DOCUMENT RESUME

ED 458 238		TM 033 399

AUTHOR	Williamson, David M.; Mislevy, Robert J.; Almond, Russell G.
TITLE	Model Criticism of Bayesian Networks with Latent Variables.
PUB DATE	2001-04-00
NOTE	26p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Seattle, WA, April 11-13, 2001).
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Artificial Intelligence; *Bayesian Statistics; *Cognitive Tests; *Mathematical Models; *Networks
IDENTIFIERS	*Latent Variables

ABSTRACT

This study investigated statistical methods for identifying errors in Bayesian networks (BN) with latent variables, as found in intelligent cognitive assessments. BN, commonly used in artificial intelligence systems, are promising mechanisms for scoring constructed-response examinations. The success of an intelligent assessment or tutoring system depends on the adequacy of the student model, representing the relationship between the unobservable cognitive variables of interest (thetas) and the observable features of task performance (x) with the probability model for x given theta expressed as a BN. The method for model fit analyses investigated in this study is appropriate for several uses in cognitive assessment. Data were generated for posited models to reflect the true BN model and several discrepancies from the true model. The study examined three indices: (1) Weaver's Surprise Index (Weaver, 1948); (2) Good's Logarithmic Score (Good, 1952); and (3) the Ranked Probability Score (Epstein, 1969). Simulation studies offer promise for the usefulness of the Ranked Probability Score and Weaver's Surprise Index as global measures and node measures to detect specific types of modeling errors in the latent structure of BNs. The introduction of this methodology and the emphasis on model criticism of BNs with latent variables provide a means of maximizing the accuracy and usefulness of BN models for a variety of applications. (Contains 4 tables, 9 figures, and 26 references.) (SLD)

Model Criticism of Bayesian Networks with Latent Variables

David M. Williamson

Robert J. Mislevy

Russell G. Almond

Educational Testing Service

Presented at the annual meeting of the National Council on Measurement in Education
Seattle, Washington
April, 2001

Model Criticism of Bayesian Networks with Latent Variables

Introduction

The past decade has brought new emphasis on cognitive approaches to measurement constituting a paradigmatic shift, even a revolution (Mislevy, 1996), in educational measurement research (Embretson, 1983, 1998; Frederiksen, Mislevy, & Bejar, 1993; Marshall, 1989; Nichols, Chipman, & Brennan, 1995). This new emphasis is altering the foundation upon which inferences are made about examinees (e.g. Frederiksen, Mislevy, & Bejar, 1993; Mislevy, 1996; Nichols, Chipman, & Brennan, 1995). The adoption of a cognitive approach to assessment, and the more complex target of inference, calls for a constructed-response format to provide the rich and complex evidence required to support complex cognitive inferences (e.g. Chipman, Nichols, & Brennan, 1995; Collins, 1990; Fiske, 1990). Yet, the scoring of such complex constructed-response data for cognitive assessment remains the greatest obstacle to successful implementation.[1]

Bayesian Networks (BN), commonly utilized in artificial intelligence systems, are a promising mechanism for scoring such constructed-response examinations. Two distinct uses for BNs in complex assessments can be envisaged: Summarizing key aspects of a given student performance, given features extracted from the raw work products, and synthesizing evidence from such evaluations across tasks. This presentation concerns the latter use. However, the use of BN in this way is currently hampered by an inability to fully critique the implemented network, particularly with regard to potential errors in modeling the inherently latent cognitive variables. (Similar challenges arise in factor analysis and item response theory).

This study investigates statistical methods for identifying errors in BN with latent variables, as found in intelligent cognitive assessments. The success of an intelligent assessment or tutoring system depends on the adequacy of the *student model*, representing the relationship between the unobservable cognitive variables of interest (θs) and the observable features of task performance (xs), with the probability model for x given θ being expressed as a BN.

The student model is constructed on the basis of a *cognitive task analysis* (CTA), an investigation of the cognitive components that contribute to task performance (Mislevy, Steinberg, Breyer, Almond, & Johnson, 1999). There is no assurance that the resulting student model is an accurate representation of the true structure of cognition, or

that it is the most useful model for the purpose of the assessment. *Model criticism* means evaluating the adequacy of a statistical model, enabling the analyst to discover hypotheses, variables, or relationships beyond those represented in the original model—to improve the structure of the BN in response to mismatches between modeled and observed data patterns (Mislevy, 1994; Mislevy & Gitomer, 1996). A BN model can be criticized at the levels of its fit as a whole (*global measures*) and of individual nodes (*node measures*).

At present, the current process of critiquing, refining, and validating a student model depends largely on examining the model from the perspective of the findings of the CTA and from theoretical considerations of cognition in the domain. The use of statistical diagnostic tools is notably lacking. Developing and using empirical tools for model criticism, therefore, is important to the continued development and implementation of BN methodologies in cognitive assessment. Statistical indices of model fit could be useful in cognitive assessment in several ways, such as (1) comparing proposed modeled structures to preliminary performance data; (2) evaluating the model-data concordance for nodes upon which examinee classification decisions are based, (3) identifying examinee performance that is inconsistent with the posited student model, and (4) confirming the appropriateness of the modeled cognitive structure and, by implication, providing evidence about the validity of that conceptualization of cognition in the domain. The methodology for model fit indices investigated in this study is appropriate for each of these uses, though the discussion in this study emphasizes the application to the evaluation of the structure of the statistical model of cognition.

Methodology

The general process for the simulations and comparisons conducted for each posited model in this study is illustrated in Figure 1, which is more fully detailed in the procedure section below. In brief, response data are generated according to a true BN model, which in applied settings would be real response data and as such would be produced under an unknown model structure. A "posited" BN model is created (in this study several such posited models were created to reflect both the true model and several particular discrepancies from the true model). Data are generated in accordance with the posited model, and a bootstrap distribution of model criticism indices calculated with these latter datasets become a null distribution for evaluating the fit statistics calculated with the data from the true model.

Indices

This study examined three indices, Weaver's Surprise Index (Weaver, 1948), Good's Logarithmic Score (Good, 1952), and the Ranked Probability Score (Epstein, 1969)[2], that have been used to evaluate the accuracy of probabilistic predictions in weather forecasting (Murphy & Winkler, 1984). Each measures of the degree of "surprise" felt when a datum is observed.[3]

Weaver's Surprise Index

Weaver (1948) developed the Surprise Index to distinguish a "rare" event from a "surprising" event. An event is surprising if its probability is small compared with the probabilities of other possible outcomes. A *surprising* event must be a *rare* event, but a *rare* event need not be *surprising*. His definition of surprise is

$$(S.I.)_i = \frac{E(p)}{p_i} = \frac{p_1^2 + p_2^2 + \ldots + p_n^2}{p_i}, \quad (1)$$

where there are n possible outcomes of a particular probabilistic event (in BN cognitive assessments with discrete variables, the n possible states of a variable), p_1-p_n are the prior probabilities of each of the n possible states, $E(p)$ is the expected value of the probability, and p_i is the prior probability of the observed state. Values increasingly greater than unity indicate increasingly surprising observations.

Good's Logarithmic Score

In a discussion of fees and rational decisions, Good (1952) introduced what we shall be refer to as Good's Logarithmic Score:

$$GL = \log(bp_i) \quad (2)$$

when the (predicted) event occurs, and

$$GL = \log b(1 - p_i) \quad (3)$$

when it does not. Here p_i is the prior probability of the event i in question before making the observation, and b is a penalty term that keeps a forecaster from long term gain by simply predicting the average frequency of occurrence. This penalty term is given by

$$b = -\sum_{j=1}^{r} x_j \log x_j \,, \tag{4}$$

where r is the number of possible outcomes and xj is the expectation of pj, that is, xj is the marginal probability associated with category j before the observation. Values of Good's Logarithmic Score near zero indicate accurate prediction, and values increasing from zero indicate poor prediction.

Ranked Probability Score

Epstein (1969) developed the Ranked Probability Score to evaluate forecasting accuracy when the states of the predicted variable are categories of an ordered variable (such as four categories of temperature in degrees Fahrenheit). Its distinguishing feature is that it considers how close (categorically) the predicted probabilistic outcome is to the observed outcome. The Ranked Probability Score is given by

$$S_j = \frac{3}{2} - \frac{1}{2(K-1)} \sum_{i=1}^{K-1} \left[\left(\sum_{n=1}^{i} p_n \right)^2 + \left(\sum_{n=i+1}^{K} p_n \right)^2 \right] - \frac{1}{K-1} \sum_{i=1}^{K} |i-j| p_i \tag{5}$$

where K represents the number of possible outcome states and j indicates the observed outcome. The Ranked Probability Score uses a linearly increasing penalty as the predicted observation becomes more distant from the observed state, implying that node categorizations are an interval scale as they progress from one extreme to the other. The values of the Ranked Probability Score vary from 0.00 to 1.00, indicating the poorest possible prediction and best possible prediction respectively. This study examined several indices, including Weaver's Surprise Index (Weaver, 1948), Good's Logarithmic Score (Good, 1952), the Ranked Probability Score (Epstein, 1969), the Quadratic Brier Score (Brier, 1950), Good's Logarithmic Surprise Index (Good, 1954), Logarithmic Score (Cowell, Dawid, & Spiegelhalter, 1993) and the Spearman correlation coefficient.

Data Generation Model

As a baseline for evaluating fit indices, we generated 1000 response patterns x from a hypothetical BN cognitive assessment—the 'Data Generation' BN—with known nodes, edges, and conditional probabilities. Although they are simulated, we refer to these vectors as 'observed' data since they represent the data that would be observed in practice. The structure of the Data Generation model, based on a hypothetical example of a student model for a general practice physician, is provided as Figure 2. The nodes θ_1 through θ_4 are latent variables representing aspects of

physician ability while the nodes X_1 through X_5 are summary observable variables from interaction with five simulated patients (with each patient represented as a different observable).

Model Criticism Computation

Our strategy was to route predictions for observable variables through the latent structure, providing an opportunity to detect problems with the latent structure even though the latent student model variables cannot be assessed directly. Errors in the student-model would manifest patterns of poor prediction for observable nodes individually or in the aggregate.

The 'observed data' were uploaded into the Data Generation BN. For each of the 1000 simulees, predictive probabilities were computed for each observable node treating the remaining observable nodes as known (i.e. for observable nodes X_1 through X_n the probability that node k is in state j is given by $P_{kj}^{*} = p(X_k = j \mid X_1,...X_{k-1}, X_{k+1},...X_n)$). The resulting probabilities for X_2 were treated as predictions to be compared to the observed state of X_2 for the simulee, as required to calculate the model criticism indices discussed above for each observed-variable node in turn for a given simulee. Carrying out this process for each of the observable nodes provided the node measures, and then aggregating across the five nodes produced a global measure for the simulee. The mean value of a node measure across the 1000 simulees served as the node measure (node-data fit) for the node in question, while the mean global measure value across the 1000 simulees served as the global measure of the model-data fit[4].

Error Models

Manipulating the Data Generation Model to introduce errors in the latent structure produced a number of alternative models to serve as the targets for investigating the utility of the model criticism indices. The errors introduced included node error, directed edge error, variable state error, and prior probability error.

The structure of the models representing the erroneous exclusion and inclusion of latent nodes are provided as Figures 3 and 4, respectively while the structure of models representing the erroneous exclusion and inclusion of weak and strong edges are provides as Figures 5, 6 and 7. In addition, three other error models were developed to represent the erroneous inclusion and exclusion of node states for a continuous latent variable and with incorrect prior probabilities assigned to a latent variable.

Procedure

Each stage of the study followed the same sequence of steps: 1) Generate a dataset (N=1000) consistent with the posited model (the model, either erroneous or true, that is the subject of model criticism). 2) Use the posited model to produce the probabilities via Bayesian Network updating software, in this case Ergo (Beinlich & Herskovits, 1990; Noetic Systems, 1996) for each observable node for both the model-consistent data (from step 1) and the 'observed' data. 3) Compute the fit indices (described above) for the observable variables at various sample sizes for both the model-consistent data and the 'observed' data and determine the distributional properties of the indices. 4) Bootstrap (Efron & Tibshirani, 1993) the model-consistent data (for posited model) to generate empirical distributions of values under the null hypothesis and determine critical values for evaluating the 'observed' data. 5) Evaluate the 'observed' data in light of these empirical distributions and critical values.

For each model (true and error models) this evaluation was conducted at sample sizes of 50, 100, 250, 500 and 1000 simulees. The larger sample sizes included the data from the smaller sample sizes. Each bootstrap data set had a sample size equal to that of the 'observed' data being evaluated, and critical values were established at the empirical values representing the 2.5% and 97.5% percentiles. This corresponds to a $p < .05$, two-tailed test. Values of the 'observed' data that exceeded these critical values were considered significant. A two-tailed test made it was possible to obtain significant results for better than expected model-data fit as well as misfit, though the latter is the primary concern of model criticism.

Results

Plots of the resultant values for the global and node measures served as the first basis for evaluating the effectiveness of the model criticism indices. For each plot (examples are provided below) the x-axis indicates sample size (e.g. 50 indicates that the observed data and each of the 1,000 bootstrapped data sets had N=50) and the y-axis indicates the empirical value of the index. The dots connected by dashed lines represent the mean values for the 'observed' data and the solid lines represent the upper and lower critical values from the bootstrap (97.5% and 2.5% of the 1,000 bootstrapped data sets, respectively).

Example Plots

To illustrate trends in the results, we provide examples for the Ranked Probability Score as applied to a node measure (the same procedure was also performed at the Global level, assessing the overall degree of model fit). The

node measure results for the Data Generation Model were predominantly within the bootstrap critical values, with an occasional value slightly beyond the cutoff, as illustrated in Figure 8.

In contrast, nodes for observables closely associated with an error in the latent structure showed more dramatic deviations from the bootstrap distributions, as illustrated in Figure 9.

Important findings include the discovery that the x with the closest proximity and greatest degree of relationship to the source of the latent structure error was nearly always the first (by sample size) to identify model error, and produced the greatest degree of discrepancy from the bootstrap parameters. Also, nodes in close proximity but with weaker associations with the source of the error seldom deviated from the bootstrap distributions.

Plot Summaries

A summary of the plots produced for the Ranked Probability Score is provided as Table 1. The 'Model' column indicates true or error model for which results are presented. The column marked 'Global' indicates the global measure results, and the columns marked X_1 through X_5 are the results for the observable variables. Numeric values in a cell indicate that at least one analysis (of the five sample sizes utilized) produced a significant deviation from the bootstrap distributions. The numeric values indicate which sample sizes produced significant deviations. Bold type represents cells where there was an error in the latent structure of the immediate parent variable, and a bold X appears in cells where there was an undetected error in the latent structure of the immediate parent variable. Cross-referencing the data in Table 1 to Figures 8 and 9 helps to clarify its interpretation. Tables 2 and 3 give similar summaries for Weaver's Surprise Index and Good's Logarithmic Score.

Discussion

Implications

These results offer promise of utility for the Ranked Probability Score and Weaver's Surprise Index as global measures and node measures to detect specific types of modeling errors in the latent structure of BNs. For global measures, major error types (node exclusions and strong edge errors) in the latent structure were detectable. For node measures (preferably used in combination) these indices helped identify major latent structure errors (node errors and strong edge errors) at moderate sample sizes, and minor latent structure errors (weak edge errors, node state errors, and prior probability errors) at large sample sizes. The results suggest utility as node measures even in the absence of model-

data misfit for global measures. Furthermore, these results suggest that as node measures these indices can identify nodes in close proximity to the latent structure error, providing the modeler some direction for appropriate modification to the student model.

To the extent that these results generalize to other such BN models with latent variables, Table 4 suggests guidelines for the use of the Ranked Probability Score (RPS), Weaver's Surprise Index (WSI), and Good's Logarithmic Score (GLS) as node measures.

Future Directions

Obviously an important direction for further research is to establish the generalizability of these results to BNs with latent variables by systematically manipulating BN features such as network size, associations, proportion of latent to observable nodes, etc. to determine whether model criticism is affected by such variations.

Conclusion

The introduction of this methodology, and more critically, the emphasis on model criticism of BNs with latent variables in general, provides a means of maximizing the accuracy and utility of BN models for a variety of applications. As methods of providing empirical support or criticism of student models in cognitive assessment, these results provide a means of ensuring that the student models developed are appropriate representations of the constellation of knowledges, processes, and strategies which contribute to task performance. This capability offers the potential of helping the analyst to create a student model from a CTA by comparing modeled structures with preliminary performance data; to revise BN structures to improve classification decisions for examinees; to provide validity evidence for the student model in the substantive domain; and to identify examinees who do not fit the model. With such applications these indices would contribute to the production of more accurate cognitive models in less time, facilitate the implementation of BN and related methodologies in future applications, and support the construct validity of the resultant cognitive assessments and intelligent tutoring systems.

References

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

Almond, R. G., Herskovits, E., Mislevy R. J., & Steinberg, L. S. (1999). Transfer of information between system and evidence models. In D. Heckerman & J. Whittaker (Eds.), Artificial intelligence and statistics 99: Proceedings of the seventh international workshop on artificial intelligence and statistics (pp. 181-186). San Francisco, CA: Morgan Kaufmann.

Beinlich, I. A., & Herskovits, E. H. (1990). Ergo: A graphical environment for constructing Bayesian belief networks. Proceedings of the conference on uncertainty in artificial intelligence. Cambridge, MA.

Bejar, I. I. (1991). A Methodology for scoring open-ended architectural design problems. Journal of Applied Psychology, 76, (4), 522-532.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. Monthly Weather Review, 78, 1-3.

Cowell, R. G., Dawid, A. P., & Spiegelhalter, D. J. (1993). Sequential model criticism in probabilistic expert systems. IEEE Transactions on Pattern Analysis and Machine Intelligence, 15, 209-219.

de Finetti, B. (1965). Methods for discriminating levels of partial knowledge concerning a test item. British Journal of Mathematical and Statistical Psychology, 18, 87-123.

Efron, B. & Tibshirani, R. (1993). An introduction to the bootstrap. New York: Chapman & Hall.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. Psychological Bulletin, 93, 179-197.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. Psychological Methods, 3, 380-396.

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

Frederiksen, N., Mislevy, R. J., & Bejar, I. I. (Eds.) (1993). Test theory for a new generation of tests. Hilsdale, New Jersey : Lawrence Erlbaum Associates.

Good, I. J. (1952). Rational decisions. Journal of the Royal Statistical Society, B, 14, 104-114.

Good, I. J. (1954). The appropriate mathematical tools for describing and measuring uncertainty. In C. F. Carter, G. P. Meredith, & G. L. S. Sheckle (Eds.), Uncertainty and Business Decisions (pp.385-388). Liverpool: University Press.

Marshall, S. P. (1989). Generating good items for diagnostic tests. In N. Frederiksen, R. Glaser, A. Lesgold, & M. G. Shafto (Eds.), Diagnostic monitoring of skill and knowledge acquisition (pp. 433-452). Hillsdale, NJ: Erlbaum.

Mislevy, RJ (1994). Evidence and inference in educational assessment. Psychometrika, 59 (4), 439-483.

Mislevy, R. J. (1996). Test theory reconceived. Journal of Educational Measurement, 33, 379-416.

Mislevy, R. J., & Gitomer, D.H. (1996). The role of probability-based inference in an intelligent tutoring system. User Mediated and User-Adapted Interaction, 5, 253-282.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (1999). On the several roles of task model variables in assessment design. In S. Irvine & P. Kyllonen (Eds.), Generating items for cognitive tests: Theory and practice. Hillsdale, NJ: Erlbaum.

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G., & Johnson, L. (1999). A cognitive task analysis, with implications for designing a simulation-based assessment system. Computers and Human Behavior, 15, 335-374.

Nichols, P. D., Chipman, S. F., & Brennan, R. L. (Eds.) (1995). Cognitively diagnostic assessment. Hilsdale, NJ: Lawrence Erlbaum Associates.

Noetic Systems (1996). ERGO [computer program]. Baltimore, MD: Noetic Systems, Inc.

Spiegelhalter, D.J., Dawin, A.P., Lauitzen, S.L., & Cowell, R.G. (1993). Bayesian analysis in expert systems. Statistical Science, 8 (3), 219-283.

Steinberg, L. S., & Gitomer, D. G. (1996). Intelligent tutoring and assessment built on an understanding of a technical problem-solving task. Instructional Science, 24, 223-258.

Weaver, W. (1948). Probability, rarity, interest, and surprise. Scientific Monthly, 67, 390-392.

Williamson, D. M., Bejar, I. I., & Hone, A. S. (1999). 'Mental model' comparison of automated and human scoring. Journal of Educational Measurement, 36 (2), 158-184.

Footnotes

[1]The interested reader is referred to the following sources for discussions of other aspects of the research program from which this work arises: cognitive psychology (Frederiksen, Mislevy, & Bejar, 1993; Steinberg & Gitomer, 1996); computer-based simulations and constructed-response tasks (Bejar, 1991; Williamson, Bejar, & Hone, 1999); probability-based reasoning (Almond, & Mislevy, 1999; Almond Herskovits, Mislevy & Steinberg, 1999); and assessment design (Mislevy, Steinberg, Breyer, Almond, & Johnson, 1999; Mislevy, Steinberg, & Almond, 1999).

[2] The Quadratic Brier Score (Brier, 1950), Good's Logarithmic Surprise Index (Good, 1954), Logarithmic Score (Cowell, Dawid, & Spiegelhalter, 1993) and Spearman correlation coefficient were also investigated but are not discussed due to lesser promise of utility.

[3]An interesting connection exists between indices of surprise (which are essentially measures of distance between probabilistic predictions and a criterion) and assessment: De Finetti (1965) proposed that students answer multiple-choice questions by assigning a probability to each option representing the student's belief that the option is the correct answer, and he provided methods of scoring such responses that increase scores as the assigned probabilities are less surprising in light of the key.

[4] By transposing the matrix of values it would be possible to utilize this procedure to evaluate the person-model fit rather than the model-data fit.

Table 1

Plot Summary for the Ranked Probability Score

Model	Global	X_1	X_2	X_3	X_4	X_5
	Level/Node					
Data Generation						
Node Exclusion	100, 250, 500, 1000	500, 1000		X	X	**100, 250, 500, 1000**
Node Inclusion					X	**100, 250, 500, 1000**
State Exclusion				X	**1000**	**500, 1000**
State Inclusion				X	X	X
Prior Probability				X	X	**500, 1000**
Strong Edge Exclusion	100, 250, 500, 1000	500, 1000				**100, 250, 500, 1000**
Strong Edge Inclusion	500, 1000		**250, 500, 1000**			500, 1000
Weak Edge Exclusion				X		100, 250, 500, 1000
Weak Edge Inclusion			X			

Table 2

Plot Summary for Weaver's Surprise Index

	Level/Node					
Model	Global	X_1	X_2	X_3	X_4	X_5
Data Generation						
Node Exclusion	100, 250, 500, 1000			**X**	**X**	**100, 250, 500, 1000**
Node Inclusion					**X**	**500, 1000**
State Exclusion				**X**	**1000**	**X**
State Inclusion				**X**	**X**	**X**
Prior Probability				**X**	**X**	**X**
Strong Edge Exclusion	250, 500, 1000	1000				**100, 250, 500, 1000**
Strong Edge Inclusion	500, 1000		**500, 1000**			100, 250, 500, 1000
Weak Edge Exclusion				**500, 1000**		
Weak Edge Inclusion			**X**			

Table 3

Plot Summary for Good's Logarithmic Score

	Level/Node					
Model	Global	X_1	X_2	X_3	X_4	X_5
Data Generation						
Node Exclusion				**X**	**X**	**X**
Node Inclusion					**X**	**X**
State Exclusion				**X**	**100, 250, 500, 1000**	**X**
State Inclusion				**X**	**100, 250, 500, 1000**	**X**
Prior Probability				**X**	**X**	**X**
Strong Edge Exclusion						**X**
Strong Edge Inclusion		100	**X**			
Weak Edge Exclusion				**X**		1000
Weak Edge Inclusion			**X**			

Table 4

Utilization as Node Measures

N	Sig. Deviation GLS	RPS	WSI	Error Types
≤250	yes	no	no	node state exclusion; node state inclusion
	no	yes	no	node inclusion; strong edge inclusion
	no	yes	yes	node exclusion; strong edge exclusion
>250 and ≤1000	yes	no	no	node state exclusion; node state inclusion
	no	yes	no	node state exclusion; prior probability error
	no	no	yes	weak edge exclusion
	no	yes	yes	node exclusion; node inclusion; strong edge exclusion; strong edge inclusion

Figure 1

Data Simulation and Reference Distribution Generation

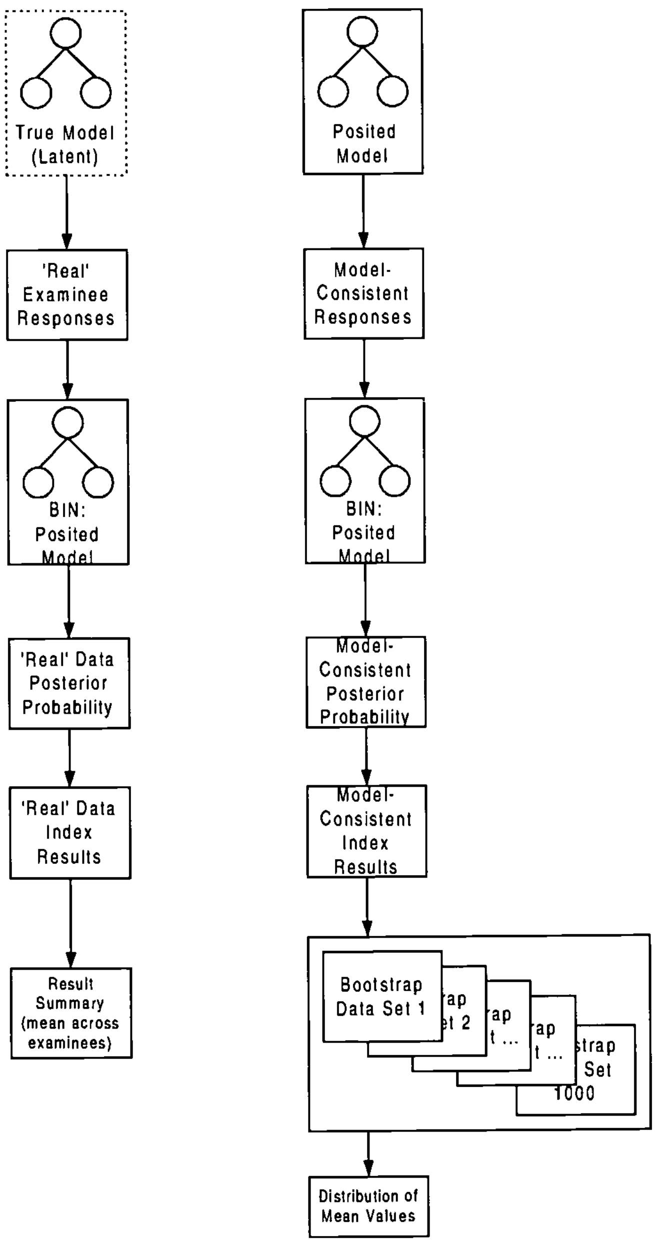


Figure 2

Data Generation Model

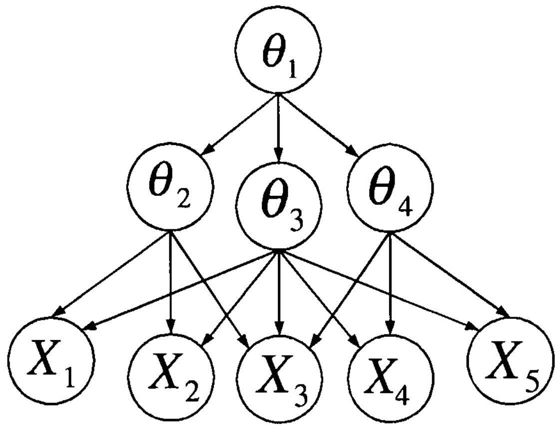


Figure 3

Node Exclusion Error Model

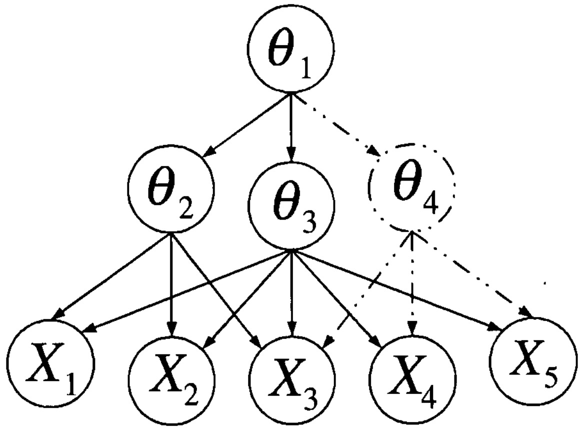


Figure 4

Node Inclusion Error Model

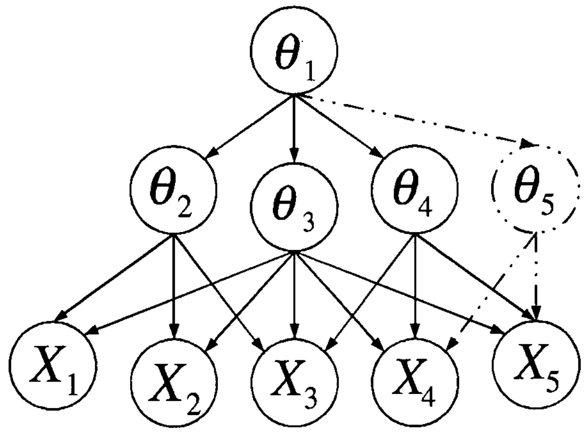


Figure 5

Weak Edge Exclusion Error Model

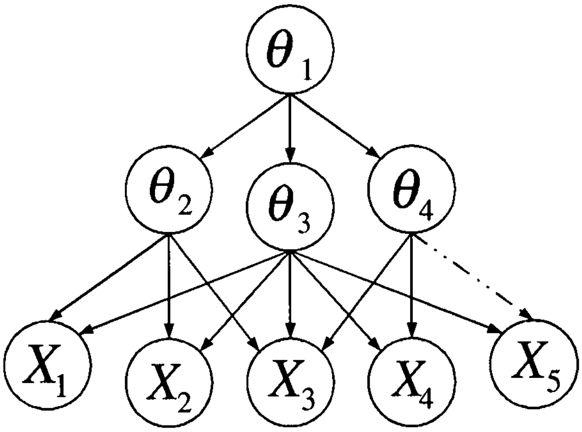


Figure 6

Strong Edge Exclusion Error Model

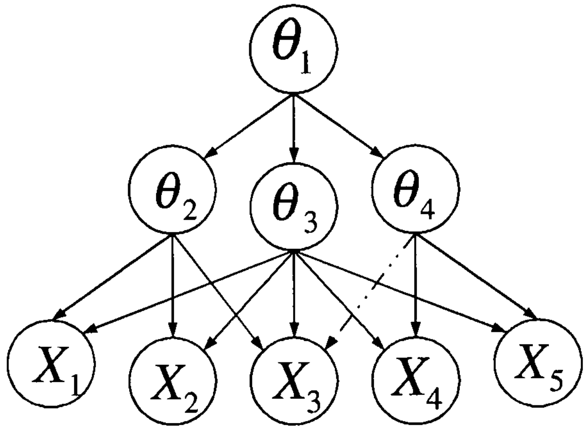


Figure 7

Edge (Strong or Weak) Inclusion Error Model

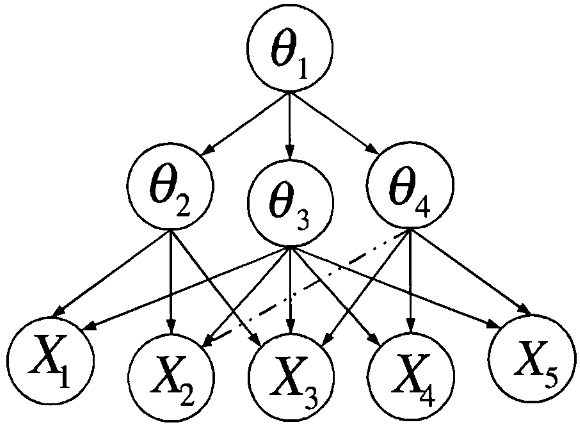


Figure 8

Patient 2 (X_2) Node Measure Ranked Probability Score Results for the Data Generation Model

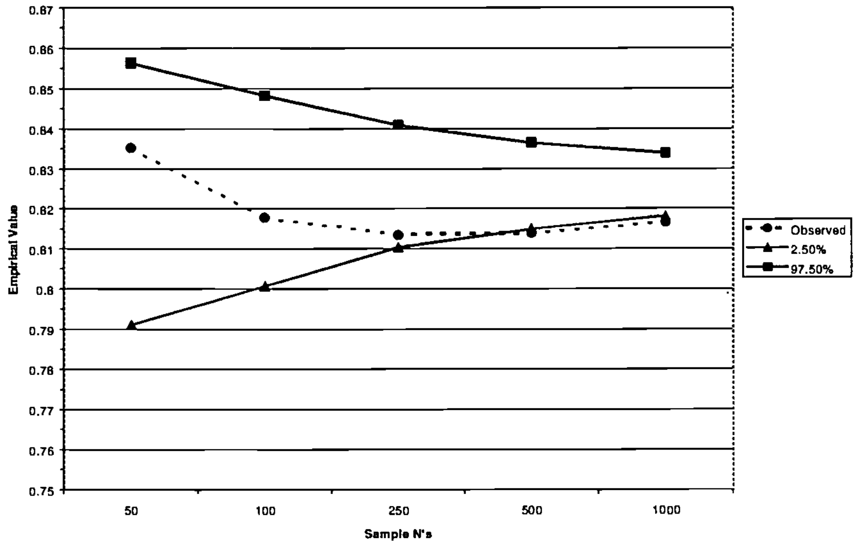
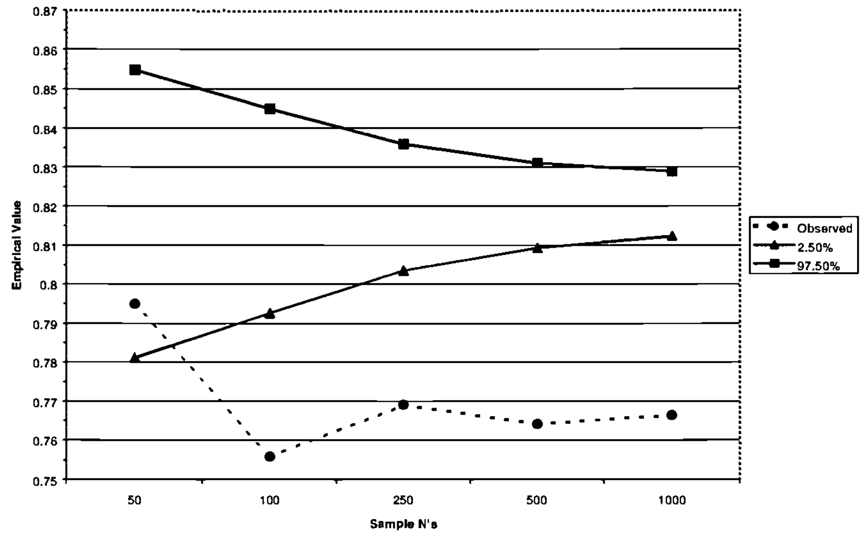


Figure 9

Patient 5 (X_5) Node Measure Ranked Probability Score Results for the Node Exclusion Model



U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

TM033399

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Model Criticism of Bayesian Networks with Latent Variables

Author(s): David M. Williamson ; Robert J. Mislevy ; Russell G. Almond

Corporate Source:	Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 [X]	Level 2A [ ]	Level 2B [ ]
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) and paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: David Williamson	Printed Name/Position/Title: David Williamson Research Scientist	
Organization/Address: ETS Rosedale Road Princeton, NJ 08541	Telephone: (609) 734-1303	FAX: (609) 734-1090
	E-Mail Address: dwilliamson@ets.org	Date: 10/30/01

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20742-5701
ATTN: ACQUISITIONS

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com

EFF-088 (Rev. 2/2000)