Practical Assessment, Research, and Evaluation

Volume 26 Article 22

2021

Model Criticism of Bayesian Networks in Educational Assessment: A Systematic Review

Irina Uglanova
HSE University

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 26 Number 22, November 2021 ISSN 1531-7714

Model Criticism of Bayesian Networks in Educational Assessment: A Systematic Review

Irina Uglanova, *HSE University*

There is increased use of Bayesian networks (BN) in educational assessment. In psychometrics, BN serves as a measurement model with high flexibility, suitable to model educational assessment data with a complex structure. BN is a novel psychometric approach and not all aspects of its application are well-known. The article aims to provide the systematization of BN model criticism methods in the field of educational assessment. The review revealed the diversity of model criticism methods and the shortages of specific research. The results demonstrate the state-of-the-art and help to navigate practitioners and researchers.

Introduction

Bayesian networks (BN) have become a widely applied modeling tool in a variety of research domains (Cruz, Desai, Dewitt, Hahn, Lagnado, Liefgreen et al., 2020). In the last decades, the increased use of BN arose in the area of educational assessment (Culbertson, 2016). In psychometrics, BN serves as a measurement model with high flexibility suitable to model educational assessment data with a complex structure (Almond, Mislevy, Steinberg, Yan, & Williamson, 2015).

The early studies in the field of educational assessment describe techniques for building BN and illustrate different aspects of BN construction, such as the estimation of probabilities within conditional probability tables or the interpretation of inferences about students (e.g., Almond, DiBello, Moulder, & Zapata-Rivera, 2007; Mislevy, Almond, Yan, & Steinberg, 1999; Mislevy, Senturk, Almond, Dibello, Jenkins, Steinberg, et al., 2002; Mislevy, Steinberg, Breyer, Almond, & Johnson, 2002).

DiCerbo, Bertling, Stephenson, Jia, Mislevy, Bauer et al. (2015) mentioned that, in the BN framework, it is convenient to include new observable indicators into the model with the same structure of latent variables. Scalise and Clarke-Midura (2018) discussed the hybridized model, which consists of both multidimensional Item Response Theory (IRT) modeling and BN with a small number of nodes. In their study, the application of BN helps to get new information from assessment systems and make more precise inferences about students.

De Klerk, Veldkamp, and Eggen (2015) demonstrated that BN is the most widely used framework for psychometric analysis in the area of simulation-based and game-based assessment. Usually, the latent characteristics, assessed through the simulation-based and game-based assessment, have an extremely complex structure; hence, a measurement model should be flexible enough to take into account the system of complex relationships among skills and student actions (Becker and Shute, 2010). Therefore, a psychometric modeling approach based on an application of the BN appeared to be useful (e.g., Levy,

2013; Shute & Wang, 2016; Xing, Li, Chen, Huang, Chao, Massicotte et al., 2020).

Notwithstanding BN's benefits, there are several challenges in its application (Levy, 2013). One of them is associated with the model criticism procedure. For psychometric paradigms with a relatively long history, such as the IRT or structural equation modeling (SEM), model criticism methods are well-known (de Ayala, 2009; Hu & Bentler, 1999). In contrast, model criticism methods are underdeveloped for novel psychometric approaches, such as BN (Crawford, 2014). Model fit analysis of BN is a theoretical and practical issue that needs to be answered to the application of BN in educational assessment (Hu & Templin, 2020).

The aim of the study is to reveal the state of the development and application of model criticism methods, within the BN approach, in the context of educational assessment. The study focuses on the application of BN as a measurement model and the following model criticism procedure. In the article, model criticism is considered in a broad sense, including model comparison, different aspects of model misfit, and the functioning of observable variables. In addition, the essential characteristics of model criticism approaches are briefly discussed.

The study is organized as follows. Firstly, we discuss BN as a measurement model and briefly describe model criticism issues within BN. Next, we present the methodology and the results of the literature review analysis.

Bayesian Networks

The BN model is a framework for modeling probabilistic relationships among latent and observable variables and performing a probabilistic inference (Pearl, 1988). More technically, BN is a directed acyclic graph, which represents a complex system of joint probability distribution among nodes that are interrelated by edges. In the context of educational assessment, the latent nodes represent cognitive characteristics, e.g., math skills or critical thinking; the observed nodes represent students' actions, e.g., an answer on a multiple-choice item or an action in a computer game. Thus, edges represent conditional dependencies between latent characteristics and observed actions. Moreover, edges might represent conditional dependencies between students' characteristics or between students' actions themselves.

In educational assessment, the most common practice to build BN is to initially define the structure of the model (Almond et al., 2015). In this case, conditional dependencies between variables are set initially into the model, and the probability inference is obtained within this structure. A given structure of BN should represent theoretical expectations of domain experts, test-developers, or psychometricians about the association between cognitive characteristics and the observed actions. However, a data-driven approach also can be applied if there is an absence of a strong theory or as a source of additional information (Yan, Almond, & Mislevy, 2004).

BN assumes Bayes' theorem application for making a probabilistic inference about the latent variable; therefore, all significant features of Bayesian statistics should be considered (Pearl, 1988). For instance, let us denote x as an observable variable, which represents student action, and θ as a latent variable, which represents a student characteristic. Then, the conditional relationship between these variables, through Bayes' Theorem, is expressed by the following equation:

$$P(\theta \mid x) = \frac{P(x \mid \theta) * P(\theta)}{P(x)} \qquad (1)$$

where P (θ | x) is the posterior probability distribution, which is the distribution of the latent variable conditional on the observed variable; P (θ) is the prior probability representing previous knowledge about the distribution of the latent variable, without considering the empirical data; P (x | θ) is a so-called likelihood that shows the plausibility of the data, given a parameter of the model; and P (x) is a probability distribution of the observed variable, unconditional on any other variable.

Following computational restrictions, all variables included in BN are discrete random variables. In educational assessment, observable variables are usually dichotomous (0 represents an incorrect answer, 1 represents a correct answer) or polytomous (where scores of partially correct answers are added). Contrary to widespread IRT or SEM models, in BN, latent variables are also considered discrete, usually described in terms of latent classes, and show different

proficiency levels (e.g., low, moderate, or high; Levy, 2009).

The Bayesian approach requires setting prior probability distribution regarding model parameters, such as distribution for latent variables. Prior distribution might be gathered from previous research, empirical data, pretesting, or experts' opinions (Almond et al., 2015).

Conditional dependencies between variables are expressed via conditional probability tables (CPT). Each cell of CPT, for observable variable x, represents the conditional probability of being at each state of variable x, given each state of latent variable θ. The values within CPT are also parameters to be estimated through the application of BN, considering prior information and empirical data.

Schematically, the process of building BN can be described in several consecutive steps. In the first step, the variables of interest are defined, and the structure of the model is set. In the second step, prior distributions of model parameters should be specified. Further, model parameters are estimated via any parameter estimation method, taking into account prior information and empirical data. Finally, the next step of the analysis is to verify the quality of the model, the so-called model criticism process. For more information about theoretical and methodological foundations of BN, see Neapolitan (2004), Pearl (1988), and for application in the context of psychometrics, see Almond et al. (2015), Mislevy (1994).

Model criticism

According to modern psychometric standards, theoretical models ought to be validated, which means the quality of theoretical models should be proven (AERA, APA, & NCME, 2014). One of the ways to gather evidence of the validity of the theoretical model is to analyze the data results from students' performances. This analysis aims to conclude whether empirical data support theoretical expectations. This procedure is called model checking or model criticism. As was mentioned by Crawford, "models are built to help evaluate what students know. The models are, themselves, evaluated to see what the model-builders (domain experts) know." (2014, p.2).

According to Sinharay (2006), model criticism analysis for BN is not straightforward. For instance, standard techniques, such as the $\chi 2$-test of the goodness of fit, cannot be applied directly due to many response patterns. Furthermore, the IRT model diagnostics are irrelevant due to discrete latent variables in BN. In summary, there is a shortage of well-studied model diagnostic techniques for the BN framework (Crawford, 2014; Hu & Templin, 2020).

BN is a relatively novel type of measurement model for the psychometric community. To the author's knowledge, there is no unified model criticism procedure that should be applied to the BN measurement model. As presented in the following sections, the authors pay attention to different aspects of model fit, such as model identification, item or global fit (Almond et al., 2015; Culbertson, 2016). The systematization of BN model criticism approaches is the motivation of the study.

Methods

A systematic review of model criticism approaches

The purpose of this study is to identify the state-of-the-art, related to model criticism techniques within Bayesian networks in educational assessment. In order to conduct the analysis, the literature discussing BN as a measurement model in educational assessment was searched. Searching the literature included two iterations. The first one looked through the Web of Science and Scopus databases and selected relevant articles. In the second iteration, references of these papers were additionally scanned, and relevant articles were selected. After that, the content analysis of the selected articles was conducted.

There were three key inclusion criteria in the study. The first criterion was that the papers should be published in English. The second criterion was that studies should be published as an article, conference paper, book or book chapter, or a chapter of a dissertation with open access. The third criterion was that the content of the studies was relevant: the studies are related to the area of educational or psychological research and performed analysis of simulated or/and real assessment datasets. The studies should discuss the psychometric analysis of the data and present model criticism analysis within BN. Overall, 25

[1] studies were analyzed in the research. The following section will discuss the directions of model criticism approaches. The summary of the systematization is presented in Appendix A

Results

The selected articles were analyzed and categorized into seven groups, based on the techniques of model criticism within the BN framework: a) classification and prediction accuracy; b) model comparisons; c) mutual information; d) inspection of conditional probability tables; e) residual analysis; f) posterior predictive model checking, and g) correlation with external criteria. Because one study may include different aspects of model criticism, one article can be assigned to several groups.

Classification and predication accuracy

In the context of educational assessment, classification and prediction accuracy demonstrates the ability of a model to categorize and predict the level of students' proficiency.

The basic concept of the prediction accuracy approach is that if the model fits the data well, the model can precisely predict the state of the parameter. In the study by Williamson, Almond, & Mislevy (2000), the accuracy of predictions was considered as the criteria for model fit assessment. The study investigated if different indices, such as Ranked Probability Score (RPS; Epstein, 1969), Weaver's Surprise Index (WSI; Weaver, 1948), and Good's Logarithmic Score (GLS; Good, 1952), are helpful for the detection of errors. Vomlel (2004) compared the prediction accuracy (expected decision error) of different models, including computer-adaptive and non-adaptive models. The criterion demonstrates if the model correctly predicts the state of students' skills. The statistic that was used is the percentage of cases that were predicted equivalently to the observed data.

Xing, Li, Chen, Huang, Chao, Massicotte et al. (2020) investigated the prediction accuracy of engineering design process assessment by comparison of the results provided by the assessment system and students' annual energy output. In the study, an external criterion (students' annual energy output) serves as the basis for the decision about the usefulness of BN. The authors computed precision and recall indexes. Precision demonstrates the correctness of the identification of students' labels; recall reflects the correctness of the identification of true cases. The indexes reveal the percentage agreement separately for high and low-performing groups. In the research by Pardos, Heffernan, Anderson, and Heffernan (2007) and Pardos, Feng, Heffernan, and Linquist-Heffernan (2007), model criticism of mathematics test was realized based on the analysis of the prediction accuracy through estimation of the absolute difference between predicted and observed score.

The classification accuracy approach is based on a similar idea that if the model fits the data well, it will correctly classify students to their proficiency levels. Almond (2015) discusses the application of two classification accuracy coefficients: Goodman and Kruskal's lambda (Goodman & Kruskal, 1954) and Cohen's $\varkappa$ (Cohen, 1960). In the study by Lee and Corter (2011), firstly, classification accuracy was checked for both misconceptions and skills, based on the percentage agreement between the predicted and observed parameters and Cohen's k statistic. Secondly, the classification accuracy coefficient (percentage agreement) was applied to investigate misconception patterns. In a study based on simulations, Rutstein (2012), applied classification accuracy analysis to detect if a model had correctly assessed the level of learning progress. The analysis was based on the percentage agreement and the adjusted Rand index (Steinley, 2004).

The diversity of classification and prediction accuracy statistics is wide; therefore, a researcher should be highly attentive in choosing one or several of them. For instance, as the criterion of accuracy, a straightforward percentage agreement analysis and special criteria were applied. However, in the article by Lee and Corter (2011), the drawbacks of percentage agreement were discussed. Cohen's k statistics are suggested as a more reliable index because it adjusts the percentage agreement for the probability of agreement by chance (Lee and Corter, 2011). Almond (2015)

[1] Following the described criteria, 23 articles were selected. Two articles were added following anonymous reviewer recommendation.

highlighted that Goodman and Kruskal's lambda and Cohen's k statistics answer different questions. Cohen's k is a statistic that estimates agreement between rates; Goodman and Kruskal's Lambda answers where the estimates of the state are more precise if the classification was applied.

In the study by Williamson, Almond, & Mislevy (2000), the primary focus was on the comparison of different prediction indices (RPS, WSI, GLS). However, the indices were not presented in other studies of the current literature review. The investigation of the efficiency of discrepancy measures within the PPMC approach by Crawford (2014) demonstrated that RPS and GLS were unable to detect any model misspecification. Additional investigations of the properties of different criteria in the context of BN application in education appeared to be helpful.

Model comparisons

The studies of the second group are unified by an application of information criteria. The articles are based on the comparison of likelihood estimates and implement information criteria, such as AIC (Akaike information criterion; Akaike, 1973), BIC (Bayesian information criterion; Schwarz, 1978), or DIC (Deviance information criterion; Spiegelhalter, Best, Carlin, & van der Linde, 2002) for model selection.

West, Rutstein, Mislevy, Liu, Choi, Levy et al. (2010) applied BIC and the bootstrapped likelihood ratio test (BLRT) to compare learning progress models in the assessment system. Lee and Corter (2011) compared models for the diagnosis of mathematical skills and misconceptions based on AIC and BIC. Song, Wang, Dai, and Ding (2018), analyzed fraction subtraction data in terms of comparison of models with different hierarchical structures, based on −2log-likelihood (−2LL) and DIC (Celeux, Forbes, Robert, & Titterington, 2006). Rutstein (2012) applied AIC, BIC, and DIC to compare models with different structures. Also, DIC was used in a parameter recovery study by Almond, Yan, and Hemat (2008). Sinharay and Almond (2007) applied DIC to compare models with different numbers of latent classes for the mixed number subtraction example.

There is a shortage of studies investigating which information criteria should be chosen for specific circumstances. In the article by Sinharay and Almond (2007), it was noticed that DIC is similar to AIC if the priors are noninformative and might be preferable if MCMC is applied.

Mutual information

In the studies of the third group, the analysis of Mutual Information (MI) values and the visualization of the Weight of Evidence (WOE), evidence balance sheets, help to "debug" assessment systems. MI shows the degree of association between two variables and helps to indicate their utility. For instance, a high level of MI between latent and observable variables shows that the observable indicator sustainably represents latent proficiency (Almond, 2015).

Similarly, WOE (Good, 1985), and its visualization evidence balance sheet (Madigan, Mosurski, & Almond, 1997), helps to identify the most influential observable variables for latent variables assessment. The WOE method demonstrates the amount of information provided by the observable indicator to estimate the level of proficiency. This method was applied for the analysis of the functioning of indicators within Physics Playground (Newton's Playground), NetPass, and Math Word Problem assessment (Almond, Kim, Shute, & Ventura, 2013; Almond, Kim, Velasquez, & Shute, 2014; Almond, Mulder, Hemat, & Yan, 2009; Kim, Almond, & Shute, 2016).

Inspection of conditional probability tables

In the articles from the fourth group, item functioning is analyzed based on the inspection of CPTs. It is based on the idea that the probability of a correct answer should be higher if a student has mastered a skill. CPTs inspection shows whether items can discriminate between students with different proficiency levels. This approach was applied by West et al. (2010) and Kim et al. (2016). DiCerbo, Xu, Levy, Lai, and Holland (2017) compared prior and posterior values of CPTs and additionally analyzed the values of CPTs of latent variables.

Residual analysis

The articles of the fifth group focus on the analysis of the residuals. In the study by Almond et al. (2009), the impact of the common stimulus that bound a set of items was investigated. Mantel-Haenszel statistic was applied to verify if the local independence across observable variables holds. Pardos, Feng, Heffernan, and Linquist-Heffernan (2007) apply Bayesian test item residuals for model criticism analysis. In the study, the

residual is the average of the differences between students' observed and predicted responses for an item.

The study of Sinharay and Almond (2007) focuses on the analysis of the difference between observed and predicted values based on the raw scores. According to the authors' interpretation, if there are too many items with high absolute values of residuals, the model should be considered poor. However, it was revealed that this approach serves to detect only an extreme level of a model misfit. Visualization techniques for this analysis were presented and discussed as a fruitful way to catch the sources of the misfit.

Item fit analysis was based on the comparison between observed and predicted proportion correct answers for items within a given proficiency level. The comparison includes graphical representation and $\chi 2$-type test statistics. Despite promising results, the authors indicated that the power of $\chi 2$-type test statistics is under-investigated and also mentioned that an approach based on Posterior Predictive Model Checking (PPMC) methods is more promising (see the section PPMC below).

Posterior predictive model checking

The sixth group included studies that applied the PPMC method (Gelman, Meng, & Stern, 1996). PPMC compares observed (realized) and model-expected simulated (posterior predictive) data with respect to a discrepancy measure. The discrepancies between realized and posterior predictive values indicate that the model is incapable of describing the observed data. One of the features of PPMC is that it provides a reference distribution of the discrepancy measure empirically from the draws of the model parameters from the posterior distribution.

Sinharay (2006) describes the model criticism approach, which includes PPMC for the analysis of item fit and inter-item association. The former analysis included three types of discrepancy measures. The first discrepancy measure characterizes the association between number-correct scores and item scores (point biserial correlations). The second discrepancy measure is based on examinees' equivalent class memberships and demonstrates the proportion of students in an equivalent class answering an item correctly. Equivalent classes represent the patterns of students' responses. The third discrepancy measure is based on observed raw scores, that form groups of students. The discrepancy measure demonstrates the proportion of students in a raw score group answering an item correctly.

The later analysis provides information about the sources of local independence violations based on the odds ratio as a discrepancy measure. The comparison between observed and predicted data was also conducted via $\chi 2$-type and G-type test statistics and a graphical method called Direct Data Display.

The utility of PPMC for model checking and, more precisely, the sensitivity of different discrepancy measures to local independence violation was investigated by Levy (2006). Six directional discrepancy measures (covariance, residual covariance, log odds ratio, standardized log odds ratio residual, the model-based covariance, and Q3 (Yen, 1984)) demonstrated the sources of local dependence and, therefore, provide necessary information for model improvement.

The PPMC approach for BN model checking was studied in more detail by Crawford (2014). The author applied discrepancy measures aimed at detecting misfit at the global, item, and person levels. The author examined the sensitivity of discrepancy measure to inadequately modeled dimensionality and the number of latent classes. Following the results of previous research, thirteen discrepancy measures were selected. As a result, standardized generalized dimensionality discrepancy measure (SGDDM; Levy, Xu, Yel, & Svetina 2015) and Q3 performed reasonably well to detect either local independence violation and the misconceptions of the number of latent classes.

The study by Rupp, Levy, DiCerbo, Sweet, Crawford, Calico, et. al. (2012) was focused on the investigation of the psychometric quality of the assessment tool for measuring the proficiency of design, configuration, and troubleshooting computer networks. In the article, different discrepancy measures within PPMC were applied; among them are univariate proportions correct; Q3; a marginal generalized dimensional discrepancy measure (GDDM; Levy & Svetina, 2011). The model criticism analysis helped to find out and clarify the sources of item misfit. More precisely, based on the values of Q3, the sources of local dependence between observable indicators were detected. To verify the scoring structure, proportion correct values and GDDM were applied, and it was

revealed that the model should be redefined by subject-matter experts.

The study by Levy and Mislevy (2016) demonstrates an unusual application of PPMC to verify a hard prerequisite relationship. The study investigated whether it is necessary to possess one skill in advance of possessing another. The authors analyzed the frequency of cases when students performed successfully on one skill and poorly on the skill that serves as a prerequisite. After that, the authors compared the realized and posterior predictive values.

The comparison with posterior predictive values serves as a reference to decide to what extent the deviation from the theoretically expected pattern leads to the intended results.

The studies of the section investigated the usefulness of different discrepancy measures, as well as the benefits and drawbacks of PPMC. Among the benefits of the PPMC are its theoretical basis, sensitivity to the uncertainty of parameter estimation, and flexibility.

Sinharay (2006) calls the method intuitive and straightforward. However, it should be noticed, PPMC is a conservative method, rather than restrictive, in terms of the tendency to not detect a misfit if it exists, rather than detect misfit if it is absent. Also, a significant feature of PPMC is the intensity of computation. Because PPMC goes along with MCMC, it might take hours to estimate and store the necessary parameter estimates (Crawford, 2014; Sinharay, 2006).

Talking about the choice of the discrepancy measure, Crawford (2014) concluded that a combination of different discrepancy measures is needed to provide useful and interpretable characteristics of model-data fit and recognize the source of the misfit. Rupp et al. (2012) mentioned that the choice of the discrepancy measure appeared as a problematic point.

Despite the advantages of PPMC as a model criticism technique, it seems like a challenge for methodologists to make it more widespread in educational assessment since it is necessary not only to discover proper ways of model criticism but also to make it convenient for practitioners.

Correlation with external criteria

Finally, the seventh group was formed by studies that focused on the validation of the inferences about students, gathered through BN application as a measurement model. The studies analyze the correlation of the BN inferences about students with an external criterion, such as the results of another test, expert opinion, or the raw score.

De Klerk, Eggen, and Veldkamp (2016) analyze the correlation between the results of the multimedia-based performance assessment tool, gathered through BN models, and the raw score. DiCerbo et al. (2017), in the field of the geometric measurement of area, synthesizing the information in the data with prior beliefs provided by experts.

Shute, Wang, Greiff, Zhao, and Moore (2016), validate the inference about students' performance in game-based problem-solving assessment by analyzing the correlation of these results with the results of Raven's Progressive Matrices. Shute and Moore (2017) investigated the validity of the physics understanding assessment, realized via the game-based simulation, by the comparison with the results of an external physics test. An analysis, which included the investigation of the correlation between the results given by the physics understanding assessment and combined pretest and posttest results, was also presented by Almond (2015).

Conclusion

The literature review was focused on the model criticism of BN, as a measurement model, in the field of educational assessment. BN provides an opportunity for the flexible modeling of complex educational assessment data. However, BN is a novel psychometric approach, and not all aspects of its application are well-known. In the context of model criticism, there are fruitful and beneficial studies that demonstrate useful model criticism approaches; nevertheless, more research in the area is still required.

The review revealed the diversity of model criticism procedures applied in educational assessment studies. It was demonstrated that model criticism for BN does not appear as a unified framework. In the review, model criticism was considered in a broad sense: the way of model checking was not a criterion for either inclusion or exclusion of the papers. The

results demonstrated that the studies focused on the model comparison, quality of inference about students, item fit, and global fit, whereas person fit analysis appeared underrepresented. The WOE helps to recognize unexpected patterns in student behavior; however, this approach focuses on investigating poor indicator functioning. Person-fit statistics were briefly discussed by Almond et al. (2015), Crawford (2014), Levy and Mislevy (2016), and Mislevy, Senturk, et al. (2002).

Almond et al. (2015), Levy and Mislevy (2016) conducted PPMC with the root mean square of the squared Pearson residuals as a discrepancy measure. Mislevy, Senturk, et al. (2002) also conducted PPMC with the fit mean square measure which is based on the difference between observed and expected responses. However, the authors highlighted that "a more serious analysis would run simulations to characterize the specificity and the sensitivity of fit indices constructed in this manner" (Mislevy, Senturk, et al., 2002, p. 28). Crawford (2014) applied Hierarchy Consistency Index (HCI) as a discrepancy measure which demonstrated promising results for person misfit identification.

The issues related to differential item functioning (an issue when the test behaves differently for different groups of the population, DIF) were discussed by Almond et al. (2015) and Sinharay (2006). Almond et al. (2015) suggest that if there is no DIF in the model, the introduction of a subpopulation membership variable should not affect the probability of correct response. Sinharay (2006) conducted PPMC with Mantel–Haenszel test as a discrepancy measure. The author highlighted that PPMC provides a benefit for researchers because it helps to obtain matched groups with respect to students` latent skills. DIF-detection and person-fit analysis are underrepresented among modern psychometric studies that apply BN as a measurement model, and more methodological research in this area is needed.

Another approach to model discrete latent and observable variables applied in modern psychometrics is a cognitive diagnostic model (CDM). CDM is a family of measurement models that provide a specific person classification based on the performance in particular domains of a target construct. Almond and Zapata-Rivera (2019) postulated that CDM could be represented as BN.

Sessoms and Henson (2018) demonstrated that across studies that applied CDM, model criticism was based primarily on relative fit indices (e.g., AIC), overall fit indices (e.g., HCI), and item fit. The authors highlighted that an important direction for CDM model criticism is the development of guidelines for model fit indexes. This direction is also relevant for BN because there is a shortage of guidelines and simulation studies that provide practitioners with recommendations about applying fit indexes.

Furthermore, it was shown that approaches suitable for model criticism for CDM could be successfully applied for model criticism for BN (Sinharay and Almond, 2007). Recently, Hu and Templin (2020) demonstrated a similarity between BN and CDM in parametrizations of the hierarchical relationship and proposed an approach of model fit analysis for BN. According to the results, the likelihood ratio test is helpful to conduct a model comparison analysis between models with differences in the hierarchical relationships between latent variables.

A promising extension of BN in modern psychometrics is a Dynamic Bayesian network (DBN; Almond et al., 2015; Reye, 2004). This model allows researchers to conduct an analysis considering dependencies between time segments. For instance, the DBN captures students` attempts to solve the task and provides detailed feedback to students immediately during the assessment procedure.

DBN is considered a perspective direction for future research (Xing et al., 2020) and has been successfully applied in educational assessment (Levy, 2019). However, the questions about the investigation of the psychometric properties of DBN remain open, and the model criticism of DBN is one of them (Reichenberg, 2018a). Reichenberg (2018a, 2018b) highlighted that there are not enough studies that investigate the adequacy of the BN model criticism methods in application to DBN. However, it provides an opportunity for researchers to investigate novel model criticism methods specifically for DBN. For instance, Reichenberg (2018a) postulates that the development of new discrepancy measures for PPMC analysis is a promising direction.

A special case of DBN is Bayesian Knowledge Tracing (BKT), a general approach in intelligent tutoring systems (Käser et al., 2014). BKT describes

the probability of transition from the state of an unknown skill to the state of mastered skill during skill acquisition (Corbett & Anderson, 1995). However, model criticism of BKT has its unique features, such as probable semantic model degeneracy, which is the issue that arises when estimated parameters that fit the data well are inconsistent with the assumptions of the model (Doroudi & Brunskill, 2017).

Moreover, the investigation of the relationship between BKT and Performance Factors Analysis (PFA) is considered a promising research direction (Galyardt & Goldin, 2015). PFA describes student learning progression based on correct and incorrect responses within the logistic regression approach (Pavlik Jr et al., 2009). Model criticism of PFA (particularly, Recent Performance Factors Analysis model, R-PFA) was conducted via stratified cross-validation analysis, AIC, and visualization technique Viz-R, which is similar to a confusion matrix, but also takes into account the distance between observed and predicted data (Galyardt & Goldin, 2015; Goldin & Galyardt, 2015).

Overall, discussing the details of model criticism of student learning models, Pelánek (2018) postulate that improvement of evaluation standards is highly demanded. Remarkably, the author highlighted that for student models such as BKT or PFA, several aspects of model criticism are needed to be conducted and reported carefully: the way of data collection, the approach to data splitting for cross-validation, the choice and computation of accuracy statistics (Pelánek, 2018).

The results of the review serve to provide systematization of model criticism methods for BN used in the educational assessment research and to discuss related areas and directions for future research. The review intended to navigate practitioners and researchers in the area of BN model criticism, assuming that it helps not to miss the benefits of the application of BN as a flexible and convenient measurement model. However, one of the limitations of the study is that despite two iterations in the literature search process, any research might be omitted, and the diversity of model criticism approaches might be underrepresented.

To conclude, the primary focus of educational assessment is to make valid inferences about student characteristics, which is possible if a proper measurement is applied. Almond and colleagues postulated that "a requisite for valid, high quality and effective assessment is harmony between the substantive theory that underlies the conceptual student model and the formal probability model supporting the assessment." (2007, p. 355). The measurement model represents the underlying structure of cognitive processes, and the model fit analysis can provide insights into it. Moreover, model criticism analysis helps us realize if observable indicators of an assessment tool represent latent proficiencies, which may shed light on the assessment tool's quality and help improve it. The role of model criticism in modern educational research is emphasized by the fact that many articles consider model fit analysis to be a self-sustained direction of future research (de Klerk et al., 2016; DiCerbo et al., 2017; Levy & Mislevy, 2004; Rutstein, 2012; West, Rutstein, Mislevy, Liu, Levy, Dicerbo et al., 2012). BN is a promising measurement model in modern educational assessment; therefore, the model criticism techniques deserve more attention.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov F. C'aki (Eds.), *Proceedings of the 2nd international symposium on information theory*. Budapest: Akademiai Kiado, 267–281.

Almond, R. G. (2015). Tips and Tricks for Building Bayesian Networks for Scoring Game- Based Assessments. In J. Suzuki and M. Ueno (Eds.) *Advanced Methodologies for Bayesian Networks. AMBN 2015. Lecture Notes in Computer Science, vol 9505*. Springer, Cham. https://doi.org/10.1007/978-3-319-28379-1_18

Almond, R. G., DiBello, L. V., Moulder, B., & Zapata-Rivera, J. D. (2007). Modeling Diagnostic Assessments with Bayesian Networks. *Journal of Educational Measurement, 44*(4), 341–359. https://doi.org/10.1111/j.1745-3984.2007.00043.x

Almond, R. G., Kim, Y. J., Shute, V. J., & Ventura, M. (2013). Debugging the evidence chain. In 2013 UAI Application Workshops: *Big Data meet Complex Models and Models for Spatial, Temporal and Network Data (Association for Uncertainty in Artificial Intelligence)*.

Almond, R. G., Kim, Y. J., Velasquez, G., & Shute, V. J. (2014). How Task Features Impact Evidence From Assessments Embedded in Simulations and Games. Measur*ement: Interdisciplinary Research and Perspectives, 12*(1–2), 1–33. https://doi.org/10.1080/15366367.2014.910060

Almond, R. G., Mislevy, R. J., Steinberg, L. S., Yan, D., & Williamson, D. M. (2015). *Bayesian Networks in Educational Assessment*. Springer New York. https://doi.org/10.1007/978-1-4939-2125-6

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. https://doi.org/10.3102/1076998609332751

Almond, R. G., Yan, D., & Hemat, L. (2008). Parameter Recovery Studies With a Diagnostic Bayesian Network Model. *Behaviormetrika, 35*(2), 159–185. https://doi.org/10.2333/bhmk.35.159

Almond, R. G., & Zapata-Rivera, J. D. (2019). Bayesian Networks. In *Handbook of Diagnostic Classification Models* (pp. 81-106). Springer, Cham.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Becker, B. J., & Shute, V. J. (Eds.). (2010). *Innovative Assessment for the 21st Century: Supporting Educational Needs*. Springer.

Celeux, G., Forbes, F., Robert, C. P., & Titterington, D. M. (2006). Deviance information criteria for missing data models. *Bayesian Analysis, 1*(4), 651–673.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46.

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.

Crawford, A. (2014). *Posterior predictive model checking in Bayesian networks* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database (UMI No. 3618300). Arizona State University, Tempe, AZ.

Cruz, N., Desai, S. C., Dewitt, S., Hahn, U., Lagnado, D., Liefgreen, A., Phillips, K., Pilditch, T., & Tešić, M. (2020). Widening Access to Bayesian Problem Solving. *Frontiers in Psychology, 11*. https://doi.org/10.3389/fpsyg.2020.00660

Culbertson, M. J. (2016). Bayesian Networks in Educational Assessment: The State of the Field. *Applied Psychological Measurement, 40*(1), 3–21. https://doi.org/10.1177/0146621615590401

De Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.

De Klerk, S., Eggen, T. J. H. M., & Veldkamp, B. P. (2016). A methodology for applying students' interactive task performance scores from a multimedia-based performance assessment in a Bayesian Network. *Computers in Human Behavior, 60*, 264–279. https://doi.org/10.1016/j.chb.2016.02.071

De Klerk, S., Veldkamp, B. P., & Eggen, T. J. H. M. (2015). Psychometric analysis of the performance data of simulation-based assessment: A systematic review and a Bayesian network example. *Computers & Education, 85*, 23–34. https://doi.org/10.1016/j.compedu.2014.12.020

DiCerbo, K. E., Bertling, M., Stephenson, S., Jia, Y., Mislevy, R. J., Bauer, M., & Jackson, G. T. (2015). An application of exploratory data analysis in the development of game- based assessments. In *Serious games analytics* (pp. 319-342). Springer, Cham.

DiCerbo, K. E., Xu, Y., Levy, R., Lai, E., & Holland, L. (2017). Modeling Student Cognition in Digital and Nondigital Assessment Environments. *Educational Assessment, 22*(4), 275–297. https://doi.org/10.1080/10627197.2017.1382343

Doroudi, S., & Brunskill, E. (2017). The Misidentified Identifiability Problem of Bayesian Knowledge Tracing. *International Conference on Educational Data Mining, EDM2017*.

Epstein, E. S. (1969). A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology, 8*(6), 985–987.

Galyardt, A., & Goldin, I. (2015). Move Your Lamp Post: Recent Data Reflects Learner Knowledge Better than Older Data. *Journal of Educational Data Mining, 7*(2), 83–108.

Goldin, I., & Galyardt, A. (2015). Viz-r: Using recency to improve student and domain models. Proceedings of the Second (2015) *ACM Conference on Learning@Scale*, 417– 420.

Good, I. J. (1952). Rational Decisions. *Journal of the Royal Statistical Society Series B (Methodological), 14*(1), 107–114. https://doi.org/10.1111/j.2517-6161.1952.tb00104.x

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

Goodman, L. A., & Kruskal, W. H. (1954). Measures of Association for Cross Classifications. *Journal of the American Statistical Association, 49*(268), 732–764. https://doi.org/10.1080/01621459.1954.10501231

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior Predictive Assessment Of Model Fitness Via Realized Discrepancies. *Statistica Sinica, 6*(4), 733–760.

Hu, B., & Templin, J. (2020). Using Diagnostic Classification Models to Validate Attribute Hierarchies and Evaluate Model Fit in Bayesian Networks. *Multivariate Behavioral Research, 55*(2), 300–311. https://doi.org/10.1080/00273171.2019.1632165

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55.

Käser, T., Klingler, S., Schwing, A. G., & Gross, M. (2014). Beyond knowledge tracing: Modeling skill topologies with Bayesian networks. *International Conference on Intelligent Tutoring Systems*, 188–198.

Kim, Y. J., Almond, R. G., & Shute, V. J. (2016). Applying Evidence-Centered Design for the Development of Game-Based Assessments in Physics Playground. *International Journal of Testing, 16*(2), 142–163. https://doi.org/10.1080/15305058.2015.1108322

Lee, J., & Corter, J. E. (2011). Diagnosis of Subtraction Bugs Using Bayesian Networks. *Applied Psychological Measurement, 35*(1), 27–47. https://doi.org/10.1177/0146621610377079

Levy, R. (2006). *Posterior predictive model checking for multidimensionality in item response theory and Bayesian networks*. (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database (UMI No. 3212601). University of Maryland, College Park.

Levy, R. (2009). The Rise of Markov Chain Monte Carlo Estimation for Psychometric Modeling. *Journal of Probability and Statistics*. https://doi.org/10.1155/2009/537139

Levy, R. (2013). Psychometric and Evidentiary Advances, Opportunities, and Challenges for Simulation-Based Assessment. *Educational Assessment, 18*(3), 182–207. https://doi.org/10.1080/10627197.2013.814517

Levy, R. (2019). Dynamic Bayesian Network Modeling of Game-Based Diagnostic Assessments. *Multivariate Behavioral Research, 54* (6), 771-794. https://doi.org/10.1080/00273171.2019.1590794

Levy, R., & Mislevy, R. (2004). Specifying and Refining a Measurement Model for a Computer-Based Interactive Assessment. *International Journal of Testing, 4*(4), 333– 369. https://doi.org/10.1207/s15327574ijt0404_3

Levy, R., & Mislevy, R. J. (2016). *Bayesian psychometric modeling*. CRC Press.

Levy, R., & Svetina, D. (2011). A generalized dimensionality discrepancy measure for dimensionality assessment in multidimensional item response theory. *British Journal of Mathematical and Statistical Psychology, 64*(2), 208-232.

Levy, R., Xu, Y., Yel, N., & Svetina, D. (2015). A Standardized Generalized Dimensionality Discrepancy Measure and a Standardized Model - Based Covariance for Dimensionality Assessment for Multidimensional Models. *Journal of Educational Measurement, 52*(2), 144–158. https://doi.org/10.1111/jedm.12070

Madigan, D., Mosurski, K., & Almond, R. G. (1997). Graphical explanation in belief networks. *Journal of Computational and Graphical Statistics, 6*(2), 160–181.

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika, 59*(4), 439–483. https://doi.org/10.1007/BF02294388

Mislevy, R. J., Almond, R. G., Yan, D., & Steinberg, L. S. (1999). Bayes nets in educational assessment: Where the numbers come from. In K. B. Laskey and H. Prade (Eds.), *Uncertainty in artificial intelligence: Proceedings of the15th conference* (pp. 437–446). San Francisco: Morgan Kaufmann.

Mislevy, R. J., Senturk, D., Almond, R. G., Dibello, L. V., Jenkins, F., Steinberg, L. S., & Yan, D. (2002). *Modeling conditional probabilities in complex educational assessments*. Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies, University of California, Los Angeles.

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G., & Johnson, L. (2002). Making sense of data from complex assessments. *Applied Measurement in Education, 15*(4), 363-389.

Neapolitan, R. E. (2004). *Learning Bayesian networks*. Englewood Cliffs: Prentice Hall.

Pardos, Z. A., Feng, M., Heffernan, N. T., & Linquist-Heffernan, C. (2007). Analyzing fine- grained skill models using Bayesian and mixed effects methods. *Educational Data Mining*, 50–59.

Pardos, Z. A., Heffernan, N. T., Anderson, B., & Heffernan, C. L. (2007). The effect of model granularity on student performance prediction using Bayesian networks. *International Conference on User Modeling*, 435–439.

Pavlik Jr, P. I., Cen, H., & Koedinger, K. R. (2009). Performance Factors Analysis – A New Alternative to Knowledge Tracing. In V. Dimitrova & R. Mizoguchi (Eds.), *Proceedings of the 14th International Conference on Artificial Intelligence in Education*. Brighton, England.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo: Morgan Kaufmann.

Pelánek, R. (2018). The details matter: Methodological nuances in the evaluation of student models. *User Modeling and User-Adapted Interaction, 28*(3), 207–235. https://doi.org/10.1007/s11257-018-9204-y

Reichenberg, R. (2018a). Dynamic Bayesian Networks in Educational Measurement: Reviewing and Advancing the State of the Field. *Applied Measurement in Education, 31*(4), 335–350. https://doi.org/10.1080/08957347.2018.1495217

Reichenberg, R. (2018b). *The Impact of Information Quantity and Quality on Parameter Estimation for a Selection of Dynamic Bayesian Network Models with Latent Variables*. ISBN 978-0-438-28760-0. Ph.D., Arizona State University. https://search.proquest.com/pqdtglobal/docview/2099139791/abstract/B867C17BE61E 43B5PQ/2

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

Rupp, A., Levy, R., Dicerbo, K. E., Sweet, S. J., Crawford, A. V., Calico, T., Benson, M., Fay, D., Kunze, K. L., Mislevy, R. J., & Behrens, J. T. (2012). Putting ECD into Practice: The Interplay of Theory and Data in Evidence Models within a Digital Learning Environment. Journ*al of Educational Data Mining, 4*(1), 49–110.

Rutstein, D. W. (2012). *Measuring Learning Progressions Using Bayesian Modeling in Complex Assessments*. Retrieved from ProQuest Dissertations and Theses database (UMI No. 3517494). University of Maryland, College Park.

Sessoms, J., & Henson, R. (2018). Applications of Diagnostic Classification Models: A Literature Review and Critical Commentary. *Measurement: Interdisciplinary Research and Perspectives, 16*, 1–17. https://doi.org/10.1080/15366367.2018.1435104

Scalise, K., & Clarke ‐ Midura, J. (2018). The many faces of scientific inquiry: Effectively measuring what students do and not only what they say. *Journal of Research in Science Teaching, 55*(10), 1469 ‐ 1496. https://doi.org/10.1002/tea.21464

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics, 6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Shute, V. J., & Moore, G. R. (2018). Consistency and validity in game-based stealth assessment. In H. Jiao & R. W. Lissitz (Eds.), *Technology enhanced innovative assessment: Development, modeling, and scoring from an interdisciplinary perspective* (pp. 31-51). Charlotte, NC: Information Age Publisher.

Shute, V. J., & Wang, L. (2016). Assessing and Supporting Hard ‐ to ‐ Measure Constructs in Video Games. *The Wiley Handbook of Cognition and Assessment* (pp. 535 ‐ 562). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118956588.ch22

Shute, V.J., Wang, L., Greiff, S., Zhao, W., & Moore, G. (2016). Measuring problem solving skills via stealth assessment in an engaging video game. Computers in Human Behavior, 63, 106–117. https://doi.org/10.1016/j.chb.2016.05.047

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

Sinharay, S., & Almond, R. G. (2007). Assessing Fit of Cognitive Diagnostic Models A Case Study. *Educational and Psychological Measurement, 67*(2), 239–257. https://doi.org/10.1177/0013164406292025

Song L., Wang W., Dai H., & Ding S. (2018) Bayesian Network for Modeling Uncertainty in Attribute Hierarchy. In M. Wiberg,, S. Culpepper, R. Janssen, J. González, & D. Molenaar. (eds) *Quantitative Psychology. IMPS 2017. Springer Proceedings in Mathematics & Statistics, vol 233*. Springer, Cham.

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*(4), 583–639. https://doi.org/10.1111/1467-9868.00353

Steinley, D. (2004). Properties of the Hubert-Arable Adjusted Rand Index. *Psychological Methods, 9*(3), 386.

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

Weaver, W. (1948). Probability, rarity, interest, and surprise. *The Scientific Monthly, 67*(6), 390–392.

West, P., Rutstein, D. W., Mislevy, R. J., Liu, J., Choi, Y., Levy, R., Crawford, A., DiCerbo, K. E., Chappel, K., & Behrens, J. T. (2010). *A Bayesian Network Approach to Modeling Learning Progressions and Task Performance.* CRESST Report 776. National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

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., Almond, R. G., & Mislevy, R. J. (2000). Model criticism of Bayesian networks with latent variables. *Uncertainty in Artificial Intelligence Proceedings 2000*, 634–643.

Xing, W., Li, C., Chen, G., Huang, X., Chao, J., Massicotte, J., & Xie, C. (2020). Automatic Assessment of Students' Engineering Design Performance Using a Bayesian Network Model. *Journal of Educational Computing Research*. https://doi.org/10.1177/0735633120960422

Yan, D., Almond, R., & Mislevy, R. (2004). *A Comparison of Two Models for Cognitive Diagnosis*. ETS Research Report Series, 2004(1), i–33. https://doi.org/10.1002/j.2333- 8504.2004.tb01929.x

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*(2), 125– 145.

Appendix A

Authors, PublicationYear	Title of the Article	Name of the Instrument	Software	Method of Parameter Estimation	Model Checking
Almond, 2015	Tips and Tricks for Building Bayesian Networks for Scoring Game-Based Assessments	Physics Playground (Newton's Playground)	R-Netica	EM-algorithm	Classification and prediction accuracy; Mutual Information; Correlation with external criteria
Almond, Kim, Shute & Ventura, 2013	Debugging the Evidence Chain	Physics Playground (Newton's Playground)	Not presented	Not presented	Mutual Information
Almond, Kim,Velasquez, & Shute, 2014	How Task Features Impact Evidence From Assessments Embedded in Simulations and Games	Math WordProblems; Physics Playground (Newton's Playground)	Not presented	Not presented	Mutual Information
Almond, Mulder, Hemat, & Yan, 2009	Bayesian Network Models for Local Dependence among Observable Outcome Variables.	NetPass	StatShop	MCMC	Mutual Information; Residual analysis
Almond, Yan & Hemat, 2008	Parameter Recovery Studies with a Diagnostic Bayesian Network Model	Information and Communication Technology (ICT) Literacy Assessment Simulation study	StatShop	MCMC	Model Comparisons
Crawford, 2014	Posterior Predictive Model Checking in Bayesian Networks	Simulation study	WinBUGS	MCMC	Posterior predictive model checking
De Klerk, Eggen, & Veldkamp, 2016	A Methodology for Applying Students' Interactive Task Performance Scores From a Multimedia-Based Performance Assessment in a Bayesian Network	Confined spaceguard (CSG) students' assessment	GeNle	Clustering algorithm	Correlation with external criteria

DiCerbo, Xu, Levy, Lai, & Holland, 2017	Modeling Student Cognition In Digital and Nondigital Assessment Environments	The Alice in Arealand game	WinBUGS	MCMC	Inspection of conditional probability tables; Correlation with external criteria
Kim, Almond, & Shute, 2016	Applying Evidence-Centered Design for the Development of Game-Based Assessments in Physics Playground	Physics Playground (Newton's Playground)	R-Netica	Gradient descent algorithm	Mutual Information; Inspection of conditional probability tables
Lee & Corter, 2011	Diagnosis of Subtraction Bugs Using Bayesian Networks	Study of Subtraction Bugs (VanLehn, 1981)	HUGIN	Not presented	Classification and prediction accuracy; Model comparisons
Levy, 2006	Posterior Predictive Model Checking for Multidimensionality in Item Response Theory and Bayesian Networks	Simulation study	WinBUGS	MCMC	Posterior predictive model checking
Levy & Mislevy, 2016	Bayesian Psychometric Modeling	Mixed number subtraction example	WinBUGS	MCMC	Posterior predictive model checking
Pardos, Feng, Heffernan, & Linquist- Heffernan, 2007	Analyzing Fine-Grained Skill Models Using Bayesian and Mixed Effects Methods	Massachusetts Comprehensive Assessment System	Not presented	Not presented	Classification and prediction accuracy; Residual analysis
Pardos, Heffernan, Anderson, & Heffernan, 2007	The Effect of Model Granularity on Student Performance Prediction Using Bayesian Networks	Massachusetts Comprehensive Assessment System	Bayes Net Toolkit for MATLAB	Not presented	Classification and prediction accuracy
Rupp, Levy, Dicerbo, Sweet, Crawford, Calico, Benson, Fay, Kunze, Mislevy, & Behrens, 2012	Putting ECD into Practice: The Interplay of Theory and Data in Evidence Models within a Digital Learning Environment	Packet Tracer	WinBUGS	MCMC	Posterior predictive model checking
Rutstein, 2012	Measuring Learning Progressions Using Bayesian Modeling in Complex Assessments	Learning Progressions' Assessment	WinBUGS	MCMC	Classification and prediction accuracy; Model comparisons

Sinharay & Almond, 2007	Assessing Fit of Cognitive Diagnostic Models	Mixed numbersubtraction example (Tatsuoka et al., 1988)	BUGS	MCMC	Model Comparisons; Residual analysis
Sinharay, 2006	Model Diagnostics for Bayesian Networks	Mixed-number subtraction example (Tatsuoka, 1990), simulated data.	WinBUGS	MCMC	Posterior predictive model checking
Shute & Moore, 2017	Consistency and Validity in Game-Based Stealth Assessment	Physics Playground (Newton's Playground)	Netica	EM-algorithm	Correlation with external criteria
Shute, Wang, Greiff, Zhao, & Moore, 2016	Measuring Problem Solving Skills via Stealth Assessment in an Engaging Video Game	Use Your Brainz	Netica	Not presented	Correlation with external criteria
Song, Wang, Dai & Ding, 2018	Bayesian Network for Modeling Uncertainty in Attribute Hierarchy	Fraction Subtraction	Not presented	Not presented	Model Comparisons
Vomlel, 2004	Bayesian Networks in Educational Testing	Math skills	HUGIN	Hugin PC algorithm, EM-algorithm	Classification and prediction accuracy
West, Rutstein,Mislevy, Liu, Choi, Levy, Crawford, DiCerbo, Chappel, & Behrens, 2010	A Bayesian NetworkApproach to Modeling Learning Progressions and Task Performance	Packet Tracer	Not presented	Not presented	Model Comparisons; Inspection of conditional probability tables
Williamson, Almond, & Mislevy, 2000	Model Criticism of Bayesian Networks with LatentVariables	Simulation study	Ergo	EM-algorithm	Classification and prediction accuracy
Xing, Li, Chen,Huang, Chao, Massicotte, & Xie, 2020	Automatic Assessment of Students' Engineering Design Performance Using a Bayesian Network	Engineering design tasks	aGrUM	Not presented	Classification and prediction accuracy

Citation:
Uglanova, I. (2021). Model criticism of Bayesian networks in educational assessment: A systematic review. *Practical Assessment, Research & Evaluation*, 26(22). Available online: https://scholarworks.umass.edu/pare/vol26/iss1/22/

Corresponding Author

Irina Uglanova
Institute of Education / Centre for Psychometrics and Measurement in Education / Laboratory for New Construct Measurement and Test Design
HSE University
Moscow, Russia

Email: iuglanova [at] hse.ru