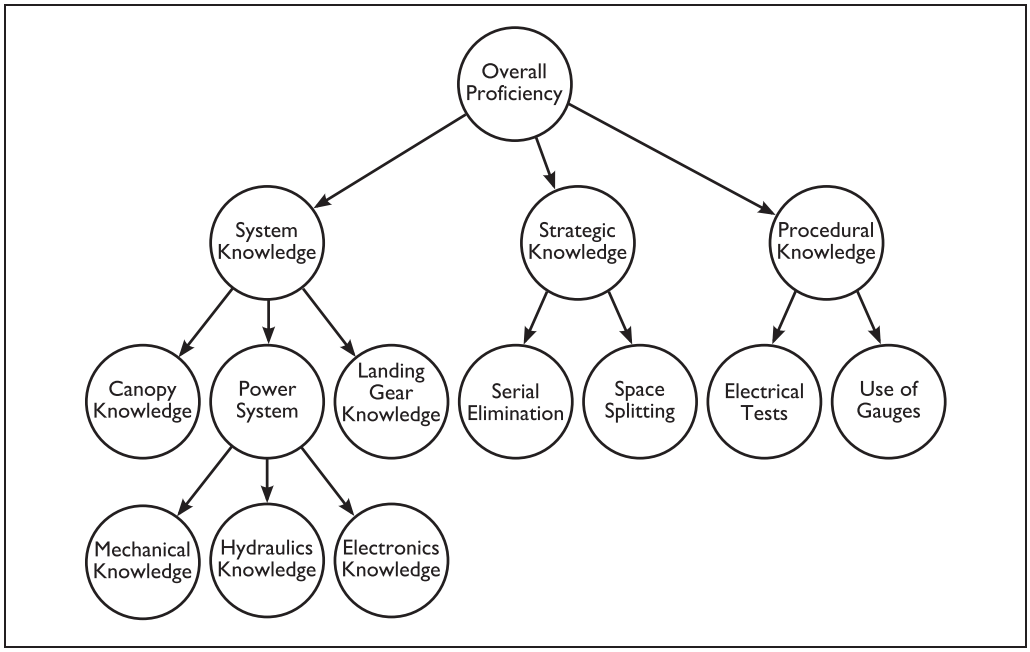

Figure 3. Network for HYDRIVE.
Source. Reprinted from Mislevy and Gitomer (1996, Figure 4) with kind permission from Springer Science+Business Media B.V.

narrow definition of individual nodes, allowing instructors to pinpoint specific knowledge items students have not mastered. Although this level of detail may be overwhelming for routine use by instructors, an ITS can leverage the diagnostic power of the model to provide timely and contextually relevant help.

Mid-level networks. A step removed from individual cognitive components, mid-level networks decompose the target domain into sub-skills, which may be related through learning progressions (one concept is prerequisite for learning another) or requirements for individual items (several concepts are necessary for completing the given item). For example, Mislevy (1994) and Vomlel (2004) have both used mid-level networks to study tests of fraction arithmetic. Their models include such sub-skills as addition with a common denominator, multiplication, finding a common denominator, simplification, and conversion between mixed numbers and improper fractions. Unlike the low-level Andes model, these graphs do not track the specific cognitive paths used by students during exercises; rather, they facilitate inference on a student's overall mastery of given sub-skills, and may even determine which strategy a student is using (as in Mislevy, 1994).

Although mid-level networks tend to have intermediate complexity in terms of number and interconnectedness of nodes, they may vary in terms of exactly how specifically their sub-skills are defined. The sub-skills above represent fairly specific and well-defined components of the fraction arithmetic domain; however, in HYDRIVE (Mislevy & Gitomer, 1996), an ITS for troubleshooting aircraft mechanical systems, the target domain is divided into relatively broad types of system knowledge, knowledge of fault-identification strategies, and kinds of procedural knowledge (Figure 3). Here, the variables do not generally represent specific cognitive procedures or facts, but small collections of knowledge or skill.

The distinction between mid-level networks and either high-level or low-level networks can be blurred in some cases. For example, the knowledge model for I-PETER (Read, Bárcena,

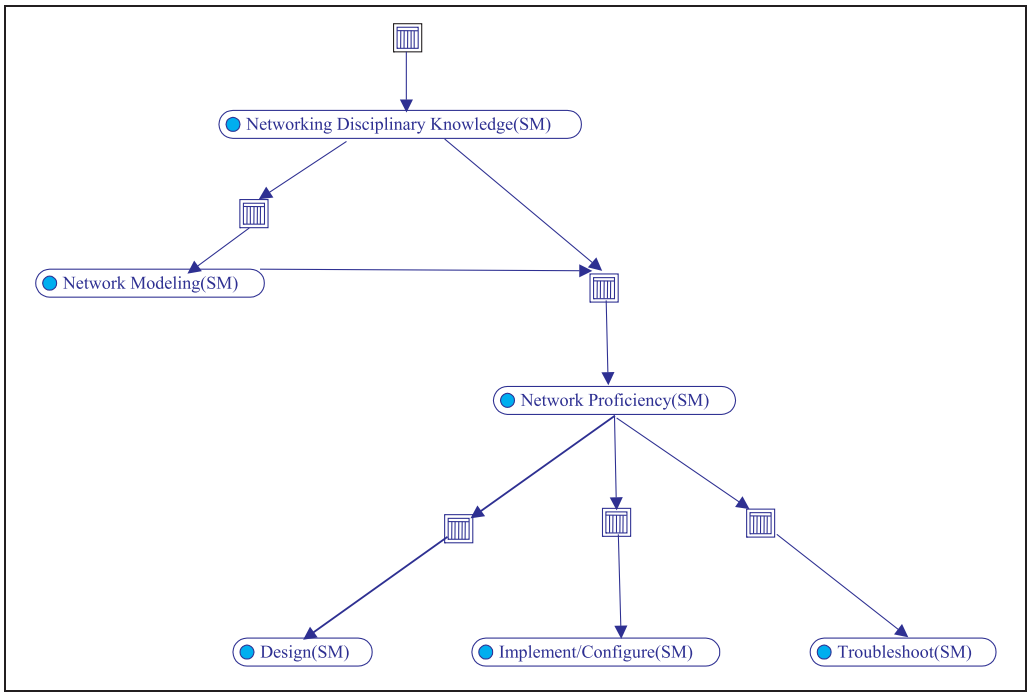

Figure 4. Network for NetPASS.
Source. Adapted from Levy and Mislevy (2004, Figure 1), reprinted by permission of Taylor & Francis Ltd., http://www.tandf.co.uk/journals
Note. SM = student model.

Barros, & Verdejo, 2002) decomposes English language knowledge along three facets: knowledge state (beginner to advanced), learning phase (mechanical reproduction vs. non-attentive application of language rules and patterns), and linguistic level (lexical, grammatical, or textual). An intermediate node is then created for each cell of the three-way table covering the three language-domain facets, leading to a proliferation of nodes seen more often in low-level networks, but with skill definitions more like those seen in high-level networks.

High-level networks. High-level networks usually have a small number of broadly defined sub-constructs that coarsely partition the knowledge domain. For example, NetPASS (Levy & Mislevy, 2004) decomposes the computer networking domain into variables such as network modeling, design, implementation, and troubleshooting (Figure 4). Similarly, the Information and Communication Technology Literacy Assessment (Almond, Yan, & Hemat, 2008) includes only four broad sub-constructs: Create/Communicate, Define/Access, Evaluate, and Manage/Integrate. Traditional IRT-based assessments can be framed as high-level BN, and high-level networks may be particularly useful when test designers already have a coarse partition of the content domain (say, for content balancing purposes).

Networks based on instructional sequences. When an assessment is closely linked with a particular course, the course's instructional sequence may be used to generate the domain decomposition. For example, Butz, Hua, and Maguire (2006) developed a mid-level network on computer programming for the ITS BITS with 29 nodes based on concepts covered in the course textbook. Similarly, Levy and Crawford (2009) simply took the nine chapters of the Networking

for Home and Small Business course as the nodes in a high-level network for a course assessment. Basing an assessment's graph on an instructional sequence, such as textbook units, facilitates rapid development of the knowledge model and resembles the structure of traditional classroom assessments, which are familiar to instructors. Moreover, instructional sequences can often suggest an initial graphical structure between concept nodes, either in terms of prerequisite relationships or simply the order in which students are exposed to concepts. If a formal learning progression theory is available, the learning progression can also serve as a basis for a network structure (e.g., West et al., 2012).

Granular Hierarchies

The link between the decomposed knowledge domain and instructional sequences suggests another source for elements of the BN. Often, instructional material is divided into units and sub-units, which form a hierarchy of knowledge aggregation. These hierarchies may be as simple as a subject–topic–concept organization of course content (e.g., Millán, Pérez-de-la Cruz, & Suárez, 2000) or may have five or more layers between ''big ideas'' and fine-grained skills (e.g., Patterson, 2011). Incorporating these hierarchies into the knowledge model may improve the utility of an educational assessment by facilitating different kinds of decision making (Martin & VanLehn, 1995). For example, an instructor may need detailed information about student mastery of particular skills to decide whether to re-teach material or assign follow-up exercises, while school-level administrators may be interested in topic- or course-level achievement to decide whether to assign students to supplementary educational services. Currently, most educational assessments are geared toward providing information optimal for only a single level, such as uni-dimensional IRT-based assessments for accountability purposes or assessments with skills-based DCMs for informing instructional decisions. With traditional narrowly focused assessments, the testing burden on students increases as more assessments are necessary to support data-based decision making at multiple levels. Assessments with knowledge models that integrate granular hierarchies, however, have the potential to provide both fine-grain information for immediate instructional decision making and rough-grain information for higher level decision making with the same assessment (Martin & VanLehn, 1995).

Moreover, models with granular hierarchies can facilitate statistically based content balancing (Collins, Greer, & Huang, 1996). Often in uni-dimensional or low-dimensional assessments, test designers impose non-statistical constraints to ensure that items from across the domain are included for construct validity. An assessment that incorporates a hierarchical knowledge model may be able to reduce the number of non-statistical constraints on item selection by optimizing the information provided on each of the fine-grain abilities. In some cases, this could even provide a stronger result, as selected items would provide strong information for their respective sub-domain content areas; whereas in non-statistical content balancing, there is generally no guarantee that each content area will be equally *informative*—only that each content area is represented by an adequate *number* of items.

The HYDRIVE knowledge model (Mislevy & Gitomer, 1996) provides an example of a granular hierarchy. The individual system or procedural skills are aggregated into three general abilities—system knowledge, strategic knowledge, and procedural knowledge—which are again aggregated for an overall proficiency. Higher order latent trait models (de la Torre & Douglas, 2004) can also be viewed as a BN with a granular hierarchy. Although aggregation relationships are not uncommon in psychometric BN in practice, little is known about the optimal number of aggregated levels or the precision of estimates of aggregated abilities. In a parameter recovery study for the Information and Communication Technology Literacy Assessment, Almond et al. (2008) determined that the global proficiency variable did not have adequate precision for

reporting purposes. Limited understanding of the precision of estimates from BN with granular hierarchies presents a current barrier to the use of BN for summative, particularly high-stakes, assessment and may restrict the utility of BN-based assessments to more-formative settings that rely only on low-level information from the hierarchy. More practical investigation into parameter recovery and ability estimate reliability is needed for a better understanding of the potential and limitations of granular hierarchies for supporting multiple levels of decision making.

Learning Structure

Once the target domain has been decomposed into the desired constituents, the next step in constructing a BN is to specify or learn the structural relationships between the given latent variables. There is currently no standard of best practice for specifying these relationships, and several possibilities exist. Even if a hierarchical aggregation structure is imposed on the individual skill variables, other relationships may be included between variables at any level in the hierarchy (Collins et al., 1996).

Types of node relationships. Perhaps the simplest and most common relationship between latent abilities is ''prerequisite of,'' though edges in the graph could also indicate ''part of'' (as in aggregation relationships), ''is correlated with,'' ''induces,'' or ''inhibits,'' among others (Almond et al., 2007). These semantic relationships bear little statistical significance, but may be helpful either in constructing or interpreting the graph. Generally aggregating latent variables are not directly observed (all child nodes are also latent variables), while latent variables in other kinds of relationships will have observed child nodes. When a node has more than one parent, the relationship between the parents must be specified in the node's conditional distribution. Possible relationships include compensatory, conjunctive, disjunctive, and inhibitor/enabler (Almond et al., 2007; Mislevy et al., 2002). A *compensatory* relationship adds the effects of the parent abilities—if one ability is low, the other ability compensates for missing skills. Compensatory relationships are particularly appropriate when there are multiple means of achieving the given skill and possessing more than one leads to a greater understanding of the derivative skill. A *conjunctive* relationship requires mastery of both antecedent skills—the skill with the *lowest* level dominates mastery of the derivative skill. Conjunctive relationships are particularly appropriate when different skills must be combined to understand the derivative skill. In contrast, a *disjunctive* relationship requires mastery on only one of the antecedent skills—the skill with the *highest* level dominates mastery of the derivative skill; however, unlike the compensatory relationship, mastery of more than one skill provides no additional advantage for mastery of the derivative skill. Finally, an *inhibitor/enabler* relationship requires a minimum level on one skill before mastery of the other antecedent skill begins to have an effect on mastery of the derivative skill. These relationships may be useful when a minimum understanding of one of the antecedent skills is necessary to begin to make sense of the combination of the antecedents to form an understanding of the derivative. These basic relationships can be combined to describe complex interactions between a number of different skills (Mislevy et al., 2002). For example, combining a disjunctive and conjunctive relationship might require mastery either of ability A or of both abilities B and C.

Context variables. By the local independence assumption, all observed variables must be conditionally independent, given the latent variables. This assumption likely will not hold when observed variables involve a common stimulus, such as multiple items for a single reading passage or science scenario and multiple measures from an extended piece of student work (such as an essay, project, or simulation). The dependencies between items with a common stimulus can be accounted for by adding a context variable to the graph (Almond, Mulder, et al., 2009;

Levy & Mislevy, 2004). A latent context variable generally has no parents and is parent to the observed variables sharing the common stimulus; it is generally an unreported nuisance parameter included only to absorb the domain-irrelevant covariance due to the common context.

Learning from data versus expert opinion. In the absence of strong theory about the relationships between domain components, one might be tempted to try to learn network structure purely empirically using data from real examinees. However, learning latent BN structure is challenging and an area of current investigation (Almond et al., 2007; Liu, 2009); for example, Almond (2010) proposed a method for deriving graphical structure from a correlations matrix. Most current educational assessments with BN knowledge models obtain their network structure from expert opinion (e.g., Butz et al., 2006; Mislevy et al., 2002) or a blend of expert opinion and exploratory empirical methods. For example, Liu (2009) used simulated data from several candidate models to train classifiers (e.g., support vector machines, artificial neural networks, classification trees, etc.), which were then used to classify real response data; the model whose classifier best fit the real data was taken to be the best model. Other methods for blending expert and empirical graph specification include comparing candidate models provided by experts via various indices for model criticism.

Once the structure has been determined, parameters for each conditional distribution must be specified. While some have obtained these parameters empirically from student data (e.g., Almond et al., 2007; Liu, 2009), others obtain conditional probability tables from content experts (e.g., Conati et al., 1997; Read et al., 2002). van der Gaag, Renooij, Witteman, Aleman, and Taal (1999) outlined a method for eliciting a large number of conditional probability distributions from experts efficiently. A middle-ground Bayesian alternative is to treat the parameters themselves as random and to use expert opinion to guide selection of hyper-priors (Mislevy et al., 2002). Additional work is needed to understand how the choice of expert versus empirically determined network parameters affects the quality of knowledge estimates.

Dynamic Networks

The modular model-building aspect of BN provides a flexible framework for modeling the changes in knowledge over time (Corbett & Anderson, 1995; Mislevy & Gitomer, 1996; Reye, 2004). In a *dynamic* BN, specific sub-graphs (model fragments) correspond to different moments in time. Typically, the latent variables in each sub-graph (time slice) represent the same set of concepts with the same structural relationships, though the observed variables present in each time slice may differ (Figure 5). In a basic dynamic BN, sub-graphs are connected via edges between corresponding latent variables in adjacent time slices; however, more complex temporal relationships are possible, such as modeling the change in one latent variable on other related latent variables (Schäfer & Weyrath, 1997). To keep the graph size manageable in applications with many time points, old information can be summarized by computing the posterior distribution for the latent variables using this as the prior distribution for the succeeding time slice (e.g., Conati et al., 2002), a procedure known as ''rollup.'' For example, the Prime Climb educational game about prime factorization maintains a time slice for each player interaction (mouse click) with rollup between game scenarios (Conati & Zhao, 2004).

Modeling the change in knowledge over time can become considerably complex as students may make progress at times and regress at others—each student's trajectory will be different and may be non-linear. To reduce the number of model parameters, many of the early ''knowledge tracing'' systems made the simplifying assumption that students do not forget concepts once they have learned them, but the implications of this assumption for the accuracy of estimates of knowledge growth has not been investigated. Although this assumption can be relaxed

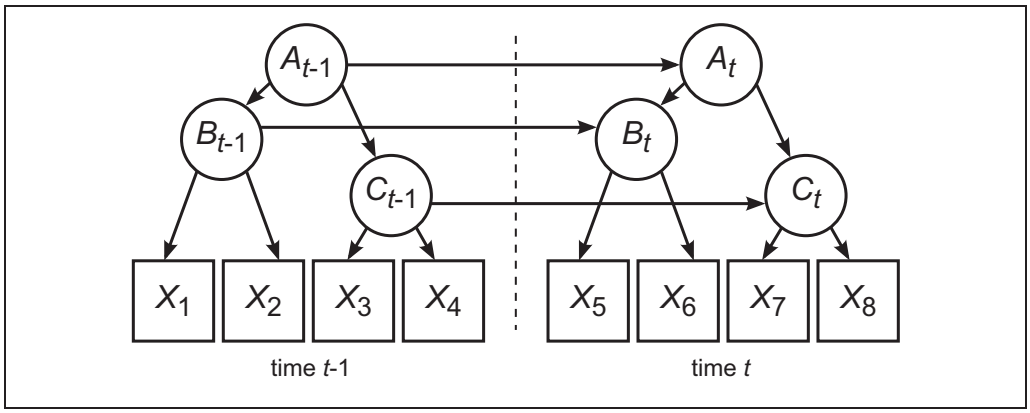

Figure 5. Simple dynamic Bayesian network with two time points.

for a more-realistic model of students' knowledge over time (Yudelson, Medvedeva, & Crowley, 2008), choosing a parametric form for students' knowledge as a function of time is currently understudied. Moreover, little is known about the recovery of the temporal model parameters or the conditions necessary for adequate calibration of dynamic networks.

Model Criticism

Once a set of candidate graphs has been identified, either through expert opinion or empirical means, the potential models can be compared via a number of techniques for model criticism. Methods for model criticism can also be used to detect model misfit. Although no consensus has yet emerged for evaluating the quality of a BN graph (Conati et al., 2002), a number of different techniques have been proposed for comparing models.

Mutual information (MI). The MI between two random variables indicates how significantly their joint distribution deviates from the joint distribution that would result if the two variables were independent. That is, MI indicates the degree to which two variables are associated. As a BN implies conditional independencies between variables, MI may be useful in comparing rival structures. For example, Liu (2009) applied MI to distinguish between learning progressions. Nodes in the BN represented all possible combinations of basic concepts, and the analytic goal was to determine in which order the concepts were most likely to be combined. For example, two competing progressions might suggest either that concepts A, B, and C are combined directly to form ABC, or that A and B are first combined before combining with C. If the MI between the individual concepts and the combination ABC is greater than the MI between ABC and potential parents AB and C, the former learning progression is more likely than the latter.

Model fit indices. Several different potential indices for model fit for BN in educational assessment have been investigated. One simple index proposed by Pardos, Heffernan, Anderson, and Heffernan (2007) is the mean absolute difference between predicted and actual number-correct test scores, where predicted test score is estimated by leave-one-out cross-validation (LOOCV) to reduce the effects of over-fitting. Williamson, Mislevy, and Almond (2001) investigated the performance of model fit statistics under several different model error conditions, including spurious/omitted nodes, spurious/omitted edges, and erroneous prior distributions, and found that the ranked probability score (Epstein, 1969) and Weaver's (1948) surprise index performed well as global measures for detecting BN graphical errors. As the true classification of

examinees on latent variables is not generally known, the traditional Pearson χ^2 fit statistic does not follow a χ^2 distribution under the null model, but the χ^2 statistic may be used simply as a model comparison criterion (Sinharay & Almond, 2007). Alternatively, the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & van der Linde, 2002) provides a Bayesian analog to traditional information criteria for model comparison based on complexity-penalized model fit.

Posterior predictive model checking (PPMC). PPMC (Rubin, 1984) can provide an empirical null distribution for fit statistics with unknown asymptotics. In PPMC, several sets of item responses are simulated from the proposed model using sets of parameters drawn from the posterior distribution. A discrepancy statistic is then calculated between the actual data and each of the simulated data sets, and the posterior predictive p value is obtained from the proportion of replicated sets whose discrepancy statistic is exceeded by the discrepancy for the actual data. Small p values suggest model misfit. Proposed discrepancy measures include proportion correct (for items or examinees), mean squared Pearson residuals, point bi-serial correlations, and odds ratios between item pairs (Sinharay, 2006), as well as a generalized dimensionality discrepancy measure (Levy & Svetina, 2011). Although each of these has been investigated for utility in determining model fit, there has been no clear determination of which discrepancy measures perform the best or how PPMC compares with other model fit techniques.

Diagnostic plots. In addition to quantitative measures of model fit, various diagnostic plots can provide an impression of whether a particular BN fits the data (Sinharay, 2006). In the direct data display, the item responses matrix is plotted as a binary or grayscale image. Several simulated item matrices from the posterior distribution for parameters are plotted next to the real-data image, and the patterns in the images are inspected for deviations. In addition, item fit plots (empirical item characteristic curves) or plots of Bayesian residuals (either for individual items or the entire test) against estimated abilities can suggest whether there are large departures from the data predicted by the proposed model. Finally, plots of prior versus posterior distributions for item parameters can suggest problems with identifiability: If a posterior distribution is not very different from the corresponding priors, there is little information available to identify the given parameter.

Adaptive Item Selection

Given increasing demands for educational assessment data, interest in CAT is growing. CAT has the potential to improve the precision of examinee ability estimates by selecting an individualized set of items most informative for the particular examinee. Although item selection algorithms exist for MIRT (C. Wang & Chang, 2011) and DCMs (McGlohen & Chang, 2008), it may be possible to use the graphical information in a BN knowledge model to improve item selection performance. As Almond and Mislevy (1999) proposed a comprehensive model for using BN in CAT, elaboration of the details of such systems has been quite limited, resulting in only a smattering of proposals for item selection and little comparison between them.

Information-Based Selection

An assessment aims primarily at reducing uncertainty in an examinee's ability. As there is an inverse relationship between uncertainty and statistical information, CAT lends itself well to item selection methods based on various information measures. For example, as in usual IRT-based CAT, items could be selected to maximize Fisher information or to obtain the greatest reduction in posterior variance (e.g., Guzmán, Conejo, & Pérez-de-la Cruz, 2007). When latent

abilities are represented discretely, Fisher information does not exist; instead, Shannon entropy provides a measure of the uncertainty in the ability estimate, and items may be selected to minimize the expected posterior entropy (Reye, 2004; Vomlel, 2004), which is equivalent to maximizing the MI between the item response X and examinee ability θ (Weissman, 2007). MI is the Kullback–Leibler (KL) divergence between the joint and marginal distributions of X and θ, and other uses of KL divergence have been applied to item selection, as well. If there is a particular hypothesis of interest about the latent ability (e.g., that ability is above a particular proficiency threshold), items may be selected to maximize the KL divergence between the item's distribution under the hypothesis and under its complement (also known as the expected weight of information; Madigan & Almond, 1996). For continuous latent variables, the hypothesis is often taken to be the current point estimate (cf. Chang & Ying, 1996).

Decision-Based Selection

When a test is intended to classify examinees into groups, whether or not the latent abilities are taken as continuous or discrete, an alternative to information-based methods is to select items that minimize the expected classification decision error. Although item selection based on decision-making criteria has been proposed for BN-based CAT (Vomlel, 2004), it does not appear that the technique has been implemented in practice to date. Similar methods from other CAT studies, such as maximizing the discrimination between adjacent latent classes (Rudner, 2002) or the Sequential Probability Ratio Test (Eggen, 1999; Weissman, 2007), could serve as a foundation for item selection in BN-based classification tests.

Utility-Based Selection

Given the complexity of computing posteriors for every item in the item bank each time an item must be selected, some have suggested using simple heuristics or utility functions. For dichotomous items and abilities, Liu (2005) suggested the difference between success probabilities given mastery or non-mastery. Collins et al. (1996) proposed the absolute difference between positive and negative predictive value. In contrast, Millán et al. (2000) suggested that both these probabilities should be maximized, and thus proposed a utility measure based on the sensitivity and specificity of the items. When items load onto more than one ability, Millán and Pérez-de-la Cruz (2002) have suggested either to sum over or take the maximum utility function from among the item's relevant abilities. Unfortunately, the comparative performance of these metrics and information-based methods is not known.

Node Priority

Many of the proposed selection techniques pick the best item from those that measure a particular latent ability, but they do not directly integrate the decision of which latent ability to measure next. One option is to select the BN node with the greatest variance (e.g., Guzmán et al., 2007). Another option is simply to iterate through the latent abilities. Liu (2005) has suggested a modified version of the iteration strategy in which the next concept is chosen to be maximally dissimilar to previously administered concepts. Although procedures such as these are simple to implement, item selection algorithms that focus on only one BN node at a time might select items that are not optimal for all nodes. For example, if items for some nodes are uniformly more informative than items for other nodes, an iterative procedure that cycles through all nodes will gather much more information about some than others.

Current Challenges and Research Directions

Although BN have been used extensively over the last 20 years in the artificial intelligence community, BN-based knowledge models have received only limited acceptance in mainstream psychometrics. As demonstrated by tutoring system exemplars (see the online appendix), BN provide a convenient way to specify complex relationships between latent cognitive variables. However, given the relatively small research base on these complex models, there remain a number of challenges and open research questions that have hindered BN from achieving more widespread use. A comprehensive exploration of these challenges falls outside the scope of this article, but the following section enumerates some of the most prominent research needs in the field.

Graph Development

Of primary concern are matters related to graph development. BN have been applied to coarse- and fine-grain skills, but there are few guidelines for how to choose the appropriate grain size for any given assessment. A good starting point is to consider the grain size necessary to support the decisions to be made based on assessment results (Martin & VanLehn, 1995). Turning to granular hierarchies does not completely resolve this issue, as one must still determine the level of detail of the finest grain size and the number of hierarchical layers, and there is likely a trade-off between the marginal utility of more-detailed information provided by increasingly fine-grained skills and the resulting measurement complexity requiring more items to assess each skill adequately. Moreover, most BN-based knowledge models have made use of discrete latent variables, but there has been little investigation into how to determine the optimal number of discrete knowledge levels or whether continuous latent variables might provide better performance. Learning the structure between knowledge nodes entirely empirically may be too complex for operational usage at present. Instead, expert opinion can certainly continue to guide the development of graphs, but a priori graphical structures need to be vetted empirically. Some initial work has been completed to develop procedures for identifying graph mis-specification, but additional testing is necessary to understand the robustness of estimates from BN-based assessments under a greater variety of contexts.

Parameter Recovery

Once a satisfactory graphical structure is in place, assessment developers face other practical questions: How many field-test examinees are necessary to achieve adequate calibration of the knowledge model structural parameters (e.g., priors, path coefficients)? How many items are necessary to achieve the desired precision for ability estimates? Carmona et al. (2005) have demonstrated improved precision of models with explicit relationships between abilities over an independence model; but, it is unclear whether the parsimony of a BN will provide improved precision over a multi-dimensional model with arbitrary-covariance matrix, or whether the potential for over-fitting of an arbitrary-covariance model outweighs the dangers of mis-specifying a BN. Extensive parameter recovery studies (for structural parameters, item parameters, and ability estimates, particularly in hierarchical models) would provide an enormous asset to practitioners looking to apply BN knowledge models in new operational assessment contexts.

Item Selection

A great deal of work remains in terms of developing item selection techniques for adopting BN-based knowledge models in CAT. Research in this area could benefit from a thorough

comparison of the performance of the various selection criteria already proposed in the BN-based CAT literature. Then, item selection investigations could move forward to answer open questions such as the following: Should item selection focus on one ability at a time, or aim for a more holistic approach to quantifying the informativeness of items? What are the characteristics of an informative item when considering the knowledge graph as a whole? Are existing item selection methods for MIRT readily adaptable to BN models? Does the type of graphical structure affect which method performs the best?

Vertical Scaling

In the continual effort to monitor educational improvement, many have become interested in vertical scales that can track student progress across many years of schooling. However, others have expressed concern about construct stability across the range of the vertical scale (e.g., S. Wang & Jiao, 2009). After all, math in kindergarten looks very different from math in high school—is kindergarten math really only low-level high-school math? Where construct heterogeneity exists, BN-based knowledge models may provide a means for maintaining fidelity to the varied content in different grades while still providing meaningful information about students' progress across grades.

Comparison of Methods

Research on BN has typically occurred in isolation from research on other methods for modeling knowledge in educational assessment, likely due in part to the fact that much of the research on BN to date has come from the artificial intelligence community as opposed to the psychometric community. To make informed choices, assessment developers and researchers would benefit enormously from a comprehensive conceptual and empirical comparison of the properties and performance of BN and other common psychometric models, including MIRT and DCMs.

Stakes of Assessment

So far, BN have been used almost exclusively in low-stakes formative assessment settings, such as in ITS. This may be due, in part, to the diagnostic nature of the information that BN provide, which is more amenable to making low-level instructional decisions than high-level summative judgments about student progress. Granular hierarchies may present a means for providing both diagnostic and summative scores from an educational assessment; however, there is considerably more work that needs to be done before BN may be considered robust enough to withstand the scrutiny that comes with high-stakes assessment scenarios. In particular, parameter recovery properties must be understood thoroughly, particularly with respect to bias that may be introduced in person parameters due to network mis-specification or poorly calibrated network or item parameters. Although BN are not currently ready for high-stakes testing (and even if they never are), teachers may benefit from the information provided by improved-precision, low-stakes interim assessments based on BN.

Authentic Educational Settings

Finally, for BN to gain traction outside the research lab in operational educational assessment, perhaps the most helpful research agenda could be the application of BN knowledge models in a large variety of content domains and an increasing number of authentic educational settings (see

the online appendix for existing applications). Only as this approach to knowledge modeling is implemented in practice will we fully understand the potential of BN for providing detailed, timely information to teachers for improved and efficient instructional decision making.

Conclusion

The increasing emphasis on standardized educational assessment in current U.S. policy discussions provides a tremendous opportunity for teachers to gain access to valuable tools to help them make instructional decisions more efficiently. But, this potential can only be realized if our educational assessments are closely aligned with instructional practice and provide sufficient diagnostic detail to merit the time spent on testing. BN provide an intuitive framework for modeling content domains at a diagnostic level (e.g., Almond, Shute, et al., 2009), and, by explicitly modeling the internal relationships in the tested domain, assessments become more powerful tools for teachers to make rapid, accurate instructional decisions based on what students know. Although many questions associated with constructing operational assessments based on BN remain, with increased and focused research, BN can gain greater traction in commonplace educational assessment practice, providing rich diagnostic information to support educational success for all students.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Almond, R. G. (2010). ''I can name that Bayesian network in two matrixes!'' *International Journal of Approximate Reasoning*, *51*, 167-178. doi:10.1016/j.ijar.2009.04.005

Almond, R. G., DiBello, L. V., Moulder, B., & Zapata-Rivera, J.-D. (2007). Modeling diagnostic assessments with Bayesian networks. *Journal of Educational Measurement*, *44*, 341-359. doi: 10.1111/j.1745-3984.2007.00043.x

Almond, R. G., & Mislevy, R. J. (1999). Graphical models and computerized adaptive testing. *Applied Psychological Measurement*, *23*, 223-237. doi:10.1177/01466219922031347

Almond, R. G., Mulder, J., Hemat, L. A., & Yan, D. (2009). Bayesian network models for local dependence among observable outcome variables. *Journal of Educational and Behavioral Statistics*, *34*, 491-521. doi:10.3102/1076998609332751

Almond, R. G., Shute, V. J., Underwood, J. S., & Zapata-Rivera, J.-D. (2009). Bayesian networks: A teacher's view. *International Journal of Approximate Reasoning*, *50*, 450-460. doi: 10.1016/j.ijar.2008.04.011

Almond, R. G., Yan, D., & Hemat, L. (2008). Parameter recovery studies with a diagnostic Bayesian network model. *Behaviormetrika*, *35*, 159-185. doi:10.2333/bhmk.35.159

Anderson, C. J., & Yu, H.-T. (2007). Log-multiplicative association models as item response models. *Psychometrika*, *72*, 5-23. doi:10.1007/s11336-005-1419-2

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*, 203-219. doi:10.1037/0033-295X.110.2.203

Butz, C. J., Hua, S., & Maguire, R. B. (2006). A web-based Bayesian intelligent tutoring system for computer programming. *Web Intelligence and Agent Systems*, *4*(1), 77-97.

Cai, L. (2010). Metropolis-Hastings Robbins-Monro Algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, *35*, 307-335. doi:10.3102/1076998609353115

Carmona, C., Millán, E., Pérez-de-la Cruz, J., Trella, M., & Conejo, R. (2005). Introducing prerequisite relations in a multi-layered Bayesian student model. In L. Ardissono, P. Brna, & A. Mitrovic (Eds.), *User modeling 2005: 10th international conference* (Vol. *3538*, pp. 347-356). New York, NY: Springer.

Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, *20*, 213-229. doi:10.1177/014662169602000303

Chung, G. K. W. K., Niemi, D., & Bewley, W. L. (2003, April). *Assessment applications of ontologies*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Collins, J., Greer, J., & Huang, S. (1996). Adaptive assessment using granularity hierarchies and Bayesian nets. In C. Frasson, G. Gauthier, & A. Lesgold (Eds.), *Intelligent tutoring systems: Third international conference* (pp. 569-577). New York, NY: Springer.

Conati, C. (2010). Bayesian student modeling. In R. Nkambou, J. Bourdeau, & R. Mizoguchi (Eds.), *Advances in intelligent tutoring systems* (pp. 281-299). New York, NY: Springer. doi:10.1007/978-3-642-14363-2_14

Conati, C., Gertner, A., & VanLehn, K. (2002). Using Bayesian networks to manage uncertainty in student modeling. *User Modeling and User-Adapted Interaction*, *12*, 371-417. doi:10.1023/A:1021258506583

Conati, C., Gertner, A. S., VanLehn, K., & Druzdzel, M. J. (1997). On-line student modeling for coached problem solving using Bayesian networks. In A. Jameson, C. Paris, & C. Tasso (Eds.), *User modeling: Proceedings of the sixth international conference* (pp. 231-242). New York, NY: Springer.

Conati, C., & Zhao, X. (2004). Building and evaluating an intelligent pedagogical agent to improve the effectiveness of an educational game. In N. J. Nunes & C. Rich (Eds.), *Proceedings of the 9th international conference on intelligent user interfaces* (pp. 6-13). Association for Computing Machinery. doi:10.1145/964445.964446

Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, *4*, 253-278. doi:10.1007/BF01099821

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*, 333-353. doi:10.1007/BF02295640

Desmarais, M. C., & Baker, R. S. J. d. (2012). A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction*, *22*, 9-38. doi:10.1007/s11257-011-9106-8

Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement*, *23*, 249-261. doi:10.1177/01466219922031365

Epstein, E. S. (1969). A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology*, *8*, 985-987. doi:10.1175/1520-0450(1969)008<0985:ASSFPF>2.0.CO;2

Gertner, A. S., Conati, C., & VanLehn, K. (1998). Procedural help in Andes: Generating hints using a Bayesian network student model. In J. Mostow & C. Rich (Eds.), *Proceedings of the fifteenth national conference on artificial intelligence* (pp. 106-111). American Association for Artificial Intelligence.

Goguadze, G., Sosnovsky, S., Isotani, S., & McLaren, B. M. (2011). Towards a Bayesian student model for detecting decimal misconceptions. In F.-Y. Yu, T. Hirashima, T. Supnithi, & G. Biswas (Eds.), *Proceedings of the 19th international conference on computers in education. Asia-Pacific Society for Computers in Education*. Retrieved from http://www.nectec.or.th/icce2011/program/proceedings/pdf/C1_F5_221F.pdf

Guzmán, E., Conejo, R., & Pérez-de-la Cruz, J.-L. (2007). Adaptive testing for hierarchical student models. *User Modeling and User-Adapted Interaction*, *17*, 119-157. doi:10.1007/s11257-006-9018-1

Jameson, A. (1995). Numerical uncertainty management in user and student modeling: An overview of systems and issues. *User Modeling and User-Adapted Interaction*, *5*, 193-251.

Koenig, A. D., Lee, J. J., Iseli, M. R., & Wainess, R. (2010). *A conceptual framework for assessing performance in games and simulations* (Tech. Rep. No.*771*). National Center for Research on Evaluation, Standards, and Student Testing. Retrieved from http://www.cse.ucla.edu/products/reports/R771.pdf

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement*, *41*, 205-237. doi:10.1111/j.1745-3984.2004.tb01163.x

Levy, R., & Crawford, A. V. (2009). *Bayesian network modeling for student-and domain-level inferences*. Unpublished manuscript.

Levy, R., & Mislevy, R. J. (2004). Specifying and refining a measurement model for a computer-based interactive assessment. *International Journal of Testing*, *4*, 333-369. doi:10.1207/s15327574ijt04043

Levy, R., & Svetina, D. (2011). A generalized dimensionality discrepancy measure for dimensionality assessment in multidimensional item response theory. *The British Journal of Mathematical and Statistical Psychology*, *64*, 208-232. doi:10.1348/000711010X500483

Liu, C.-L. (2005). Some theoretical properties of mutual information for student assessments in intelligent tutoring systems. In M.-S. Hacid, N. V. Murray, Z. W. Ras, & S. Tsumoto (Eds.), *Foundations of intelligent systems: 15th international symposium* (pp. 524-534). New York, NY: Springer.

Liu, C.-L. (2009). Selecting Baysian-network models based on simulated expectation. *Behaviormetrika*, *36*, 1-25. doi:10.2333/bhmk.36.1

Madigan, D., & Almond, R. G. (1996). On test selection strategies for belief networks. In D. Fisher & H.-J. Lenz (Eds.), *Learning from data: Artificial intelligence and statistics V* (pp. 89-98). New York, NY: Springer.

Martin, J., & VanLehn, K. (1995). A Bayesian approach to cognitive assessment. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 141-165). Mahwah, NJ: Lawrence Erlbaum.

McGlohen, M., & Chang, H.-H. (2008). Combining computer adaptive testing technology with cognitively diagnostic assessment. *Behavior Research Methods*, *40*, 808-821. doi:10.3758/BRM.40.3.808

Millán, E., & Pérez-de-la Cruz, J. L. (2002). A Bayesian diagnostic algorithm for student modeling and its evaluation. *User Modeling and User-Adapted Interaction*, *12*, 281-330. doi:10.1023/A:1015027822614

Millán, E., Pérez-de-la Cruz, J. L., & Suárez, E. (2000). Adaptive Bayesian networks for multilevel student modelling. In G. Gauthier, C. Frasson, & K. VanLehn (Eds.), *Intelligent tutoring systems: 5th international conference* (pp. 534-543). New York, NY: Springer.

Mislevy, R. J. (1994). *Probability-based inference in cognitive diagnosis* (Tech. Rep. No. RR-94-93). Princeton, NJ: Educational Testing Service.

Mislevy, R. J., Almond, R. G., DiBello, L., Jenkins, F., Steinberg, L., Yan, D., & Senturk, D. (2002). *Modeling conditional probabilities in complex educational assessments* (Tech. Rep. No. *580*). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Mislevy, R. J., Almond, R. G., Yan, D., & Steinberg, L. S. (2000). *Bayes nets in educational assessment: Where do the numbers come from?* (Tech. Rep. No. *518*). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Mislevy, R. J., & Gitomer, D. H. (1996). The role of probability-based inference in an intelligent tutoring system. *User Modeling and User-Adapted Interaction*, *5*, 253-282. doi:10.1007/BF01126112

Pardos, Z., Heffernan, N., Anderson, B., & Heffernan, C. (2007). The effect of model granularity on student performance prediction using Bayesian networks. In C. Conati, K. McCoy, & G. Paliouras (Eds.), *User modeling 2007: 11th international conference* (pp. 435-439). New York, NY: Springer. doi:10.1007/978-3540-73078-1_60

Patterson, J. A. (2011). *Deconstructing a domain into its cognitive attributes: Test construction and data analysis* (Unpublished doctoral dissertation). University of Illinois, Urbana–Champaign.

Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*(2), 146-178. doi: 10.3102/10769986024002146

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

Read, T., Bárcena, E., Barros, B., & Verdejo, F. (2002). Adaptive modelling of student diagnosis and material selection for on-line language learning. *Journal of Intelligent and Fuzzy Systems*, *12*(3–4), 135-149.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.

Reye, J. (2004). Student modelling based on belief networks. *International Journal of Artificial Intelligence in Education*, *14*(1), 63-96.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, *12*, 1151-1172. doi:10.1214/aos/1176346785

Rudner, L. M. (2002, April). *An examination of decision-theory adaptive testing procedures*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guilford Press.

Schäfer, R., & Weyrath, T. (1997). Assessing temporally variable user properties with dynamic Bayesian networks. In A. Jameson, C. Paris, & C. Tasso (Eds.), *User modeling: Proceedings of the sixth international conference* (pp. 377-388). New York, NY: Springer.

Shute, V. J. (2005). Stealth assessment in computer-based games to support learning. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 503-524). Charlotte, NC: Information Age.

Sinharay, S. (2006). Model diagnostics for Bayesian networks. *Journal of Educational and Behavioral Statistics*, *31*, 1-33. doi:10.3102/10769986031001001

Sinharay, S., & Almond, R. G. (2007). Assessing fit of cognitive diagnostic models a case study. *Educational and Psychological Measurement*, *67*, 239-257. doi:10.1177/0013164406292025

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*, 583-639. doi:10.1111/1467-9868.00353

Tatsuoka, K. K. (1983). Rule-space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*, 345-354. doi:10.1111/j.17453984.1983.tb00212.x

Tepperman, J., Lee, S., & Alwan, A. (2011). A generative student model for scoring word reading skills. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*, 348-360. doi:10.1109/TASL .2010.2047812

Ting, C.-Y., & Phon-Amnuaisuk, S. (2012). Properties of Bayesian student model for INQPRO. *Applied Intelligence*, *36*, 391-406. doi:10.1007/s10489-010-0267-7

van der Gaag, L. C., Renooij, S., Witteman, C. L. M., Aleman, B. M. P., & Taal, B. G. (1999). How to elicit many probabilities. In K. B. Laskey, H. Prade, & G. F. Cooper (Eds.), *Uncertainty in artificial intelligence: Proceedings of the fifteenth conference* (pp. 647-654). San Francisco, CA: Morgan Kaufmann.

Vomlel, J. (2004). Bayesian networks in educational testing. *International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems*, *12*(Supp. 1), 83-100. doi:10.1142/S021848850400259X

Wang, C., & Chang, H.-H. (2011). Item selection in multidimensional computerized adaptive testing—Gaining information from different angles. *Psychometrika*, *76*, 363-384. doi:10.1007/s11336011 -9215-7

Wang, S., & Jiao, H. (2009). Construct equivalence across grades in a vertical scale for a K-12 large-scale reading assessment. *Educational and Psychological Measurement*, *69*, 760-777. doi:10.1177/00131 64409332230

Weaver, W. (1948). Probability, rarity, interest, and surprise. *Scientific Monthly*, *67*, 390-392. doi: 10.2307/22339

Weissman, A. (2007). Mutual information item selection in adaptive classification testing. *Educational and Psychological Measurement*, *67*, 41-58. doi:10.1177/0013164406288164

West, P., Rutstein, D. W., Mislevy, R. J., Liu, J., Levy, R., Dicerbo, K. E., & Behrens, J. T. (2012). A Bayesian network approach to modeling learning progressions. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning progressions in science: Current challenges and future directions* (pp. 257-292). New York, NY: Springer.

Williamson, D. M., Mislevy, R. J., & Almond, R. G. (2001, April). *Model criticism of Bayesian networks with latent variables*. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

Yudelson, M. V., Medvedeva, O. P., & Crowley, R. S. (2008). A multifactor approach to student model evaluation. *User Modeling and User-Adapted Interaction*, *18*, 349-382. doi:10.1007/s11257007-9046-5