



Dynamic Bayesian Networks in Educational Measurement: Reviewing and Advancing the State of the Field

Ray Reichenberg



T. Denny Sanford School of Social and Family Dynamics, Arizona State University

ABSTRACT

As the popularity of rich assessment scenarios increases so must the availability of psychometric models capable of handling the resulting data. Dynamic Bayesian networks (DBNs) offer a fast, flexible option for characterizing student ability across time under psychometrically complex conditions. In this article, a brief introduction to DBNs is offered, followed by a review of the existing literature on the use of DBNs in educational and psychological measurement with a focus on methodological investigations and novel applications that may provide guidance for practitioners wishing to deploy these models. The article concludes with a discussion of future directions for research in the field.

Advances in technological capacity and accessibility in recent years have opened up possibilities for authentic assessments couched within rich environments that may yield proficiency score estimates that are more predictive of an examinee's real-world performance capabilities than what are produced by more traditional assessments such as paper/pencil exams (Shute, Leighton, Jang, & Chu, 2016; Zapata-Rivera & Bauer, 2012). These assessments often take the form of games (e.g., Chung et al., 2010), simulations (e.g., Almond, Mulder, Hemat, & Yan, 2009; Williamson, Mislevy, & Bejar, 2006), or intelligent tutoring systems (ITSs; e.g., Mislevy & Gitomer, 1995; VanLehn, 2008) and, in many cases can be easily embedded within the course of classroom activity.

Dynamic Bayesian networks (DBNs; Reye, 2004) are a promising tool for modeling student proficiency under these rich measurement scenarios. These scenarios often present assessment conditions far more complex than what is seen with more traditional assessments and require assessment arguments and psychometric models capable of integrating those complexities. An assessment embedded within a level of an educational game, for example, might include a single item or task that is repeated many times depending on whether the player responds correctly (i.e., passes the level). There is also an element of feedback inherent in such a scenario: the player knows with each attempt whether or not they completed the task successfully. In many cases, these applications might also require on-the-fly updating of student proficiency estimates to facilitate task selection, creation, or augmentation. These conditions are not necessarily conducive to modeling the player's proficiency using more widely adopted assessment modeling frameworks such as item response theory (IRT). IRT models, for example, often assume latent constructs consisting of multiple, locally independent tasks each administered a single time with no feedback being given to the respondent. Furthermore, these games might pursue the goal of modeling the player's growth or change in proficiency, a goal not often represented in most applications of IRT. IRT models have been proposed for modeling growth (see Anderson, 1985; Culpepper, 2014; Embretson, 1991; and Von Davier, Xu, & Carstensen, 2011 as examples).

CONTACT Ray Reichenberg  r.reichenberg@asu.edu  T. Denny Sanford School of Social and Family Dynamics, Arizona State University, 951 S. Cady Mall, Tempe, AZ 85287, USA

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/home.

© 2018 Taylor & Francis

However, it is not clear that these models, as currently developed, would be suitable for certain complex assessment situations that are of interest here, characterized by feedback to students that cast doubt on the conditional independence assumptions.

DBNs offer a number of strengths in the context of the complex assessment scenarios detailed previously. Almond, Mislevy, Steinberg, Yan, and Williamson (2015, pp. 14–16) presented reasons for considering the use of Bayesian networks (BNs), all of which apply to DBNs. Of particular importance in light of the demands presented by complex assessment scenarios are DBN's ability to handle very large and/or complex models while remaining computationally efficient (i.e., they are fast). This efficiency means not only that a researcher may save time in estimating model parameters, but also that the model can be queried in real-time for diagnostic updates (Almond et al., 2015). This real-time updating makes these models well-suited for use in computer adaptive testing (CAT; Almond & Mislevy, 1999), game/simulation-based assessment, diagnostic assessment, and, potentially in classroom-based formative assessment scenarios where teachers need to make decisions on-the-fly (Almond, Shute, Underwood, & Zapata-Rivera, 2009). Although more prevalent methods such as IRT also have utility in some of these areas (e.g., CAT), the computational efficiency advantages DBNs provide often set them apart and, in some cases, such as when dealing with very large and/or complex systems of variables, may position DBNs as the only feasible option. Finally, scores resulting from DBNs might be more easily understood by consumers (e.g., researchers, assessment designers, teachers, parents, students) given that latent proficiency variables in these models are often assumed to be categorical, a choice that simplifies the interpretation and representation of those proficiencies (Almond, Shute, Underwood, & Zapata-Rivera, 2009, Almond et al., 2015).

Unfortunately, the body of literature related to the use of DBNs in educational measurement and related fields such as psychological measurement is rather sparse. If DBNs are to gain wider use, thus leveraging their apparent strengths under conditions such as those previously mentioned, then the knowledge base surrounding their use must be made more robust and understanding of their structure, function, strengths, and potential utility among both researchers and practitioners must be increased. In this early stage of adoption, the additions to the literature should follow one of two threads: (a) methodological investigations aimed at better understanding the psychometric properties of DBNs and providing practitioners with guidelines for use or (b) novel applications of DBNs to measurement problems—the latter of which may serve as a roadmap for applied researchers seeking to use DBNs in their research. In light of these needs, the purpose of this article is twofold:

- (1) To review the existing literature in educational measurement related to the previously mentioned lines of research ([a] and [b] above).
- (2) To identify and discuss opportunities for further research on the use of DBNs in measurement contexts.

Exploration of these topics will require a brief overview of BNs and DBNs focusing on aspects relevant to psychometrics and will include an illustrative example using real data. The current article will focus on DBNs that are specified using categorical observed and latent variables and will not consider models using continuous variables (such as Kalman filters), although approaches for using DBNs with continuous variables exist and may retain advantages relative to other modeling frameworks under such specifications (see Ghahramani & Hinton, 2000; Johns and Woolf (2006) applied a similar model in an educational research context). The assumption of a fixed structure will also be inherent in subsequent discussions. Although the literature on machine learning and data mining offer methods for learning the structure of a model given some data, most psychometric applications are predicated on the notion of an assessment designed with some target structure in mind.

An Overview of (Dynamic) Bayesian Networks

As DBNs represent a longitudinal extension of BNs, a brief discussion of the latter is warranted to ground discussion of the former.

Bayesian Networks

A BN is a multivariate distribution of discrete variables, commonly depicted as an acyclic directed graph (aka directed acyclic graph) to express the dependence and conditional independence assumptions in the model for the joint distribution (Jensen, 1996). More concretely, a BN models the probability of an event or state such as a latent proficiency conditioned on a set of observed states, events, or characteristics such as item responses. A BN consists of a set of variables (often represented as “nodes” in the graph) and a set of “edges” between the variables. These edges are directed (i.e., single-headed arrows) and define the structure of the network. Under the representations considered here, each variable included in the model may take on a finite set of mutually exclusive states (i.e., they are categorical). More comprehensive overviews of BNs can be found in Nielsen and Jensen (2009), Neapolitan (2004), and Pearl (1988). See Almond et al. (2015), Culbertson (2015), and González-Brenes, Behrens, Mislevy, Levy, and DiCerbo (2016) for didactic treatments and reviews of BNs in educational assessment.

Dynamic Bayesian Networks

DBNs represent a longitudinal extension of BNs. In the simplest case, a DBN is a series of cross-sectional, time-specific BNs connected by a spine linking the latent proficiencies at each time point. Figure 1 represents this notion graphically. The joint probability distribution for this simple example can be expressed as

$$P(\theta_{t1}, \theta_{t2}, X_{t1}, X_{t2}) = P(\theta_{t1})P(\theta_{t2}|\theta_{t1})P(X_{t1}|\theta_{t1})P(X_{t2}|\theta_{t2}) \quad (1)$$

Note that this graph (as well as Equation 1) suggests three sets of parameters that need to be either estimated or specified: (a) the prior state of the proficiency node at time $t1$ ($P(\theta_{t1})$), (b) the conditional probability distribution (CPD) of the observed node at time $t1$ given the state of the latent proficiency node at time $t1$ ($P(X_{t1}|\theta_{t1})$; the “observation model”), and (c) the CPD for the proficiency node at time $t2$ given the CPD of the proficiency node at time $t1$ ($P(\theta_{t2}|\theta_{t1})$; the “transition model”). When using DBNs, it is almost always the case that the observation model is assumed to be static across time slices (i.e., $P(X_{t1}|\theta_{t1}) = P(X_{t2}|\theta_{t2})$). The state of the proficiency node at time $t2$ is dependent on the state of the proficiency at time $t1$ and has encoded in it all the evidence that has been observed up to that point as well as the initial or prior beliefs about the latent proficiency that were present before any observations were made. More generally, the CPD for the

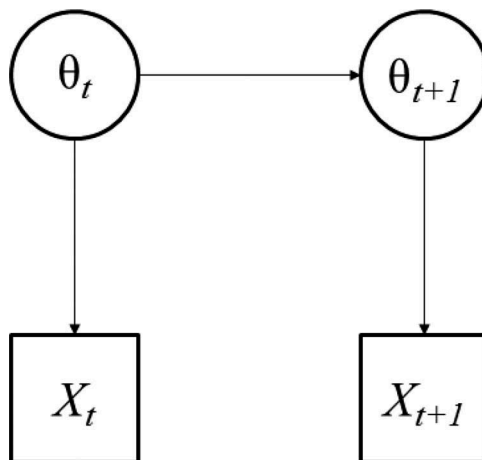


Figure 1. Graphical representation of a simple dynamic Bayesian network.

latent proficiency at any time point $t + 1$ depends only on the state of the proficiency at time t and does not depend on the sequence of proficiency states that preceded time t . As with most psychometric models, applications of BNs almost always carry with them an exchangeability assumption (thus the lack of any subscripts referring to individual cases in Equation 1) in that we assume, *a priori* that each examinee is no different from any other examinee. That is to say that the model structure is assumed to hold for each respondent.

In many classic applications, the state of the proficiency variable can take on one of two values, usually interpreted as defining higher and lower proficiency groups. Similar to how these latent variables have been conceived of in the diagnostic classification literature (Rupp, Templin, & Henson, 2010), the levels are sometimes ascribed names such as “master” or “non-master,” conveying interpretations with respect to the mastery of a skill that the latent variable is intended to represent. This binary classification yields a 2×2 conditional probability table (CPT) for the probability of “transitioning” from one state to the next between two adjacent time points. It is the values in this table, or “transition matrix,” that define the “transition model” discussed previously. Table 1 presents an example transition matrix for two adjacent time points. The values in the matrix represent conditional probabilities, such that each row gives the conditional probability distribution for mastery or non-mastery at time $t + 1$ given mastery or non-mastery at time t . Note that in Table 1, $P(\theta_{t2} = NM|\theta_{t1} = M) = 0$ and that $P(\theta_{t2} = M|\theta_{t1} = M) = 1$. This embodies a “once a master, always a master” assumption, where a respondent never regresses to non-mastery once mastery is achieved.

If we further assume an observed variable (e.g., a performance task presented to an examinee) consisting of two outcomes (i.e., “correct” and “incorrect”), then we end up with a 2×2 CPT for the probability of the examinee demonstrating either of the outcomes on the performance task conditioned on their membership in either of the categories on the proficiency variable. These probabilities define the “observation model.” Table 2 presents an example CPT for a situation such as that described above. In this example, the probability of a student who has mastered the content correctly endorsing the item is 0.70 while the probability of a non-master correctly endorsing the item is only 0.20.

An Illustrative Example

Figure 2 presents a simplified version of a DBN resulting from the *Save Patch* game (Chung et al., 2010; the scoring model by Levy, 2014 is described in a later section) in which players are asked to place sections of rope in order to navigate an avatar (named *Patch*) across an expanse. The game assesses a variety of mathematical skills across its various levels. The example presented here is adapted from an early version of a level from the game that was designed to assess understanding of whole numbers. Response data from 852 students over a maximum of 10 attempts was used to calibrate the model. Only the first three time slices are presented in Figure 2 (and subsequent figures) for the sake of simplicity. Student responses were categorized as the expected successful solution (*StandardSolution*), an unexpected successful solution (*AlternateSolution*), a complete attempt that

Table 1. Sample transition matrix for a dynamic Bayesian network.

	$\theta_{t2} = NM$	$\theta_{t2} = M$
$\theta_{t1} = NM$	0.7	0.3
$\theta_{t1} = M$	0	1

Table 2. Sample conditional probability table (CPT) for a dynamic Bayesian network.

	$\theta_t = NM$	$\theta_t = M$
$P(X_t = 1 \theta_t)$	0.20	0.70
$P(X_t = 0 \theta_t)$	0.80	0.30

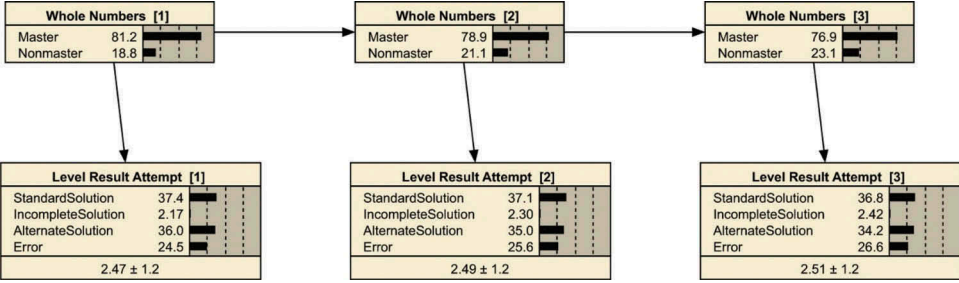


Figure 2. Netica representation of the first three time points for a calibrated DBN.

did not successfully navigate *Patch* across the expanse (*Error*), or an incomplete attempt (*IncompleteSolution*). In terms of proficiency estimates, students were characterized as having either mastered (*Master*) or not mastered (*Nonmaster*) the assessed skill. Tables 3 and 4 present the estimated CPTs for the measurement model and transition model, respectively, for the calibrated network. Note that “once a master, always a master” assumption is not encoded in this model due software limitations.

Once the parameter estimation step is complete, the model can be employed for conducting inference at the level of the individual examinee. Figures 3, 4, and 5 present the example model after task performance at Time 1, Time 2, and Time 3, respectively, have been observed for an examinee. The incorrect (*Error*) response at Time 1 dramatically reduces the probability of mastery in our estimate of the student’s proficiency at the first time point as well as our estimates of the probability that they will have mastered the content following their attempts at two future time points. Correspondingly, our beliefs about the student’s likelihood of providing a correct response (*StandardSolution* or

Table 3. Measurement model CPT for the illustrative example.

	$\theta_t = NM$	$\theta_t = M$
$P(X_t = SS \theta_t)$	0.263	0.400
$P(X_t = IS \theta_t)$	0.069	0.011
$P(X_t = AS \theta_t)$	0.021	0.438
$P(X_t = E \theta_t)$	0.647	0.151

Note. SS = Standard Solution; IS = Incomplete Solution; AS = Alternate Solution; E = Error.

Table 4. Transition model CPT for the illustrative example.

	$\theta_{t2} = NM$	$\theta_{t2} = M$
$\theta_{t1} = NM$	0.919	0.081
$\theta_{t1} = M$	0.047	0.953

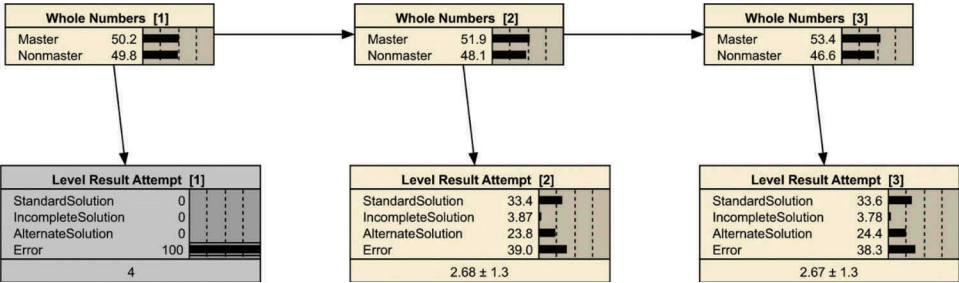


Figure 3. Netica representation of a DBN with one observed item response.

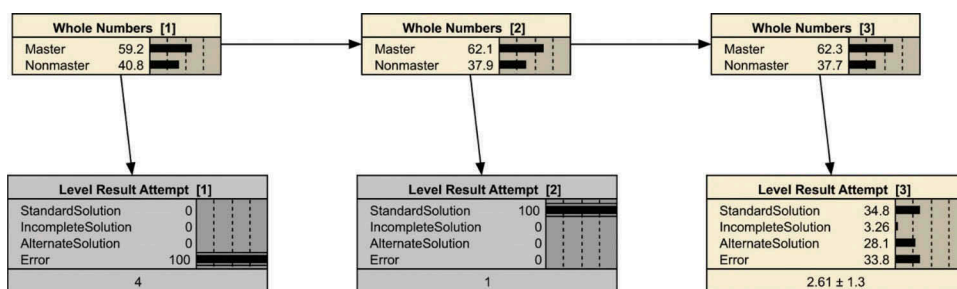


Figure 4. Netica representation of a DBN with two observed item responses.

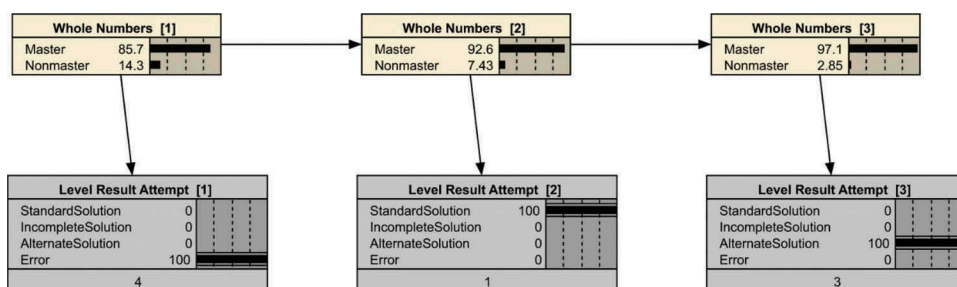


Figure 5. Netica representation of a DBN with observed responses to the first three items.

AlternateSolution) at either of the two future time points also trend toward skepticism. By the time we reach the conclusion of the observed task performance at Time 3, however, our beliefs about the student's probability of having mastered the content have improved significantly due to the two correct responses observed at Time 2 and 3. The solution provided by the student at Time 3, in particular (*AlternateSolution*) has a dramatic effect on the posterior distribution for the latent variable due to the disparity in the probability of a master ($P(X_t = AS|\theta_t = M) = 0.438$) versus a non-master ($P(X_t = AS|\theta_t = NM) = 0.021$) providing such a response.

Estimation

An important question when using DBNs, as with any statistical model, is how to obtain the parameter values. Three approaches are most common with the use of DBNs—eliciting probabilities from content experts, employing some algorithm to estimate or “learn” the parameters, or some combination of the two. The first typically involves some structured process such as the one suggested by Renooij (2001). The second often comes in one of two flavors—maximum likelihood via the expectation-maximization (EM) algorithm (see Myung, 2003) or Markov Chain Monte Carlo estimation (MCMC; see Gilks, Richardson, & Spiegelhalter, 1996. Murphy (2002) provides an overview of these approaches. The third option might be the closest analog to common psychometric practice where subject matter experts help to define features of the model (e.g., levels of the latent variable, task scoring rules, model structure, hypothesized causal relationships between the latent variables) following which pilot data are used to estimate parameters and critique the model. This approach may also be more useful in the event that the amount of available pilot data (e.g., small sample size) is potentially insufficient for estimating model parameters with the desired level of precision. The aforementioned work by Levy (2014) provides an applied example of the melding of evidence from subject matter experts and pilot data.

In the case of the illustrative example, the values presented in Tables 3 and 4 were estimated using the EM algorithm (see Levy, 2014 for an applied example using MCMC estimation) in the Netica software

Table 5. Comparison of relevant model features.

	Features					
	Time Points (<i>t</i>)	LVs per <i>t</i>	OVs per <i>t</i>	Nature of LVs	Nature of OVs	Decision node(s)?
Generalized DBN	Multiple	Multiple	Multiple	Categorical or Continuous	Categorical or Continuous	Yes
Latent transition analysis	Multiple	One	Multiple	Categorical	Categorical	No
Particle filter/Kalman filter/etc.	Multiple	One	One	Continuous	Categorical or Continuous	No
Markov decision process	Multiple	One	One	Categorical	Categorical	Yes
Hidden Markov model	Multiple	One	One	Categorical	Categorical or Continuous	No
BKT model	Multiple	One	One	Dichotomous	Dichotomous	No
Latent class analysis	One	One	Multiple	Categorical	Categorical	N/A
Generalized BN	One	Multiple	Multiple	Categorical or Continuous	Categorical or Continuous	N/A

Note. LVs = Latent variables; OVs = Observed variables.

package (Norsys, www.norsys.com). The parameter estimates for the measurement and transition models were constrained to be static (i.e., invariant with respect to time) in order to aid with model identification. Minimally informative start values were supplied for the measurement model CPT in order to help avoid issues with label-switching (Rodriguez & Walker, 2014; Stephens, 2000).

Related Modeling Frameworks

There exist models that are very similar to or, in some cases, even equivalent to DBNs in their structure and/or function. Examples include latent transition analysis models (Collins & Lanza, 2013), state-space models such as hidden Markov models (Ghahramani, 2001; Murphy, 2002), and the Bayesian knowledge tracing model commonly used with ITSs (Corbett & Anderson, 1995; Corbett, Koedinger, & Anderson, 1997), and Markov decision processes (Barber, 2012; Boutilier, Dean, & Hanks, 1999). These models can all be considered as special cases of a DBN. Table 5 presents a comparison of these models, as well as others, in terms of their relevant features. Unfortunately, the literature stemming from these models may not offer much guidance to researchers seeking to employ more generalized DBNs as these models are simple representations relative to the potential size and complexity of the models that can be specified under the DBN framework.

Existing Research on DBNs in Educational Measurement

Methodological Studies

To date, the author is not aware of any methodological studies in the literature that focused on DBNs in an educational or psychological context (or any other context, for that matter). There are examples that focus on BNs (see Almond, Yan, & Hemat, 2007; Culbertson, 2014; Guo, Gao, Di, & Yang, 2015 as examples) but it is not clear to what extent the conclusions drawn in those studies can be applied to DBNs. In particular, those studies necessarily ignore the aspects of DBNs that separate them from BNs (i.e., those associated with the transition model). Additionally, although one can find a few examples of methodological investigations using the Bayesian knowledge tracing (BKT) model (Coetzee, 2014 [sample size recommendations]; Pardos, Bergner, Seaton, & Pritchard, 2013 [meeting Massive Open Online Course challenges]; Qui, Qi, Lu, Pardos, & Heffernan, 2010 [modeling delay in attempts]; Yudelson, Koedinger, & Gordon, 2013 [modeling student-specific variability]) it may be the case that those conclusions do not generalize to the more flexible DBN framework. Continuing with this example, the BKT model typically includes a single latent ability with two performance categories (i.e., binary) identified by a limited number of tasks. In that way, BKT is akin to the

“simple” DBN presented in Figure 3. This specification suggests that the methodological work using BKT probably ignores model complexity as a factor influencing model performance. That work would then have limited utility to a researcher using a more complex design such as that presented by Levy (2014; discussed in the next section) which contains many latent proficiencies as well as hidden nodes representing misconceptions the examinee may harbor.

Novel Applications of DBNs to Educational Measurement Problems

This section aims to review examples of DBNs being applied to novel problems in educational assessment. This review is not intended to be comprehensive but rather is intended to provide readers new to DBNs with examples of how DBNs might be deployed in practice.

Levy (2014) detailed a DBN-based scoring model for the educational game *Save Patch*—a modified excerpt of which was presented earlier as an illustrative example of a DBN. During the game, players (examinees) complete levels of the game by using math skills to navigate from the beginning of the level to the end. The player must strategically place ropes (using various mathematical operations and understandings) such that the game character, Patch, can traverse the board to the target destination. If Patch successfully reaches the target destination, the player proceeds to the next level, where again a setup is provided, along with resources. If an attempt at a level is unsuccessful in that Patch does not make it all the way to the destination, the student remains at the current level and tries again. The game is designed to measure multiple skills concerning proficiency working with rational numbers. Earlier levels target simpler skills, with later levels targeting more complex skills. As suggested by the description of gameplay, feedback occurs during the game in that successfully completing a level leads to being presented a new level, while unsuccessfully completing a level leads to being presented the same level for another attempt. That is, in contrast to common assessments, the player knows whether their past attempt was successful or not. The performance of a player on an attempt is characterized in terms of a polytomous variable taking on values across 18 categories, most of which correspond to various solution strategies or errors linked to particular misconceptions. The study used an adaptation of the Scaling Individuals and Classifying Misconceptions model (Bradshaw & Templin, 2014), which specifies discrete latent variables for both the proficiencies as well as the misconceptions. Results of the study suggested that the DBN framework using MCMC estimation is suitable for use with game-based assessment but noted issues with estimation resulting from data sparseness due in part to the fact that not all variables were assessed in each level. This suggests a need for games that are designed with robust psychometric analyses in mind such that there is a synergy between the conditions that make the game-experience engaging and educational and conditions that produce data that are suited for the available analytic techniques.

Carmona, Castillo, and Millán (2008) developed a DBN for characterizing students’ learning styles based on the learning objects that those students choose to interact with as well as their reported rating of those objects (scored 1–4 “stars”). Their models considered four learning styles: a preference for visual versus verbal input, active versus reflective processing, sensing versus intuitive perception, and sequential versus global understanding (Felder & Silverman 1988). Each learning object rating is defined as a function of the type of task, the format of the task (e.g., audio, video, text), and the student’s preferred learning style. Prior information for learning style preference was obtained via the Index of Learning Style Questionnaire (Felder & Spurlin, 2005). The researchers chose to model each of the learning styles separately due to the computational complexity of modeling all four styles simultaneously (reportedly requiring the estimation of 162,000 parameters). Each time slice in the DBN is initialized upon the selection of a new learning task by the student. In this way, the beliefs about the student’s preferred style are updated as information about new choices come in. The work was exploratory in that little validation or model critiquing was conducted. Parameters were specified using expert opinions and only a small sample of student data was collected.

Conati and Maclaren (2009) designed a DBN-based model for detecting the emotional state of users interacting with an educational game (*Prime Climb*; developed by the Electronic Games for Education in Math and Science (E-GEMS) group at the University of British Columbia) with the goal of improving student outcomes based on the theory that increased emotional engagement leads to increased attention and learning (Conati, 2002; Ortony, Clore, & Collins, 1988). Their model included both predictive and diagnostic components. The predictive portion uses inputs for user traits, user actions, user-defined goals, and progress toward those goals to predict the user's emotional state while the diagnostic component uses biometric data (e.g., galvanic response) and user expressions to diagnose emotional state. The model was developed using data collected over several rounds of user studies. These data provided the basis for specifying the structure and parameter values for the model. Results suggested that the predictive portion of the model yielded predictions that were more accurate than what would be expected by simply choosing the most likely emotional state (i.e., the population mode). The diagnostic portion of the model was not specifically evaluated in the article.

Interactive Narrative Environments is another area of research in which DBNs have been applied. These narrative environments might be found in role-playing games (RPGs) centered on learning or exploration. For example, in *Crystal Island*, an exploration RPG that teaches the scientific method, the user interacts with a variety of items and characters to determine the source of an infectious disease. The information presented to the user in the form of story elements and dialogue choices is dynamically adapted based on a variety of game and user characteristics. Rowe and Lester (2010) presented a DBN for updating beliefs about the user's knowledge based on their interactions with the environment. These beliefs are used to tailor the narrative elements that the user is presented with. Each time slice of their model contained four knowledge nodes— narrative knowledge, strategic knowledge, scenario solution knowledge, and content knowledge. Each of these latent nodes are identified by a variety of in-game observables. In total, their model consisted of 135 dichotomous observables, 100 edges, and more than 750 CPT elements. The structure and conditional probabilities were specified by the researchers using expert opinions. A small ($N = 167$) data collection was undertaken to test the model. Posterior category membership was compared to the results of a knowledge posttest for the purposes of assessing the accuracy of the model. A model that assigned a uniformly distributed, random probability to each of the knowledge nodes was used for comparison. As one would expect, the target model significantly outperformed the random model in terms of accuracy. The authors note that accuracy might be improved by collecting data from a larger sample to learn the model parameters as opposed to “hand-authoring” (Rowe & Lester, 2010, p. 5) the model.

Using the same gaming environment as Rowe and Lester (2010), Sabourin, Mott, and Lester (2013) developed an early detection system for a learner's self-regulated learning (SRL) capabilities using a DBN guided by research showing that student with low SRL abilities may need scaffolding when operating in a largely self-guided environment such as *Crystal Island*. Early detection of a student's SRL status provides an opportunity for that scaffolding to occur.

Iseli, Koenig, Lee, and Wainess (2010) validated a DBN used for automated scoring of complex tasks. The tasks in this case were simulations of damage control scenarios aboard Navy ships. They were interested in comparing the performance of DBN-based performance scoring software and subject matter experts trained in scoring the tasks in terms of their ability to identify satisfactory performance with the goal of determining whether the automated scoring approach might eventually be able to replace human raters. Their network was specified in collaboration with subject matter experts. Data for a pilot study were collected from 30 university psychology students. All told, there was a high degree of agreement between the automated and judge-scored simulations, although the automated scoring algorithm seemed unable to view the examinee's performance from a holistic perspective. This, as the authors note, is evidence of the difficulty in developing a DBN that approaches a full representation of human knowledge even for a very

specific domain. A more general domain would likely result in very long lead times to develop the DBN scoring model.

When viewed together, one can see that the groundwork is currently being laid for fully adaptive and automated versions of games or simulations. In such an environment, the content-relevant aspects of the game experience such as the context, domain content, performance tasks, and even the scoring of complex tasks might be adapted to the user's ability level and interest set to increase the efficiency of knowledge assessment as well as to maximize learner engagement. This adaptation might be possible with only a relatively small amount of background information on the learner (e.g., via a short survey) and log data from a short period of the learner's gameplay experience. Given the apparent suitability of DBNs to modeling student characteristics in these environments, it stands to reason that additional work is needed to better understand the psychometric properties of these models such that the psychometric tools available to researchers in these areas are able to keep pace with their ambition.

Opportunities for Further Research on the Use of DBNs in Measurement

The potential for further research using DBNs in educational measurement is enormous given the potential benefits of DBNs and the lack of penetration they have seen in the field to date. This section presents issues of general concern to psychometric modeling and, for each, a discussion of the current state of DBN research related to that topic as well as some avenues for future investigation.

Reliability in the Context of DBNs

Reliability is a cardinal issue in the realm of psychometric modeling. Almond and colleagues (2015) offered some coverage of reliability indices for BNs. It is not clear whether these approaches are appropriate for dynamic models. Future research should examine the performance of these indices as well as possibly presenting others. In particular, that research might focus on models with hierarchical structures and the examination of indices for such structures such as those recently popularized in the context of structural equation modeling (SEM) (Green & Yang, 2015; Revelle & Zinbarg, 2009).

Estimation Issues

Most examples of DBNs applied to problems in the social sciences utilize expert opinion to set model parameters. There are guidelines for eliciting those opinions (Almond, 2010; Renooij, 2001) but no universal framework yet exists. Even less guidance is available for those wishing to use empirical estimation techniques. Few rigorous parameter recovery studies have been undertaken to establish benchmarks for sample size or to compare estimation routines (see Coetzee, 2014 for an example of such a study). Furthermore, it is not clear how missing data, degree of model misspecification, choice of prior distribution, or network topology affect parameter recovery. Finally, a framework for melding empirical estimation with expert opinion, using expert guidance to define priors which are then updated in light of observed data, could be established and validated to capitalize on the strengths of the two approaches and mitigate their weaknesses.

Model Critiquing

Related to the specification or estimation of model parameters is the evaluation of model fit. Evaluating data-model fit is an important step in critiquing a model regardless of how the elements of that model were arrived upon (e.g., data-driven approaches, expert opinion). Many of the general classes of options for model criticism of BNs are similar those found in the IRT and SEM literature—graphical evaluation

(see Sinharay & Almond, 2007; Sinharay, Almond, & Yan, 2004), fit indices (see Williamson, Almond, & Mislevy, 2000), and data generation techniques (e.g., poster predictive model checking; see Gelman, Meng, & Stern, 1996; Levy, Mislevy, & Sinharay, 2009; Sinharay, 2006). It should be noted that, to the author's knowledge, the entirety of the literature in this area has focused on BNs and not necessarily DBNs. It stands to reason that many of the methods employed for BNs should be adequate for DBNs but that assumption has not been evaluated directly to date. Beyond these approaches, there exists an opportunity to develop new indices such as novel discrepancy measures for use with a posterior predictive model checking (PPMC)-based approach to critiquing and to explore issues that are poorly understood in the context of DBNs but well studied in the context of other methods. For example, as is the case in SEM, it is easy to construct BNs and DBNs that are weakly identifiable from the data, if they are identifiable at all. Relatively little is known, however, about model identification with BNs/DBNs in terms of how to spot violations (e.g., identifiability rules) or what impact violations have on estimation and inference.

Differential Item Functioning

Differential item functioning (DIF) represents a particular type of data–model misfit wherein the model for members of certain subgroups differs when compared to the model for members of other subgroups. This problem is well known and has been thoroughly (although not exhaustively) examined in the IRT and psychometric literature (Hambleton, 2006) and much of the same logic applied with those models can be transferred to BNs. As Almond et al. (2015) have pointed out, BNs are well-suited to the essential task of DIF analysis—examining the conditional dependence or, hopefully independence of a performance task and an indicator of group membership conditioned on a latent proficiency variable. There exists some literature on DIF approaches for BNs (Sinharay, 2006) but no investigations of the application of those conclusions to DBNs have yet been undertaken nor has there been any discussion in the literature of multiple group approaches to DBN modeling similar to what might be used in an SEM framework. In particular, it is not clear whether those methods can be adapted for the evaluation of differential functioning of temporal parameters (e.g., “Is the probability of transitioning from non-mastery to mastery the same for males as it is for females?”).

Temporal Parameters and Associated Assumptions

The separating factor between BNs and DBNs is represented by specification of the transition matrix, which sets the probability of a learner transitioning from one latent state to another across adjacent time slices. It is these values that are necessarily ignored when studying BNs and as such it is arguable the least studied aspect of DBNs in the realm of measurement. It remains to be seen how sample size, for example, affects the estimation of these parameters relative to the values in the within-time CPTs. Coetzee (2014) offered a brief investigation of the temporal parameters for the BKT model under limited conditions and found that standard errors were typically higher for these parameters than for other parameters in the model (i.e., priors, guess/slip). Additionally, it is often assumed that (a) learners do not regress from one knowledge state to a lesser state with the passage of time and that (b) the values in the transition matrix are static across time. These assumptions are almost certainly false in many measurement contexts. Work needs to be conducted to evaluate the robustness of model inferences to violation of these assumptions as well as to develop and validate model extensions that do not require such assumptions.

DBNs and CAT

There would seem to be a fantastic opportunity to trade on the fast and flexible nature of DBNs in CAT. There is already some literature on the use of BNs with CAT (Millan, Trella, Perez-de-la-Cruz, & Conejo, 2000), but none that has been extended to DBNs. Research should be undertaken to

determine whether the existing CAT item selection and stopping algorithms might apply to BN/DBNs as well as to determine what impact network structure has on CAT performance.

Sample Size Planning

No literature currently exists on power analysis or sample size planning for DBNs for facilitating estimation or model criticism. It would seem that popular tools from other modeling frameworks, such as the Monte Carlo simulation techniques available for SEM models (Muthén & Muthén, 2002), could be used for DBNs with similar effect. Software options for employing those methods have yet to be developed, however.

DBNs and High Stakes Assessment

As some have pointed out (e.g., Almond et al., 2015) the nature of high stakes assessment (i.e., summative; explicit; no need for immediate feedback) does not necessarily capitalize on the strengths of BNs and DBNs. That said, the advantages of psychometrically rich assessment environments are apparent. Given these advantages, it is only natural that teachers, administrators, policy makers, and other stakeholders might desire to use these assessments for making high stakes decisions about student outcomes and abilities. Ideally, this would involve ground-up development of assessments designed for the rigors of high stakes assessment. Unfortunately, it is highly likely that in lieu of new measures, applications may use either existing assessments not intended for high stakes decision making, or other measures that were intended for higher-stakes inferences but with other psychometric models. Regardless, research needs to be undertaken to investigate the applicability of DBNs in these large-scale, high-stakes settings where summative information takes precedent over formative or diagnostic information. This research might be considered as the culmination of the topics listed previously in that it is only after we have a firmer understanding of the psychometric properties of DBNs as well as how best to specify models, estimate their parameters, and critique those estimates that we can be comfortable using those models to impact students' educational pathways in high-stakes scenarios.

Improving Accessibility of DBNs for Practitioners

The relative advantages that DBNs offer over other, more widely used methods beg the question, "Why aren't these models more prevalent in educational research?" The answer to this question has three facets. First, the advantages of DBNs do not serve every researcher. As mentioned above, to use BNs and DBNs for large-scale assessment may serve to undermine their strengths. While useful, these models do not necessarily offer any advantages over the more common IRT/SEM models under these conditions. The types of research where DBNs tend to excel are the same types of research (game/simulation-based assessment, stealth assessment, etc.) that have only come to the mainstream in the past decade or so. Second, there are very few, if any, methodological studies that have focused on the use of DBNs in an educational context and/or offered guidelines for use such as recommended sample sizes, the optimal number of tasks, best practices for accurate estimation of model parameters, how to handle DIF, and so on. The bulk of the current article has focused on this gap in the literature and has argued for the need to fill that gap. Third, while there are a variety of software options for estimating and updating DBNs, most come with a steep learning curve and/or a high financial outlay and very few are capable of handling parameter learning for models containing latent variables (Netica [Norsys, www.norsys.com; EM estimation] and WinBUGS [Lunn, Thomas, Best, & Spiegelhalter, 2000; MCMC estimation] are two such examples). As research scenarios conducive to the use of DBNs become more prevalent, interest in DBNs will likely increase. This increased interest will necessitate the development and implementation of workshops and graduate-level training courses oriented toward the applied researcher wishing to become familiar with DBNs for use in their own research. A significant portion of this training will need to focus on software

implementation. There is and will continue to be opportunities for the development of additional software options, such as *R* (R Core Team, 2017) packages, for example, which include a full suite of options for working with DBNs, particularly those containing latent variables.

Conclusion

DBNs offer relatively clear benefits as compared to more widely applied methods under certain psychometric modeling conditions (Almond et al., 2015, pp. 14–16). As has been discussed, the penetration of DBN-related research in the psychometric literature is not reflective of their potential utility. One goal of this article was to review what little literature there is with respect to methodological investigations and novel applications of DBNs as well as to draw comparisons between DBNs and other, potentially more familiar, methods (see Table 5). This latter goal serves two purposes: (a) to increase awareness and understanding of DBNs among practitioners and researchers (i.e., DBNs are, in many cases a more generalized version of methods that audience may already use), and (b) to provide examples of how DBNs may be employed in practice. A second intent of this work was to discuss areas of need for future research on the use of DBNs for psychometric applications. At present, little is known about the psychometric properties of DBNs and best practices for their use. Further work needs to be done to better understand the suitability of DBNs for use with complex assessment scenarios—particularly those intended for use in making moderate-to-high stakes decisions. Additionally, attention should be given to the coherence of the psychometric and design aspects of complex assessments such as games and simulations (see Behrens, Mislevy, DiCerbo, and Levy, 2012; Mislevy, 2011); and Mislevy et al., 2014 for discussions on this topic). It is only after addressing these open questions that DBNs may find greater use among the psychometric community, thus leveraging their apparent strengths.

References

- Almond, R. G. (2010). I can name that Bayesian network in two matrixes! *International Journal of Approximate Reasoning*, 51(2), 167–178. doi:10.1016/j.ijar.2009.04.005
- 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., Mislevy, R. J., Steinberg, L., Yan, D., & Williamson, D. (2015). *Bayesian networks in educational assessment*. New York, NY: Springer.
- 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(4), 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(3), 450–460. doi:10.1016/j.ijar.2008.04.011
- Almond, R., Yan, D., & Hemat, L. (2007). Parameter recovery studies with a diagnostic Bayesian network model. *Behaviormetrika*, 35(2), 159–185.
- Anderson, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika*, 50, 3–16. doi:10.1007/BF02294143
- Barber, D. (2012). *Bayesian reasoning and machine learning*. Retrieved from <http://web4.cs.ucl.ac.uk/staff/D.Barber/textbook/090310.pdf>
- Behrens, J. T., Mislevy, R. J., DiCerbo, K. E., & Levy, R. (2012). Evidence centered design for learning and assessment in the digital world. In M. Mayrath, J. Clarke-Midura, & D. H. Robinson (Eds.), *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research* (pp. 13–53). Charlotte, NC: Information Age Publishing.
- Boutillier, C., Dean, T., & Hanks, S. (1999). Decision-theoretic planning: Structural assumptions and computational leverage. *Journal of Artificial Intelligence Research*, 11(1), 94.
- Bradshaw, L., & Templin, J. (2014). Combining item response theory and diagnostic classification models: A psychometric model for scaling ability and diagnosing misconceptions. *Psychometrika*, 79(3), 403–425. doi:10.1007/s11336-013-9350-4
- Carmona, C., Castillo, G., & Millán, E. (2008). *Designing a dynamic Bayesian network for modeling students' learning styles*. Advanced Learning Technologies, 2008. ICALT'08. Eighth IEEE International Conference on (pp. 346–350). IEEE. Retrieved from http://ieeexplore.ieee.org/abstract/xpls/abs_all.jsp?arnumber=4561705

- Chung, G., Baker, E. L., Vendlinski, T. P., Buschang, R. E., Delacruz, G. C., Michiuye, J. K., & Bittick, S. J. (2010). *Testing instructional design variations in a prototype math game*. In R. Atkinson (Chair), Current perspectives from three national R&D centers focused on game-based learning: Issues in learning, instruction, assessment, and game design. Structured poster session at the annual meeting of the American Educational Research Association, Denver, CO.
- Coetzee, D. (2014). Choosing sample size for knowledge tracing models. *Educational Data Mining (Workshops)*, London, UK.
- Collins, L. M., & Lanza, S. T. (2013). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences* (Vol. 718). Hoboken, NJ: John Wiley & Sons.
- Conati, C. (2002). Probabilistic assessment of user's emotions in educational games. *Applied Artificial Intelligence*, 16(7–8), 555–575. doi:10.1080/08839510290030390
- Conati, C., & Maclaren, H. (2009). Empirically building and evaluating a probabilistic model of user affect. *User Modeling and User-Adapted Interaction*, 19(3), 267–303. doi:10.1007/s11257-009-9062-8
- Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4), 253–278. doi:10.1007/BF01099821
- Corbett, A. T., Koedinger, K. R., & Anderson, J. R. (1997). Intelligent tutoring systems. *Handbook of Human-Computer Interaction*, 5, 849–874.
- Culbertson, M. J. (2014). *Graphical models for student knowledge: Networks, parameters, and item selection*. University of Illinois at Urbana-Champaign. Retrieved from <https://www.ideals.illinois.edu/handle/2142/49372>
- Culbertson, M. J. (2015). Bayesian networks in educational assessment the state of the field. *Applied Psychological Measurement*, 40(1), 0146621615590401.
- Culpepper, S. A. (2014). If at first you don't succeed try, try again: Applications of sequential IRT models to cognitive assessments. *Applied Psychological Measurement*, 38(8), 632–644. doi:10.1177/0146621614536464
- Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, 56, 495–515. doi:10.1007/BF02294487
- Felder, R. M., Silverman, L. K. (1988). Learning and teaching styles in engineering education. *Engineering Education*, 78(7), 674–681.
- Felder, R. M., & Spurlin, J. (2005). Applications, reliability and validity of the index of learning styles. *International Journal of Engineering Education*, 21(1), 103–112.
- Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6(4): 733–760.
- Ghahramani, Z. (2001). An introduction to hidden Markov models and Bayesian networks. *International Journal of Pattern Recognition and Artificial Intelligence*, 15(01), 9–42. doi:10.1142/S0218001401000836
- Ghahramani, Z., & Hinton, G. E. (2000). Variational learning for switching state-space models. *Neural Computation*, 12(4), 831–864.
- Gilks, W., Richardson, S., & Spiegelhalter, D. (1996). *Markov chain Monte Carlo in practice*. Boca Raton, FL: CRC Press.
- González-Brenes, J. P., Behrens, J. T., Mislevy, R. J., Levy, R., & Dicerbo, K. E. (2016). Bayesian Networks. In A.A. Rupp & J.P. Leighton (Eds.) *The Wiley Handbook of Cognition and Assessment* (pp. 328–353). John Wiley & Sons: Hoboken, NJ.
- Green, S. B., & Yang, Y. (2015). Evaluation of dimensionality in the assessment of internal consistency reliability: Coefficient alpha and omega coefficients. *Educational Measurement: Issues and Practice*, 34(4), 14–20. doi:10.1111/emip.12100
- Guo, Z., Gao, X., Di, R., & Yang, Y. (2015). *Learning Bayesian network parameters from small data set: A spatially maximum a posteriori method*. Workshop on Advanced Methodologies for Bayesian Networks (pp. 32–45). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-319-28379-1_3
- Hambleton, R. K. (2006). Good practices for identifying differential item functioning. *Medical Care*, 44(11), S182–S188. doi:10.1097/01.mlr.0000245443.86671.c4
- Iseli, M. R., Koenig, A. D., Lee, J. J., & Wainess, R. (2010). *Automated assessment of complex task performance in games and simulations*. Proceedings of the Interservice/Industry Training, Simulation and Education Conference. Retrieved from <http://cresst.org/wp-content/uploads/R775.pdf>
- Jensen, F. V. (1996). *An introduction to Bayesian networks* (Vol. 210). London: UCL press.
- Johns, J., & Woolf, B. (2006). *A dynamic mixture model to detect student motivation and proficiency*. Proceedings of the National Conference on Artificial Intelligence. (Vol. 21, No. 1, p. 163). Menlo Park, CA; Cambridge, MA; London: AAAI Press; MIT Press. 1999, 2006.
- Levy, R. (2014). Dynamic Bayesian network modeling of game based diagnostic assessments. CRESST Report 837. *National Center for Research on Evaluation, Standards, and Student Testing (CRESST)*. Retrieved from <http://eric.ed.gov/?id=ED555714>
- Levy, R., Mislevy, R. J., & Sinharay, S. (2009). Posterior predictive model checking for multidimensionality in item response theory. *Applied Psychological Measurement*, 33(7), 519–537. doi:10.1177/0146621608329504
- Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS-a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10(4), 325–337. doi:10.1023/A:1008929526011

- Millán, E., Trella, M., Pérez-de-la-Cruz, J. L., & Conejo, R. (2000). Using bayesian networks in computerized adaptive tests. In *Computers and Education in the 21st Century* (pp. 217–228). Springer, Dordrecht.
- Mislevy, R. J. (2011). *Evidence-centered design for simulation-based assessment*. CRESST report 800 (No. CRESST REPORT 800). National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Retrieved from <http://eric.ed.gov/?id=ED522835>
- Mislevy, R. J., & Gitomer, D. H. (1995). The role of probability# based inference in an intelligent tutoring system. *ETS Research Report Series*, 1995(2), i–27.
- Mislevy, R. J., Oranje, A., Bauer, M. I., Von Davier, A., Hao, J., Corrigan, S., ... John, M. (2014). *Psychometric considerations in game-based assessment*. Redwood City, CA: GlassLab.
- Murphy, K.P. (2002). Dynamic Bayesian networks: Representation, inference, and learning (doctoral dissertation). Retrieved from: <https://pdfs.semanticscholar.org/60ed/db80f54c796750a8173f2abea3bc85a62322.pdf>.
- Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, 9(4), 599–620. doi:10.1207/S15328007SEM0904_8
- Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, 47(1), 90–100. doi:10.1016/S0022-2496(02)00028-7
- Neapolitan, R. E., (2004). *Learning Bayesian networks* (Vol. 38). Upper Saddle River, NJ: Pearson Prentice Hall.
- Nielsen, T. D., & Jensen, F. V. (2009). *Bayesian networks and decision graphs*. Berlin, Germany: Springer Science & Business Media.
- Ortony, A., Clore, G. L., & Collins, A. (1988). The cognitive structure of emotions. 10.1017. CBO9780511571299. doi:10.3168/jds.S0022-0302(88)79586-7
- Pardos, Z., Bergner, Y., Seaton, D., & Pritchard, D. (2013). Adapting Bayesian knowledge tracing to a massive open online course in edx. *Educational Data Mining 2013*. Retrieved from <http://www.educationaldatamining.org/conferences/index.php/EDM/2013/paper/download/1030/996>
- Pearl, J. (1988). *Probabilistic inference in intelligent systems*. San Mateo, CA: Morgan Kaufmann.
- Qiu, Y., Qi, Y., Lu, H., Pardos, Z., & Heffernan, N. (2010). Does time matter? modeling the effect of time with Bayesian knowledge tracing. *Educational Data Mining 2011*. Retrieved from <http://www.educationaldatamining.org/conferences/index.php/EDM/2011/paper/download/897/863>
- R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <http://www.R-project.org/>
- Renooij, S. (2001). Probability elicitation for belief networks: Issues to consider. *The Knowledge Engineering Review*, 16 (03), 255–269. doi:10.1017/S0269888901000145
- Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijsma. *Psychometrika*, 74(1), 145–154. doi:10.1007/s11336-008-9102-z
- Reye, J. (2004). Student modelling based on belief networks. *International Journal of Artificial Intelligence in Education*, 14(1), 63–96.
- Rodríguez, C. E., & Walker, S. G. (2014). Label switching in Bayesian mixture models: Deterministic relabeling strategies. *Journal of Computational and Graphical Statistics*, 23(1), 25–45. doi:10.1080/10618600.2012.735624
- Rowe, J. P., & Lester, J. C. (2010). Modeling user knowledge with dynamic Bayesian networks in interactive narrative environments. *AIIDE*. Retrieved from <https://pdfs.semanticscholar.org/08a8/8dc4db85164de01f8falcfe3013066c49f7d.pdf>
- Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guilford Press.
- Sabourin, J., Mott, B., & Lester, J. (2013). Utilizing dynamic bayes nets to improve early prediction models of self-regulated learning. International Conference on User Modeling, Adaptation, and Personalization. (pp. 228–241). Springer. Retrieved from http://link.springer.com.ezproxy1.lib.asu.edu/chapter/10.1007/978-3-642-38844-6_19
- Shute, V. J., Leighton, J. P., Jang, E. E., & Chu, M.-W. (2016). Advances in the science of assessment. *Educational Assessment*, 21(1), 34–59. doi:10.1080/10627197.2015.1127752
- Sinharay, S. (2006). Model diagnostics for Bayesian networks. *Journal of Educational and Behavioral Statistics*, 31(1), 1–33. doi:10.3102/10769986031001001
- Sinharay, S., Almond, R., & Yan, D. (2004). Assessing fit of models with discrete proficiency variables in educational assessment. *ETS Research Report Series*, 2004(1). Retrieved from <http://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2004.tb01934.x/full>
- Sinharay, S., & Almond, R. G. (2007). Assessing fit of cognitive diagnostic models a case study. *Educational and Psychological Measurement*, 67(2), 239–257. doi:10.1177/0013164406292025
- Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4), 795–809. doi:10.1111/rssb.2000.62.issue-4
- VanLehn, K (2008). Intelligent tutoring systems for continuous, embedded assessment. In C. Dwyer (Ed.) *The Future of Assessment: Shaping Teaching and Learning* (pp. 113–138). Routledge: Abingdon-on-Thames, UK
- Von Davier, M., Xu, X., & Carstensen, C. H. (2011). Measuring growth in a longitudinal large-scale assessment with a general latent variable model. *Psychometrika*, 76(2), 318–336. doi:10.1007/s11336-011-9202-z

- Williamson, D. M., Almond, R. G., & Mislevy, R. J. (2000). *Model criticism of Bayesian networks with latent variables*. Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence (pp. 634–643). Stanford, CA: Morgan Kaufmann Publishers Inc.
- Williamson, D. M., Mislevy, R. J., & Bejar, I. I. (2006). *Automated scoring of complex tasks in computer-based testing*. Mahwah, NJ: Lawrence Erlbaum Associates. Inc.
- Yudelson, M. V., Koedinger, K. R., & Gordon, G. J. (2013). *Individualized Bayesian knowledge tracing models*. International Conference on Artificial Intelligence in Education (pp. 171–180). Springer. Retrieved from http://link.springer.com/10.1007/978-3-642-39112-5_18
- Zapata-Rivera, D. & Bauer, M. (2012). Exploring the role of games in educational assessment. In M.C. Mayrath, J. Clark-Midura, D.H. Robinson, & G. Schraw (Eds.) *Technology- Based Assessments for Twenty-First Century Skills: Theoretical and Practical Implications from Modern Research* (pp. 147–169). Information Age Publishing: Charlotte, NC.

Copyright of Applied Measurement in Education is the property of Taylor & Francis Ltd and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.