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

Irina Uglanova

HSE University

Follow this and additional works at: https://scholarworks.umass.edu/pare

Part of the Educational Assessment, Evaluation, and Research Commons

Recommended Citation
Available at: https://scholarworks.umass.edu/pare/vol26/iss1/22

This Article is brought to you for free and open access by ScholarWorks@UMass Amherst. It has been accepted for inclusion in Practical Assessment, Research, and Evaluation by an authorized editor of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.
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,
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:

\[
P(\theta | x) = \frac{P(x | \theta) * P(\theta)}{P(x)} \tag{1}
\]

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
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 prediction 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 κ (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)
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 PPCM method (Gelman, Meng, & Stern, 1996). PPCM 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 PPCM 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 PPCM 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 PPCM 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 PPCM 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 PPCM 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


https://doi.org/10.1111/1467-9868.00353


https://doi.org/10.1142/S021848850400259X


https://doi.org/10.1177/0735633120960422

https://doi.org/10.1002/j.2333-8504.2004.tb01929.x

### Appendix A

<table>
<thead>
<tr>
<th>Authors, Publication Year</th>
<th>Title of the Article</th>
<th>Name of the Instrument</th>
<th>Software</th>
<th>Method of Parameter Estimation</th>
<th>Model Checking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Almond, 2015</td>
<td>Tips and Tricks for Building Bayesian Networks for Scoring Game-Based Assessments</td>
<td>Physics Playground (Newton's Playground)</td>
<td>R-Netica</td>
<td>EM-algorithm</td>
<td>Classification and prediction accuracy; Mutual Information; Correlation with external criteria</td>
</tr>
<tr>
<td>Almond, Kim, Shute &amp; Ventura, 2013</td>
<td>Debugging the Evidence Chain</td>
<td>Physics Playground (Newton's Playground)</td>
<td>Not presented</td>
<td>Not presented</td>
<td>Mutual Information</td>
</tr>
<tr>
<td>Almond, Kim, Velasquez, &amp; Shute, 2014</td>
<td>How Task Features Impact Evidence From Assessments Embedded in Simulations and Games</td>
<td>Math WordProblems; Physics Playground (Newton's Playground)</td>
<td>Not presented</td>
<td>Not presented</td>
<td>Mutual Information</td>
</tr>
<tr>
<td>Almond, Mulder, Hemat, &amp; Yan, 2009</td>
<td>Bayesian Network Models for Local Dependence among Observable Outcome Variables.</td>
<td>NetPass</td>
<td>StatShop</td>
<td>MCMC</td>
<td>Mutual Information; Residual analysis</td>
</tr>
<tr>
<td>Almond, Yan &amp; Hemat, 2008</td>
<td>Parameter Recovery Studies with a Diagnostic Bayesian Network Model</td>
<td>Information and Communication Technology (ICT) Literacy Assessment Simulation study</td>
<td>StatShop</td>
<td>MCMC</td>
<td>Model Comparisons</td>
</tr>
<tr>
<td>Crawford, 2014</td>
<td>Posterior Predictive Model Checking in Bayesian Networks</td>
<td>Simulation study</td>
<td>WinBUGS</td>
<td>MCMC</td>
<td>Posterior predictive model checking</td>
</tr>
<tr>
<td>De Klerk, Eggen, &amp; Veldkamp, 2016</td>
<td>A Methodology for Applying Students’ Interactive Task Performance Scores From a Multimedia-Based Performance Assessment in a Bayesian Network</td>
<td>Confined spaceguard (CSG) students' assessment</td>
<td>GeNle</td>
<td>Clustering algorithm</td>
<td>Correlation with external criteria</td>
</tr>
<tr>
<td>Author(s)</td>
<td>Title</td>
<td>Method(s)</td>
<td>Tools/simulation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------------------</td>
<td>----------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DiCerbo, Xu, Levy, Lai, &amp;</td>
<td>Modeling Student Cognition In Digital and Nondigital Assessment</td>
<td>The Alice in Arealand game</td>
<td>WinBUGS, MCMC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Holland, 2017</td>
<td>Environments</td>
<td></td>
<td>Inspection of conditional probability tables; Correlation with external criteria</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kim, Almond, &amp; Shute,</td>
<td>Applying Evidence-Centered Design for the Development of Game-Based Assessments in Physics Playground</td>
<td>Physics Playground (Newton's Playground)</td>
<td>R-Netica, Gradient descent algorithm</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2016</td>
<td></td>
<td></td>
<td>Mutual Information; Inspection of conditional probability tables</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lee &amp; Corter, 2011</td>
<td>Diagnosis of Subtraction Bugs Using Bayesian Networks</td>
<td>Study of Subtraction Bugs (VanLehn, 1981)</td>
<td>HUGIN, Not presented</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Classification and prediction accuracy; Model comparisons</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Levy, 2006</td>
<td>Posterior Predictive Model Checking for Multidimensionality in Item Response Theory and Bayesian Networks</td>
<td>Simulation study</td>
<td>WinBUGS, MCMC</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Posterior predictive model checking</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Levy &amp; Mislevy, 2016</td>
<td>Bayesian Psychometric Modeling</td>
<td>Mixed number subtraction example</td>
<td>WinBUGS, MCMC</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Posterior predictive model checking</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pardos, Feng, Heffernan,</td>
<td>Analyzing Fine-Grained Skill Models Using Bayesian and Mixed Effects Methods</td>
<td>Massachusetts Comprehensive Assessment System</td>
<td>Not presented, Not presented</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heffernan, 2007</td>
<td></td>
<td></td>
<td>Classification and prediction accuracy; Residual analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pardos, Heffernan,</td>
<td>The Effect of Model Granularity on Student Performance Prediction Using Bayesian Networks</td>
<td>Massachusetts Comprehensive Assessment System</td>
<td>Bayes Net Toolkit for MATLAB, Not presented</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anderson, &amp; Heffernan,</td>
<td></td>
<td></td>
<td>Classification and prediction accuracy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rupp, Levy, Dicerbo,</td>
<td>Putting ECD into Practice: The Interplay of Theory and Data in Evidence Models within a Digital Learning Environment</td>
<td>Packet Tracer</td>
<td>WinBUGS, MCMC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sweet, Crawford, Calico,</td>
<td></td>
<td></td>
<td>Posterior predictive model checking</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Benson, Fay, Kunze,</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mislevy, &amp; Behrens, 2012</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rutstein, 2012</td>
<td>Measuring Learning Progressions Using Bayesian Modeling in Complex Assessments</td>
<td>Learning Progressions’ Assessment</td>
<td>WinBUGS, MCMC</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Classification and prediction accuracy; Model comparisons</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Author(s)</td>
<td>Title</td>
<td>Methodology</td>
<td>Software</td>
<td>Model Comparisons; Residual analysis</td>
<td>Posterior predictive model checking</td>
</tr>
<tr>
<td>-----------------------------------</td>
<td>------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------------</td>
<td>----------------</td>
<td>--------------------------------------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td>Sinharay &amp; Almond, 2007</td>
<td>Assessing Fit of Cognitive Diagnostic Models</td>
<td>Mixed numbersubtraction example (Tatsuoka et al., 1988)</td>
<td>BUGS</td>
<td>MCMC</td>
<td></td>
</tr>
<tr>
<td>Sinharay, 2006</td>
<td>Model Diagnostics for Bayesian Networks</td>
<td>Mixed-number subtraction example (Tatsuoka, 1990), simulated data.</td>
<td>WinBUGS</td>
<td>MCMC</td>
<td>Posterior predictive model checking</td>
</tr>
<tr>
<td>Shute &amp; Moore, 2017</td>
<td>Consistency and Validity in Game-Based Stealth Assessment</td>
<td>Physics Playground (Newton's Playground)</td>
<td>Netica</td>
<td>EM-algorithm</td>
<td></td>
</tr>
<tr>
<td>Shute, Wang, Greiff, Zhao, &amp; Moore, 2016</td>
<td>Measuring Problem Solving Skills via Stealth Assessment in an Engaging Video Game</td>
<td>Use Your Brainz</td>
<td>Netica</td>
<td>Not presented</td>
<td></td>
</tr>
<tr>
<td>Song, Wang, Dai &amp; Ding, 2018</td>
<td>Bayesian Network for Modeling Uncertainty in Attribute Hierarchy</td>
<td>Fraction Subtraction</td>
<td>Not presented</td>
<td>Not presented</td>
<td>Model Comparisons</td>
</tr>
<tr>
<td>Vomlel, 2004</td>
<td>Bayesian Networks in Educational Testing</td>
<td>Math skills</td>
<td>HUGIN</td>
<td>Hugin PC algorithm, EM-algorithm</td>
<td></td>
</tr>
<tr>
<td>West, Rutstein, Mislevy, Liu, Choi, Levy, Crawford, DiCerbo, Chappel, &amp; Behrens, 2010</td>
<td>A Bayesian Network Approach to Modeling Learning Progressions and Task Performance</td>
<td>Packet Tracer</td>
<td>Not presented</td>
<td>Not presented</td>
<td>Model Comparisons; Inspection of conditional probability tables</td>
</tr>
<tr>
<td>Williamson, Almond, &amp; Mislevy, 2000</td>
<td>Model Criticism of Bayesian Networks with Latent Variables</td>
<td>Simulation study</td>
<td>Ergo</td>
<td>EM-algorithm</td>
<td></td>
</tr>
<tr>
<td>Xing, Li, Chen, Huang, Chao, Massicotte, &amp; Xie, 2020</td>
<td>Automatic Assessment of Students' Engineering Design Performance Using a Bayesian Network</td>
<td>Engineering design tasks</td>
<td>aGrUM</td>
<td>Not presented</td>
<td></td>
</tr>
</tbody>
</table>
Citation:

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