Analyzing Knowledge Structure: An Application of Graphical Models to a Medical Licensure Exam

Michael Culbertson, *University of Illinois, Urbana-Champaign*
Feiming Li, *National Board of Osteopathic Medical Examiners*

Presented at the Annual Meeting of the
National Council on Measurement in Education,
April 16, 2012, Vancouver, BC

Introduction

Supporting student success well in a global economy under the very real-world constraints of scarce resources requires certain economies of instructional practice. Teachers with ever-more-limited time and ever-increasing demands need to make critical decisions concerning educational interventions for individual students quickly and accurately. Supporting student success thus entails supporting teachers by providing timely, high-quality, detailed information about students' current progress. Most of today's educational assessment tools, however, usually provide only high-level information about broad educational domains (such as mathematics, reading, science). Measurement models for rapid educational response need to provide analysis of student achievement on a variety of sub-domains, and must eventually do so with a modest number of items.

Several classes of models are common choices for providing sub-domain ability estimation, including Multidimensional Item Response Theory (MIRT; Reckase, 2009) and diagnostic classification models (DCM; e.g. Rupp, Templin, and Henson, 2010). However, most of these models leave the structure of the underlying latent abilities unspecified, and it may be possible to achieve greater measurement precision by explicitly modeling the sub-domain relationships. Graphical models serve as one means for modeling these relationships.

Purpose

This paper applies graphical modeling techniques to analyze the sub-domain structure of a medical licensing exam. While not unknown in educational measurement (e.g. Mislevy et al., 2002; Levy and Mislevy, 2004), graphical models have not yet been widely adopted. This paper thus represents one of few applications of graphical modeling to analyze the latent knowledge structure of an operational exam. The paper illustrates the graph-development process in an exploratory analysis of operational data and investigates the relative precision of measurement models with unstructured (MIRT) and structured knowledge components through a small simulation study.

Theoretical Framework

Graphical Modeling

Graphical modeling techniques, such as Bayesian Networks (BN) and Structural Equation Modeling (SEM), consist of assigning the data a graph (a set of vertices and edges connecting them) and a corresponding joint probability distribution. In BN, the joint distribution is specified as a set of conditional distributions for each variable, given its parents, and Bayesian inference is typically used. In SEM, the joint distribution is usually taken as multivariate normal with a graph-implied covariance matrix and frequentist inference. In either case, graphical models are convenient for specifying complex relationships between variables due to the conditional independence relationships they imply and their intuitive visual representation.

In educational measurement, correlations of sub-domains may not be completely arbitrary, following instead some underlying structure. Ability in some sub-domains (say, reciprocation and fraction multiplication) may be taken as pre-requisite for achievement in others (fraction division). A graphical measurement model can be constructed to encode these relationships (Fig. 1). The graph specifies the joint distribution of latent abilities, and any usual IRT model may be given for the conditional distribution of the items.

A fundamental graphical modeling task, then, is determining the relationships among the variables. This may involve expert opinion to generate an initial model, modification indices and Wald tests to refine paths of the model, and information criteria or likelihood-ratio tests to compare candidate models.

Measurement Model

In this work, observed variables follow a Rasch model:

$$P(X_{ij} = 1 \mid \theta_i) = \frac{\exp(L_j \theta_i - b_j)}{1 + \exp(L_j \theta_i - b_j)}$$

where X_{ij} is the response from examinee i to item j, θ_i is a vector of latent abilities, L_j is a row-vector of 1s and 0s indicating the topics assigned to item j, and b_j is the item difficulty. A unidimensional Rasch model is used in the operational scoring of this exam, as is common among current licensure exams. Thus, the Rasch model was selected for this analysis to provide results more directly comparable with operational practice.

Knowledge Structure

Here, latent variables (sub-domain abilities) are specified according to a standard path model with uncorrelated normal error terms:

$$\theta_i = A\theta_i + \xi_i = (I_p - A)^{-1} \xi_i$$

where A is a matrix of path coefficients encoding the graph structure, $\xi_i \sim N_p(0, \Sigma)$, Σ is a diagonal matrix of error variances, p is the number of latent variables, and I_p is the identity matrix.

METHODS

Data

Graphical model development techniques were applied to item response data from a computer-based, non-adaptive medical licensure exam in a particular subject area, Obstetrics and Gynecology. Responses to 252 dichotomous items from 13 overlapping forms were analyzed for 776 examinees. Each examinee responded to 52–67 items, and each item had 55–465 responses (over 75% of items had at least 100 responses). Items for this particular exam are categorized in two ways during operational assessment development: by subject matter (general gynecology, gynecological oncology, normal obstetrics, abnormal obstetrics, and reproductive endocrinology) and by task context (patient history/examination, diagnosis/management, preventative care, and disorders of the breast). Items are assigned topics in each categorization by expert opinion, and the cross-categorization forms the basis for content balancing during form construction. Since these topics are fundamental to exam construction, they were taken as the 9 sub-domains measured in the exam. Examinees had 3–34 responses for each topic.

Graph Development

The modeling exercise began with three initial models (Fig. 3): an independence model (*A*=0), *a priori* subject paths (Normal Obstetrics → Abnormal Obstetrics, General Gynecology → Gynecological Oncology), and *a priori* context paths (History/Examination → Diagnosis/Management → Preventative Care). Then, model improvements were investigated by freeing path coefficients as identified by modification indices (Sörbom, 1989; Buse, 1982), which estimate the change in the log-likelihood if a particular parameter constraint were freed. For constrained parameter *a*, the modification index is:

$$MI_a = \frac{g_a^2}{2(k_a - d_a' H^{-1} d_a)}$$

where $g_a = \partial F / \partial a$, $k_a = \partial^2 F / \partial a \partial a$, $d_a = \partial^2 F / \partial a \partial \eta'$, F is the log-likelihood, η is a vector of free parameters, and $H = \partial^2 F / \partial \eta \partial \eta'$ is the model Hessian. For a given test model, paths with the largest modification indices were considered candidates, and one of the candidates was freed in the subsequent model. Models were then compared using information criteria (AIC, BIC).

Model Estimation

One impediment to the wide-spread adoption of measurement models with complex knowledge structures has been the difficulty in estimating high-dimensional latent-variable models (the "curse of dimensionality"). These models are typically estimated using the Expectation-Maximization Algorithm (EM; Dempster, Laird, and Rubin, 1977) or Markov Chain Monte Carlo (MCMC; e.g. Gilks, Richardson, and Spiegelhalter, 1996). While EM is very efficient for estimating model parameters, its Expectation step requires integration over the latent variables, which is very expensive for high-dimensional problems. MCMC is relatively efficient at sampling from high-dimensional joint probability distributions, but is less efficient for estimating model parameters.

The recently proposed Metropolis-Hasting Robbins-Monro Algorithm (MH-RM; Cai, 2010a; Cai, 2010b) achieves dramatically reduced computation times relative to EM and MCMC by combining their relative strengths: A Markov chain is constructed to sample efficiently from the posterior latent ability distribution, and the Robbins-Monro algorithm provides gradient-based maximization of the likelihood using draws from the MC. MH-RM is easy to code and fast relative to EM and pure MCMC in high-dimensional latent spaces. Although the MH-RM algorithm was not originally developed for latent path models, the extension is transparent, as demonstrated here.

Simulation Study

The simulation study compares the precision of ability estimates under measurement models with unstructured and structured relationships between latent abilities. Simulations investigate scenarios with a great deal of information for each attribute and scarce information. Two tests of 10 sub-domains were constructed: one with 150 items (15 per sub-domain) and one with 30 items (3 per sub-domain). Two knowledge structures were simulated: a relatively complex structure (Fig. 1) and a relatively simple structure (Fig. 2). Path coefficients were assigned reasonable values to represent a range of sub-domain relationships. Item difficulties were assigned randomly from a standard normal distribution, and sub-domain abilities (ξ) for 1,000 examinees were drawn from a standard multivariate normal distribution. The test lengths and knowledge structures were crossed to yield four simulation conditions. Dichotomous item responses were simulated using the Rasch measurement model with a latent path structure, as described above. Then, the item responses were fit with the path model (using the correct graph), a multidimensional Rasch model with unstructured covariance matrix, and an independence model. Mean squared error in the estimated abilities was calculated under each model for comparison. In order to detect over-fitting, a new set of abilities and item responses were generated for each condition, abilities were estimated from the new item responses using the previously calibrated model parameters, and mean squared error was calculated for the new estimates.

RESULTS

By AIC, all models performed better than the independence model, even if they contained only one path relationship (Table 1). The *a priori* subject paths model performed worse than the independence model by BIC; otherwise, AIC and BIC put the models in the same order. Adding the path Diagnosis → Abnormal Obstetrics to the independence model, as suggested by modification indices, improved model fit, but adding a second path (General Gynecology → Disorders) provided no additional improvement. The *a priori* task context model provided the best fit, which was not improved by adding an additional path (General Gynecology → History). This suggests that an unstructured model (all possible correlations) would be overly unconstrained and that the modest path model provides a much more parsimonious account of the relationships between sub-domain knowledge.

In the simulation study, independence models provided 6–7% larger mean squared error (MSE) than the correct model in a cross-validation sample for simple structure and 13–14% larger MSE for complex structure, as would be expected since the latent variables were not independent. Interestingly, the MSE for the fully-correlated (unstructured) models was only marginally larger than for the correct model, providing little evidence of over-fitting. However, the path model provided a much more parsimonious fit than the unstructured model.

Implications for Practice

Supporting students requires more-detailed information about their educational progress than provided by low-dimensional assessments. Graphical models are a simple yet powerful means of specifying the structural relationships of test sub-domains. This work illustrates (1) the application of a new estimation algorithm to latent path modeling, (2) the development of a graphical model for the knowledge structure of an operational test, and (3) the relative measurement precision of structured and unstructured models for examinee knowledge. It serves as an example for others seeking to satisfy the call for detailed and imminently useful assessment results.

References

Buse, A. (1982). The Likelihood Ratio, Wald, and Lagrange Multiplier Tests: An expository note. *American Statistician.* 36, 153–157.

Cai, L. (2010a). High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. *Psychometrika,* 75, 33–57.

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

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B,* 39, 1–38.

Gilks, W.R., Richardson, S., and Spiegelhalter, D.J. (1996). *Markov Chain Monte Carlo in Practice.* London: Chapman and Hall.

Levy, R. and Mislevy, R. (2004). Specifying and refining a measurement model for a computer-based interactive assessment. *International Journal of Testing.* 4, 333–369.

Mislevy, R., Almond, R., DiBello, L., Jenking, F., Steinberg, L., Yan, D., and Senturk, D. (2002). *Modeling Conditional Probabilities in Complex Educational Assessments.* Technical Report CSE-580. Center for the Study of Evaluation, University of California, Los Angeles.

Rackase, M. D. (2009). *Multidimensional Item Response Theory.* New York: Springer.

Rupp, A., Templin, J., and Henson, R. (2010). *Diagnostic Measurement: Theory, Methods, and Applications.* New York: Guliford Press.

Sörbom, D. (1989). Model Modification. *Psychometrika,* 54, 371–384.

FIGURES

Figure 1: A Hypothetical Graph for Fraction Arithmetic

Latent variables for sub-domains are represented by circles. Sample observed variables (items) are represented by rectangles. Edges represent pre-requisite relationships.

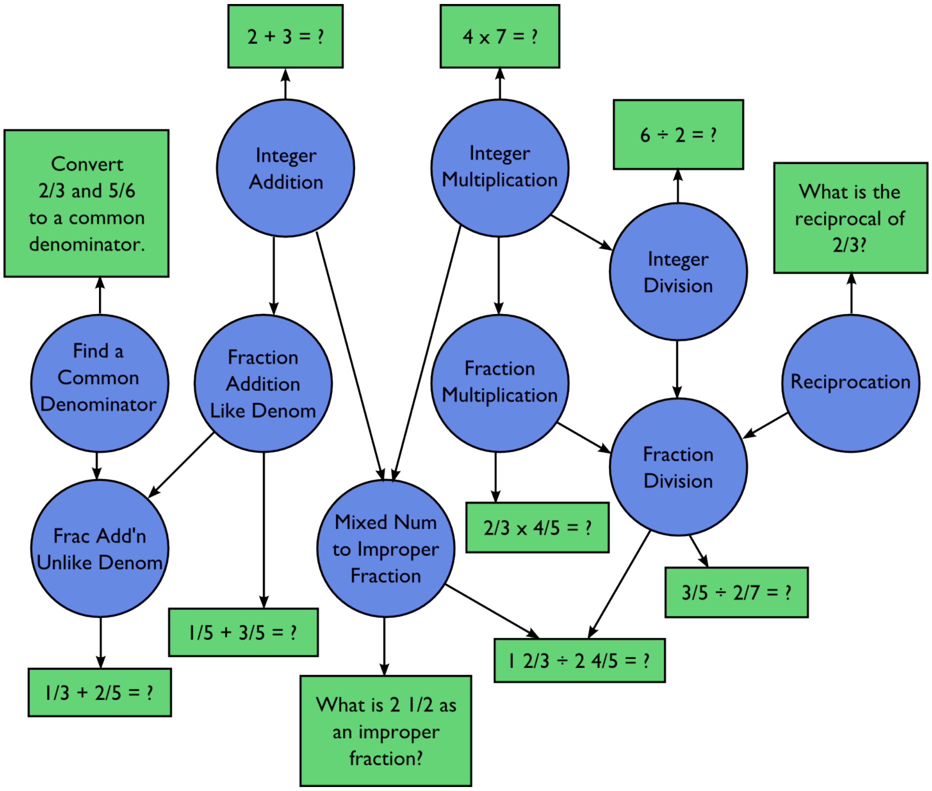

Figure 2: Simple Structure for Simulation Study

Eight latent variables are divided into two independent groups, each conditional on a higher-order ability. Squares illustrate sample observed variables.

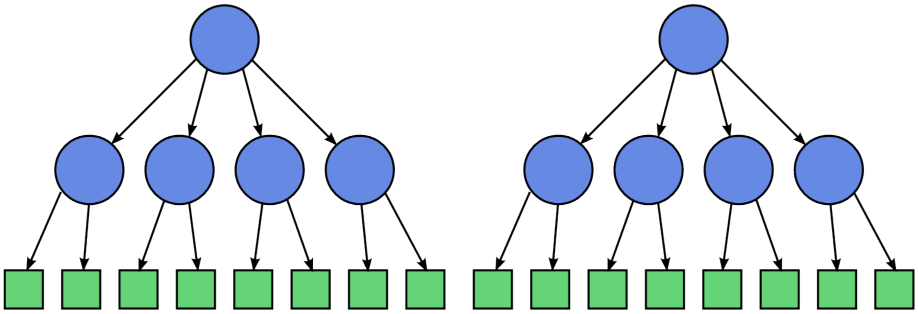

Figure 3: Sample Path Diagram for Medical Licensure Exam

Latent variables (circles) come from two categorizations of each item (squares): content area and medical task context. Paths from two of the three initial models are shown.

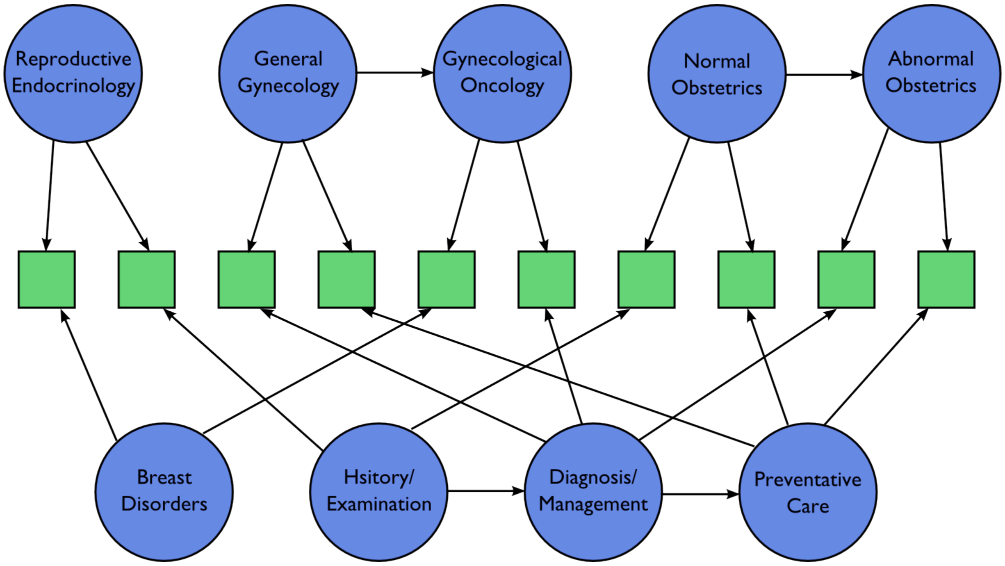

Table 1: Model Comparison

Model	AIC	BIC
Independence	60653	61868
Gen Gyn → Gyn Onc, Normal Ob → Abnormal Ob	60651	61875
Diagnosis → Abnormal Ob, Gen Gyn → Disorders	60611	61835
Diagnosis → Abnormal Obstetrics	60560	61779
Both sets of theoretical paths	60542	61776
Gen Gyn → History → Diagnosis → Prevention	60473	61702
History → Diagnosis → Prevention	60469	61693

Table 2: Simulation Study Results

The table lists the increase in cross-validation mean squared error of ability estimates, relative to error in the structured (correct path) model.

Configuration	Unstructured	Independence
Simple Structure (3 items/ability)	1.4%	6.8%
Simple Structure (15 items/ability)	0.1%	5.9%
Complex Structure (3 items/ability)	1.9%	12.8%
Complex Structure (15 items/ability)	0.4%	14.0%