ED 476 861 TM 034 955

AUTHOR	Levy, Roy; Mislevy, Robert J.
TITLE	Specifying and Refining a Complex Measurement Model.
PUB DATE	2003-04-00
NOTE	83p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, April 22-24, 2003).
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC04 Plus Postage.
DESCRIPTORS	*Bayesian Statistics; Cognitive Processes; Markov Processes; *Mathematical Models; Monte Carlo Methods; Simulation; Statistical Analysis

ABSTRACT

This paper aims to describe a Bayesian approach to modeling and estimating cognitive models both in terms of statistical machinery and actual instrument development. Such a method taps the knowledge of experts to provide initial estimates for the probabilistic relationships among the variables in a multivariate latent variable model and refines these estimates using Markov Chain Monte Carlo procedures. This process is described in terms of NetPASS, a complex simulation based assessment in the domain of computer networking. The paper describes a parameterization of the relationships in NetPASS via an ordered polytomous item response model and details the updating of the model with observed data via Bayesian statistical procedures ultimately being provided by Markov Chain Monte Carlo estimation. (Contains 12 tables, 9 figures, and 47 references.) (Author/SLD)

Running head: Specifying a Complex Measurement Model

Specifying and Refining a Complex Measurement Model

Roy Levy and Robert J. Mislevy

University of Maryland, College Park

[Paper presented at the annual meeting of the National Council On Measurement In Education, Chicago, IL, April 2003]

Correspondence concerning the paper should be directed to the first author at:
1230 Benjamin Building
University of Maryland
College Park, MD 20742-1115
levyr@wam.umd.edu

Abstract

In this paper we aim to describe a Bayesian approach to modeling and estimating cognitive models both in terms of statistical machinery and actual instrument development. Such a method taps the knowledge of experts to provide initial estimates for the probabilistic relationships among the variables in a multivariate latent variable model and refines these estimates using Markov chain Monte Carlo procedures. This process is described in terms of NetPASS, a complex simulation based assessment in the domain of computer networking. We describe a parameterization of the relationships in NetPASS via an ordered polytomous item response model, and detail the updating of the model with observed data via Bayesian statistical procedures ultimately being provided by Markov chain Monte Carlo estimation.

Key words: Bayesian inference networks, Markov chain Monte Carlo, measurement model, complex assessment

Specifying and Refining a Complex Measurement Model

Instruments in educational measurement have taken on a variety of forms ranging from the more familiar e.g., multiple choice formats, to the unique, e.g., computer simulation of a real-world application. Different formats yield different work products, e.g., a scan-tron sheet with circles filled in, essays to be scored by raters, and portfolios. While methods for drawing inferences from examinees' work products to their knowledge, skills, and abilities exist for the more popular assessment instruments, new and innovative assessment instruments are often left to develop inferential rules individually. Nonstandard and complex tasks result in complex work products. Drawing proper inferences from these work products requires models that accumulate and incorporate information in order to produce a score that is interpretable and valid for inferences about students. It is these models that we investigate in this paper. More specifically, we focus on a method of specifying and refining models that allow for updating beliefs and reaching conclusions about examinees based on observable variables that are extracted from multiple, complex work products.

Drawing from Schum (1987) we maintain that probability based reasoning can be applied to all forms of inference, more specifically, to inference in educational measurement, and is particularly useful for inferences from innovative and complex assessment instruments (Mislevy, 1994). In what follows we shall couch such probability based reasoning in a more general assessment framework, describe such reasoning in detail, and illustrate these methods in practice via an example from a complex assessment of the cognitive development of students in the Cisco Networking Academy Program.

Specifically, the development of NetPASS, a measurement device to be utilized to assess cognitive development of students in the third semester of Cisco Networking Academy

Program's sequence of courses on computer networking, will be discussed. However, it should be stressed that while the particulars of NetPASS will be described in detail, the process of instrument and model development can be reinstantiated in settings that, on the surface, may appear to have little in common with NetPASS.

Assessment Context

In what follows we summarize the application of the Evidence Centered Design framework (Mislevy, Steinberg, & Almond, in press) to an assessment of computer networking proficiency designed for Cisco Learning Institute's (CLI) Cisco Networking Academy Program (CNAP).

CLI collaborates with high schools, community colleges, and vocational schools to provide education on the fundamentals of computer networking. The CNAP is four-semester curriculum teaching the principles and practice of designing, implementing, and maintaining computer networks capable of supporting local, national, and global organizations. Instruction is provided in classrooms as well as through online curriculum and activities; assessments are likewise conducted through classroom exercises and online testing. The CNAP uses the World Wide Web for both instruction and, more importantly for our purposes, assessment administration and data maintenance. World Wide Web usage facilitates global access to educational resources, with approximately 150,000 students in 60 countries participating in the CNAP and an average of 10,000 administrations of online assessments per day. This high-volume global access presents both opportunities and challenges for educational research. Computer networking demands considerable technical knowledge as well as strategic and procedural expertise to become accomplished at common networking tasks. Despite the importance of these domain abilities current web-administered assessments consist of multiple-choice items primarily assessing declarative knowledge. As such, the then-current online assessments were inadequate for determining student understanding of some of the most important aspects of networking ability, namely those aspects involved in hands-on applications of designing, installing, and maintaining networks. The absence of standardized assessment of critical elements of ability force a reliance on local evaluation efforts, which may be prone to substantial variability in curriculum emphasis and evaluation standards.

The limited scope of standardized assessment and the potential for substantial variability in student capabilities in critical areas of computer networking ability is being addressed through a redesign of the CNAP networking assessment program. Work in redesigning the networking assessment program to more appropriately measure critical computer networking abilities combines current simulation technology and remote connection capabilities to produce an online assessment exercising the cognitive aspects of network design, implementation, and troubleshooting. The initial outcome is a prototype assessment, called NetPASS, using network simulations with realistic interactive tasks to measure students' abilities and provide targeted educational feedback. This feedback includes reporting on the students' knowledge of networking, their mastery of various networking skills, their ability to carry out procedures and strategies for networking tasks, and their misconceptions about network functionality and operational procedures. An Evidence Centered Design process (Mislevy, Steinberg, & Almond, in press) was employed in this redesign, providing the framework necessary to meet the design needs for such a complex computerized assessment in a technical domain.

Assessment Framework: Evidence Centered Design

Technology exists such that assessments can incorporate sophisticated and intricate simulations that require examinees to draw on relevant knowledge, skills, and abilities to solve meaningful tasks. To qualify as an assessment, a simulation system must evoke, record, and interpret observable evidence in a justifiable manner; this requires that the desired inferences play a role in the development of the assessment at each stage. Ad hoc procedures such as constructing an assessment, considering it, and then asking, "How do I score it?" or "How do I interpret the results?" are insufficient.

Until recently, the grounding in sound design principles and corresponding methods for developing, administering, scoring, and maintaining such innovative assessments was not sufficiently developed to assure that resultant assessments would meet professional standards. Evidence Centered Assessment Design (ECD; Mislevy, Steinberg, & Almond, in press) leverages knowledge of cognition in the domain and sound design principles to provide a robust framework for designing assessments, be they simple and familiar or complex and innovative. The application of ECD to simple designs provides a re-usable assessment blueprint, which facilitates developing assessments that meet professional measurement standards. For complex and innovative measurement needs, such as the one considered in this work, the ECD framework provides for assessment design that maintains an evidentiary focus, to guide the professional through the complexities of innovative design. Even a coarse treatment of ECD is beyond the scope of this paper. For more on ECD the reader is referred to Mislevy, Steinberg, and Almond (in press); for a thorough explication of the application of ECD to the NetPASS assessment, the reader is referred to Williamson et al., (2003). For our purposes, it will be sufficient to summarize several key aspects of ECD, namely those that most lend themselves to the construction of an evidentiary argument.

The purpose of any assessment is to gain relevant information about the examinees; just what information is relevant and how to obtain such information guides the development of an assessment instrument. More technically, we define *claims* as specific statements about the examinee's knowledge, skills, and abilities made on the basis of observable evidence from the assessment. Once these claims of interest are established, they form the basis for the nature and the extent of the required evidence; following Schum (1987), observable *data* only qualifies as *evidence* when it is relevant to some claim.

Following the work of Messick (1994), we advocate a construct-centered approach wherein the assessment is built from the foundation up by considering the construct(s) of interest. Indeed:

> A construct-centered approach [to assessment design] would begin by asking what complex of knowledge, skills, or other attributes should be assessed, presumably because they are tied to explicit or implicit objectives of instruction or are otherwise valued by society. Next, what behaviors or performances should reveal those constructs, and what tasks or situations should elicit those behaviors? Thus the nature of the construct guides the selection or construction of relevant tasks as well as the rational development of construct-based scoring criteria and rubrics. (pg. 17)

What claims would we like to make? Given these claims, what evidence is required to enable us to make such claims? Given the evidence we want to collect, what tasks should we present to examinees that will yield work products from which we can extract such evidence? To address these questions, we rely on a Cognitive Task Analysis (CTA; Mislevy et al., 1999). A CTA is a detailed approach to laying the foundation for evidentiary arguments regarding examinee proficiencies. In general, the CTA aims to identify knowledge, abilities, and strategies examinees utilize to address tasks. A CTA may include investigating

- essential features of tasks for eliciting behaviors
- internal representations of the task situations
- relationships between internal representations and the problem-solving behavior
- processes used to solve problems
- characteristics of the tasks that impact the problem-solving processes.

In performing the CTA specific to NetPASS, which grounds the assessment, the assessment team of psychometricians and subject matter experts (SMEs) analyzed the performance of CNA students at different levels of expertise under standard conditions across a range of tasks. Ultimately, the goal of performing a CTA is to determine the answers to Messick's questions posed above: (1) what are the knowledge, skills, and abilities required to solve problems? (2) what behaviors manifest themselves at different levels of domain proficiency? and (3) what features of tasks evoke those distinct behaviors? Corresponding to the three objectives in the CTA, a Conceptual Assessment Framework (CAF) consists of three main models. In addition, the CAF (see Figure 1) contains a fourth model; each is described below. For more on the CTA and is importance to grounding and designing assessments, see Mislevy et al. (1999); for more on the CTA performed in designing NetPASS see Williamson et al., (2003).

The Student Model

The Student Model addresses the question of what claims we would like to make; in Messick's words "what complex of knowledge, skills, or other attributes should be assessed?" The variables in the Student Model represent the knowledge, skills, abilities, or other relevant attributes that are of interest to those administering or gaining information from the assessment. The Student Model can be in different states depending on the different values of the variables. The state of the Student Model represents belief regarding an examinee's proficiencies. It is based on these variables that we make decisions, evaluate, and/or plan future instruction.

Depending on the nature or purpose of the assessment, there may be few or many variables of interest. For more general applications, such as a pass/fail decision, one Student Model variable may suffice; for more focused applications, many Student Model variables may be required, for instance in the case where feedback about student performance on many distinct and differentiable aspects is preferred. When there are multiple Student Model variables, it is of utmost importance to capture the relationships among them. One of the main purposes of this paper is to examine several such relationships prominent in NetPASS from a probability-based inferential standpoint, which will be discussed in detail below. A few more quite general points about the Student Model and the variables therein are in order. First, the assessment design team develops descriptions of the ways knowledge represented in the domain and of those features that provide evidence about the relevant knowledge and skills possessed. Second, SMEs knowledgeable about the ways students accumulate and demonstrate knowledge, skills, and abilities can design assessment instruments in which multiple aspects of such knowledge, skills, and abilities are required in predictable ways. Such assessments can be as simple as a two-stage question in which the successful completion of the second part requires the successful completion of the first part. The required relationships between aspects of knowledge, skills, and abilities constitute evidentiary structures via which inferences can be made from observable to Student Model variables. Finally, because we cannot observe the Student Model variables, we do not ever their values with certainty; we utilize probability based reasoning to estimate their values and to estimate our uncertainty.

Evidence Models

Evidence Models address the question of what observable evidence is required to enable us to make claims about non-observable constructs; in Messick's words "what behaviors or performances should reveal those constructs [of interest]?" An Evidence Model defines how the observable features of work products constitute evidence about Student Model variables. These models consist of (1) the evaluation component containing *rules of evidence* in which features of the work product are identified as observable variables that amount to evidence for inferences about examinees, and (2) the *statistical*, or *measurement model,* which defines how these observables should influence belief about values of the Student Model variables. The former can take on many forms e.g., an answer key, a rubric, etc. The latter will be the focus of this work. Thus as the major aim of this paper is the explication of and the relationships between variables, be they Student Model variables or observable variables, the vast majority of what follows will presume the first component of Evidence Models – the rules for extracting evidence from work products – is already in place, and that what remains are observable variables that will serve as evidence.

The measurement model contains deductive reasoning structures that facilitate the estimates of likely values of observable variables (data) given the state of the Student Model and also support the inductive reasoning from observed variables to probabilities of the states of the Student Model variables (Mislevy, 1994). The probability model quantifies the relationship between evidence and a claim. This relationship is the backbone of the justification of an assessment in terms of its validity and utility.

Task Models

Task Models address the question of what tasks we should present to examinees to yield work products from which we can extract evidence; in Messick's words "what tasks or situations should elicit those behaviors?" Task Models define the content and structure of tasks such that examinees will respond to the tasks with work products from which evidence can be extracted. Specific elements of Task Models include (1) performance situations, (2) material presented to the examinee, and (3) student work produced in responding to the task. While it is possible for an assessment to employ only one Task Model, most assessment use many Task Models. Regardless of the number Task Models, we can think of them as templates to construct or generate tasks. For instance, an assessment with 10 multiple choice items and 10 free response items might have just two Task Models, one for each format. Tasks that share the same Task Model share similarities in task requirements and the student work produced in response.

Summary

In an existing assessment, Task Models serve as templates for the various types of tasks. The Assembly Model defines the strategy or in some cases the algorithm used to select and present tasks to an examinee. The Evidence Model employs the evaluation component to extract the relevant features of the work products produced in response to the task that will count as observables. The Evidence Model then takes those observables and enters them into the measurement model. Beliefs about Student Model variables are then updated by propagating the evidence throughout the Student Model. The state of the Student Model – the current probabilities of the values of the Student Model variables – represents our belief about the examinee.

The main focus of this paper is to describe in detail the construction and refinement of the measurement model, the mechanism through which observable evidence affects our beliefs about the knowledge, skills, and abilities of examinees. We now turn to a more complete elaboration of the Student Model and the Evidence Model, with an eye toward the NetPASS application. To that end we begin with a brief description of Bayes' Theorem, followed by the extension to a Bayesian Inference Network, which will then in turn be used to represent the Student Model and the Evidence Models.

Bayes' Theorem

To develop the inferential machinery to be employed, we begin by introducing several probabilistic relationships. It is well known that the joint probability of two events is equal to the product of the probability of one event occurring and the probability of the second event occurring, given that the first event occurred[1]:

$$P(X,Y)=P(Y)\times P(X\mid Y) \tag{1}$$

where $P(X,Y)$ is the joint probability of event *X* and event *Y* occurring, $P(Y)$ is the probability of event *Y* occurring, and $P(X\mid Y)$ is the (conditional) probability of event *X* occurring given that event *Y* occurred. Let us now impose some structure on the variables. Let *X* be a variable whose probability distribution $P(X\mid Y)$ depends on some variable *Y*. Let $P(Y)$ be the probability distribution for *Y*, prior to observing *X*; this distribution represents our belief about the value of *Y*

[1] Note that the terminology 'second event' does not necessarily refer to a temporal order among events

before observing *X* and is known as the *prior distribution* for *Y*. Once *X*, some datum, is observed, Bayes' Theorem implies that the updated distribution of *Y* is:

$$P(Y \mid X) = \frac{P(Y) \times P(X \mid Y)}{P(X)} \tag{2}$$

where $P(X)$ acts as a normalizing constant whose analytical derivation or approximation is often problematic or computationally intractable. Removing $P(X)$ from the denominator on the right side of eq. (2) renders the two sides unequal but proportional:

$$P(Y \mid X) \propto P(Y) \times P(X \mid Y)\,. \tag{3}$$

That is, the *posterior distribution* of *Y*, conditional on the observed value of *X*, is proportional to the product of (1) the prior distribution of *Y* and (2) the conditional distribution of *X* given *Y*, $P(X \mid Y)$, also known as the *likelihood* of *Y*. Again, if probability distributions are expressions of our beliefs, the change in our beliefs about *Y* from the time before *X* is observed to the time after *X* is observed is a function of this likelihood expression. These likelihoods – the relative probabilities of the observed value (of *X*) given the possible states (of *Y*) that may have produced the observed value – allow for the deductive reasoning about the possible values of *X*, given the value of *Y*. And as just shown, these likelihoods also allow, via Bayes' Theorem, for inductive reasoning about the possible values of *Y*, once some datum, *X*, is observed (Jensen, 1996; Mislevy, 1994).

When the number of variables in a problem increases, the application of Bayes' Theorem in its form given in eq. (3) becomes computationally intractable. However, more efficient techniques to represent variables and apply Bayes' Theorem across a large system of variables have been developed in the form of Bayesian Inference Networks. We put forth a less technical discussion of Bayesian Inference Networks and focus on their application to NetPASS. For a more technical discussion of Bayesian Inference Networks see Jensen (1996; 2001) and Spiegelhalter et al., (1993); for the original proofs of the computational algorithms underlying them, see Lauritzen and Spiegelhalter (1988).

Bayesian Inference Networks

In the context of the CAF, a Bayesian Inference Network (BIN) (Jensen, 1996; 2001) serves as the statistical model for updating Student Model variables (see Martin & VanLehn, 1995 and Mislevy, 1994 on the use of BINs in assessment). BINs support probability-based reasoning as a means of transmitting complex observational evidence throughout a network of interrelated variables. As such, a BIN provides the means for *deductive reasoning* (from generals, such as knowledge, to particulars, such as observable behaviors) and *inductive reasoning* (from particulars to generals) required for producing appropriate inferences from complex observations (Schum, 1987). The relationships of variables in a BIN constitute the reasoning structures of the network. As in the preceding discussion of Bayes' Theorem, the likelihoods within the network that define the deductive reasoning structures—likely values of data given states of the student model—support subsequent inductive reasoning from the observed data to probabilities of the states of student model variables (Mislevy, 1994).

A BIN is a graphical model (of which Figure 2 is an example) of a joint probability distribution over a set of random variables, and consists of the following elements (Jensen, 1996):

- A set of variables (represented by ellipses and referred to as *nodes*) with a set of *directed edges* (represented by arrows) between nodes indicating the statistical dependence between variables. Nodes at the source of a directed edge are referred to as "parents" of nodes at the destination of the directed edge, their "children." In Figure 2, for example, *Design* is a child of *Network Proficiency* while both *Network Disciplinary Knowledge* and *Network Modeling* are parents of *Network Proficiency*. Similarly, *Network Disciplinary Knowledge* is an ancestor of *Design*, while *Design* is a descendent of *Network Disciplinary Knowledge*.

- The absence of an edge between two variables indicates a conditional independence between them, given variables on the path(s) between them. For example, the variables *Design* and *Network Disciplinary Knowledge* are independent if the value of *Network Proficiency* is known. For a more general discussion of this concept, D-separation, see Jensen (1996; 2001).
- For discrete variables, each has a set of exhaustive and mutually exclusive states. For continuous variables, the distribution of variable values is defined by a probability distribution.
- The variables and the directed edges together form what is commonly referred to as a directed acyclic graph (DAG; Brooks, 1998; Edwards, 1998; Jensen, 1996; Pearl, 1988). These graphs are directed in that the edges follow a 'flow' of dependence in a single direction (i.e., the arrows are always unidirectional rather than bi-directional). The graphs are acyclic in that following the directional flow of directed edges from any node it is impossible to return to the node of origin.
- For each endogenous variable, there is an associated set of conditional probability distributions corresponding to each possible pattern of values of the parents. These distributions are graphically represented squares; the connections between variables are routed through these relationships.

- For each exogenous variable, an unconditional probability table or distribution must be specified; in Figure 2, *Network Disciplinary Knowledge* has no parents, hence, a distribution for its values must be specified. This distribution is also represented graphically with a square; note that there are no directed edges flowing into the square.

As described below, both the Student Model and the Evidence Models can be conceived of as BINs. Before turning to the probability framework used to represent these models, let us pursue further into the potential characteristics of the manner in which variables may relate to one another.

Relationships Among Variables

This section sketches out a variety of evidentiary structures among the Student Model and observable variables. Four general relationships are defined, followed by more specific relationships. Though certainly not an exhaustive set of all possible structures, these structures appear repeatedly in the NetPASS assessment, and so the structures will first be described and later illustrated by examples from NetPASS.

General Relationships

Here we introduce very general concepts that will be used in the explication of more specific relationships in NetPASS.

- Independence relationships: Two variables are independent of one another when the level of one variable does not affect the other variable; that is, the (conditional) distribution of one variable conditional on the other remains constant at all levels. In BINs, updating the probabilities for the values of one variable causes no change in the probabilities for the levels of the other variable, though such a change may impact the probabilities for the levels of variables that are descendants of both.

- Dependence relationships: Two variables are dependent if the level of one variable affects the other variable; that is, conditional distributions of one variable at levels of the other variable vary from level to level. In BINs, updating the probabilities for the values of one variable causes changes in the probabilities for the levels of the other variable. We further define *direct* dependence as the case where a relationship is conceived of between two variables directly, as opposed to via a third variable.
- Conditional independence relationships: Two variables are dependent though their dependence is explained by a third variable; that is, keeping the third variable constant, the variables are independent. More formally, if B is conditionally independent of C (given A),

  $$P(B \mid A) = P(B \mid A, C) \qquad (4)$$
- Conditional dependence relationships: Two variables are dependent above and beyond that which can be explained by a third variable; that is, even at constant levels of the third variable, the two variables are related. Negotiating conditional dependence, i.e., achieving conditional independence, plays a key role in constructing BINs, particularly for assessment purposes.

Bivariate Relationships in Modeling Skills Involved in an Assessment

Let bivariate relationships be those between two variables; in the framework of BINs, this corresponds to the case of modeling a variable as a child of a single parent. Two bivariate relationships appearing in NetPASS are presented below in the context of relating cognitive skills.

- Direct Dependence: There is an expected relation between two skills where the value of one dictates the expectation of the other in the form of a probability distribution.
- Ceiling: There is an expected relation between two skills where the value of one not only dictates the expectation of the other, but sets a maximum value that the other one can take. One instance of a ceiling is where the ceiling is set to be the value of the first skill. That is, define an expected distribution for B based on the value of A such that B cannot exceed A.

Multivariate Relationships in Modeling Performance on an Assessment

Multivariate relationships are those involving at least three variables; in the framework of BINs, this corresponds to the case of modeling a variable as a child of multiple parents. Most of the multivariate relationships discussed here can be properly thought of as generalizations of the bivariate relationships discussed above. To illustrate the applicability of various relationships to different aspects of assessment we present the multivariate relationships in the context of modeling performance; that is, performance is modeled as a child of multiple cognitive skills and abilities.

- Conjunctions: A generalization of the ceiling relationship. Multiple skills impact performance such that the minimum value of the skills defines the ceiling for performance; the absence of any of these required skills causes expectation of lower levels of performance. Conjunctions correspond to the logical term 'and,' indicating that the *joint* occurrence or instantiation is required.
- Compensatory Relationships: A generalization of the direct dependence relationship. Multiple skills impact performance such that the increase in *any* of these skills (not just the lowest, as in conjunctions) causes expectation of an increase in performance.
- Conditional dependence relationships: Multiple skills affect performance. Conditional dependence relationships occur among observable variables, indicating that the observable variables are related in ways *above and beyond* those determined by their parent skills. The consequences of ignoring these relationships can be deleterious in estimating the values of variables and the precision of the estimates (Mislevy & Patz, 1995; Patz, Junker, & Johnson, 2000).

It should be noted that these basic structures represent just a small portion of the limitless number of ways to model relationships. For other common structures, see Mislevy, Senturk, et al. (in press). Again, though estimates of these relationships can come from data, familiarity and understanding of the knowledge, skills, and abilities of the domain of interest can contribute both to defining their form and values.

The Probability Framework

Gelman et al. (1995, p. 3) define the first step in conducting a Bayesian analysis as setting up a full probability model, specifically, a joint distribution of all quantities, observable and unobservable. Furthermore they note "the model should be consistent with knowledge about the underlying scientific problem and the data collection process." In assessment, this "knowledge" is knowledge about the domain of interest, specifying the (1) targeted knowledge, skills and abilities, (2) ways in which such knowledge, skills, and abilities are manifestly demonstrated in performance, and (3) characteristics of situations that provide the opportunity to observe such performance. As discussed above, the Student Model, Evidence Models, and Task Models provide this very knowledge.

The Probability Model

The Student Model contains unobservable variables characterizing examinee proficiency on the knowledge, skills, and abilities of interest. For the i^{th} examinee, let

$$\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{iP}) \tag{5}$$

be the vector of *P* Student Model variables. The complete Student Model for all examinees is denoted $\boldsymbol{\theta}$.

Task Models define those characteristics of a task that need to be specified. Such characteristics are expressed by Task Model variables; for task *j*, these variables are denoted by the vector

$$\mathbf{Y}_j = (Y_{j1}, \ldots, Y_{jL}) \tag{6}$$

where *L* is the number of Task Model variables. The full collection of Task Model variables is denoted $\mathbf{Y}$.

The evaluation component of Evidence Models defines how to extract relevant features from an examinee's response to a task (work products) to yield the values of observable variables. Let

$$\mathbf{X}_j = \left(X_{j1},\ldots,X_{jM}\right) \tag{7}$$

be the vector of *M* potentially observable variables for task *j*. X_{imj} is then the value of observable *m* from the administration of task *j* to examinee *i*. The complete collection of values of observable variables, that is, the values for all observables from all tasks for all examinees is denoted as $\mathbf{X}$. As the focus of this paper is not on the generation of tasks from Task Models, nor is it the extracting of observables from work products via the evaluation component of Evidence Models, let us assume these important procedures have been completed, leaving us with a set of observables.

The BIN for the Student Model is a probability distribution for $\boldsymbol{\theta}_i$. An assumption of exchangeability results in a common prior distribution, i.e., before any responses to tasks are observed the Student Model is in the same state for all examinees. Beliefs about the expected values and associations among the Student Model variables are expressed through the structure of the model and higher level hyperparameters $\boldsymbol{\lambda}$. Thus, for all examinees,

$$\boldsymbol{\theta}_i \sim P(\boldsymbol{\theta}_i \mid \boldsymbol{\lambda}) \tag{8}$$

The higher level parameters, $\boldsymbol{\lambda}$, define the prior expectations. In the absence of a strong theory regarding the prior distribution of examinee proficiencies, as is the case with NetPASS, these parameters should be set such that $P(\boldsymbol{\theta}_i \mid \boldsymbol{\lambda})$ is vague.

For any given examinee, the statistical model defines how the observable variables, X_{imj}, are dependent on that examinee's values of the Student Model variables, $\boldsymbol{\theta}_i$. Let π_{mjk} be the probability or responding to observable *m* from task *j* with a value of *k*. The collection of these, for any particular observable, is then

$$\pi_{mj} = (\pi_{mj1}, \pi_{mj2}, \ldots, \pi_{mjK}) \tag{9}$$

where *K* is the number of different values observable *m* from task *j* may take on. π_{mj} is then the probability structure associated with observable *m* from task *j*, i.e., the conditional probability of X_{imj} given $\boldsymbol{\theta}_i$. More formally, if

$$\pi_{mjk} = P(X_{imj} = x_{imjk} \mid \boldsymbol{\theta}_i), \tag{10}$$

the distribution of the values for observable *m* from task *j* for examinee *i* is then

$$X_{imj} \sim P(X_{imj} \mid \boldsymbol{\theta}_i, \pi_{mj}) \tag{11}$$

In short, for any examinee, the distribution for the observables is defined by the values of the Student Model variables and the conditional distributions of observables given Student Model variables. Thus if we knew both the values of the Student Model variables and the conditional distribution of observables given Student Model variables, we would know the distribution of the observables. Of course in practice, the situation with the Student Model variables and the observables is reversed: we have values for the observables but not the Student Model variables; hence the use of Bayes' Theorem to reason from observations to Student Model variables.

When there are a large number of levels of Student Model variables and/or of the observable, there are a very large number of π_{mjk}'s. It may be the case that further structure exists for modeling the π_{mj}'s. More formally, we may express this as

$$\pi_{mj} \sim P(\pi_{mj} \mid \eta_{mj}) \tag{12}$$

where η_{mj} are higher level hyperparameters for observable *m* (e.g., characteristics of the appropriate Evidence Model and the task *j* from which *m* is obtained); prior beliefs about such parameters are expressed through higher-level distributions, $P(\eta_{mj})$. The complete set of conditional probability distributions for all Evidence Models for all observables is denoted $\boldsymbol{\pi}$; the complete set of prior probabilities for those distributions is denoted $\boldsymbol{\eta}$.

The joint probability of all parameters can be expressed as

$$P(\boldsymbol{\lambda},\boldsymbol{\eta},\boldsymbol{\theta},\boldsymbol{\pi},\mathbf{X})=P(\boldsymbol{\lambda})\times P(\boldsymbol{\eta}\mid\boldsymbol{\lambda})\times P(\boldsymbol{\theta}\mid\boldsymbol{\lambda},\boldsymbol{\eta})\times P(\boldsymbol{\pi}\mid\boldsymbol{\lambda},\boldsymbol{\eta},\boldsymbol{\theta})\times P(\mathbf{X}\mid\boldsymbol{\lambda},\boldsymbol{\eta},\boldsymbol{\theta},\boldsymbol{\pi}) \qquad (13)$$

This expression can be simplified in light of additional knowledge and assumptions we bring to the assessment context.

Simplification of the Probability Model

Eq. (11) states that the distribution of an examinee's response, X_{imj}, is defined by $\boldsymbol{\theta}_i$ and π_{mj}. As such, the set of examinee responses, $\mathbf{X}$, are conditionally independent given $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$. The distribution of the responses in not contingent on the other parameters in the model, $\boldsymbol{\lambda}$ and $\boldsymbol{\eta}$. The fifth term on the right side of eq. (13) therefore simplifies to

$$P(\mathbf{X}\mid\boldsymbol{\lambda},\boldsymbol{\eta},\boldsymbol{\theta},\boldsymbol{\pi})=P(\mathbf{X}\mid\boldsymbol{\theta},\boldsymbol{\pi}) \qquad (14)$$

Taking advantage of the other conditional independence relationships described above and rearranging terms, the full joint distribution can be parsimoniously represented as:

$$P(\boldsymbol{\lambda},\boldsymbol{\eta},\boldsymbol{\theta},\boldsymbol{\pi},\mathbf{X})=P(\boldsymbol{\lambda})\times P(\boldsymbol{\theta}\mid\boldsymbol{\lambda})\times P(\boldsymbol{\eta})\times P(\boldsymbol{\pi}\mid\boldsymbol{\eta})\times P(\mathbf{X}\mid\boldsymbol{\theta},\boldsymbol{\pi}). \qquad (15)$$

Intuitively, the (the last term on the) right side of this equation states that the probability distribution of the observable is defined by the Student Model variables and the conditional distribution for that observable. The Student Model variables are distributed conditionally on higher-level parameters (the second term), which have their own distribution (the first term). The conditional probability distributions are distributed conditionally on higher-level parameters (the fourth term), which have their own distribution (the third term). In setting up the full model, our goal then becomes to define the various terms in eq. (15). We have already mentioned that we might properly think of observable variables as conditional on latent Student Model variables. In a complex assessment, such as NetPASS, which includes multiple Student Model variables that are related, there becomes the need to model the dependencies among the Student Model variables. We therefore extend the notion of modeling observables as conditional on Student Model variables to modeling the Student Model variables as conditional on other Student Model variables. Much of the discussion regarding modeling obverables conditional on Student Model variables via the π_{mj} terms can be extended to modeling Student Model variables as conditional on others via their own conditional probability distributions. Before turning to the specification of the Student Model, we introduce a more efficient manner for modeling conditional dependencies.

Samejima's Graded Response Model

One procedure for modeling the conditional probabilities of a variable given its parent is by directly estimating the probabilities themselves (Spiegelhalter et al., 1993). This procedure quickly becomes unwieldy as the number of levels of the parent(s) or child increases. We therefore seek a more efficient way to model the conditional probabilities. To that end, we turn toward item response theory (IRT) for parsimonious ways of modeling conditional probabilities.

The Graded Response Model

Typical models for modeling variables as conditional on other variables are IRT models, the most common of which are unidimensional models for binary responses (Hambleton and

Swaminathan, 1985). The two-parameter logistic unidimensional model for binary (0/1) responses is

$$\mathrm{logit}(P(X_{ij}=1\,|\,\theta_i)) = a_j(\theta_i - b_j) \tag{16}$$

where a_j and b_j are the scale and location parameters, respectively, that define item j.[2] Samejima's Graded Response Model (GRM; 1969) extends this to the more general case where the outcome variable X_{ij} is not restricted to be binary, but instead is polytomous, though still ordinal. That is, the observable, X_{ij} can take on any integral value from 1 to *K*. Define the probability that the response is in category *k* or above as

$$P(X_{ij} \geq k) = \mathrm{logit}^{-1}(a_j(\theta_i - b_{jk})) \tag{17}$$

for *k*=2,…,*K* and b_{jk} is the location parameter associated with separating the *k*th from the *k*-1th category. Note that

$$P(X_{ij} \geq 1) = 1, \tag{18}$$

for the probability of response being in the lowest category or above is 1, and

$$P(X_{ij} \geq K+1) = 0, \tag{19}$$

for the probability of response being above the highest category is 0. The probability of response being in a category alone (and not in that category or above) may be calculated by subtracting different instantiations of eq. (17). More technically, the probability of response being in category *k* is

$$P(X_{ij} = k) = P(X_{ij} \geq k) - P(X_{ij} \geq k+1) \tag{20}$$

[2] Eq. (16), and much of the subsequent discussion of IRT models, follows customary presentations of IRT and indexes responses in terms of examinees (*i*), items (*j*), and (where relevant) categories (*k*). In terms of the language of ECD and the notation used above, we can think of items as being different tasks where there is only one observable extracted from the task.

Let us illustrate this and return to the issue of location parameters and response categories by way of an example. Consider the case where there are 3 response categories. There are $K-1 = 2$ location parameters. The first location parameter marks the location of the separation between response category 1 and response category 2; the second location parameter marks the location of the separation between response category 2 and response category 3. To continue the illustration, let us fix these location parameters at –2 and +2 and fix the scale parameter to +1. Since the value of eq. (17) for k=1 is 1 (eq. (18)), let us instantiate eq. (17) for k=2,...,K:

$$P(X_{ij} \geq 2) = \text{logit}^{-1}(1 \times (\theta_i - (-2))) \tag{21}$$

$$P(X_{ij} \geq 3) = \text{logit}^{-1}(1 \times (\theta_i - (+2))) \tag{22}$$

Instantiate eq. (20) to obtain the probabilities for being in each response category.

$$P(X_{ij} = 1) = P(X_{ij} \geq 1) - P(X_{ij} \geq 2) = 1 - P(X_{ij} \geq 2) \tag{23}$$

$$P(X_{ij} = 2) = P(X_{ij} \geq 2) - P(X_{ij} \geq 3); \tag{24}$$

$$P(X_{ij} = 3) = P(X_{ij} \geq 3) - P(X_{ij} \geq 3+1) = P(X_{ij} \geq 3) \tag{25}$$

These probabilities of response are, ultimately, functions of theta. Figure 3 plots the probabilities of each response for any value of theta obtained from a GRM with $a_j = 1$ and $\mathbf{b} = (-2, +2)$.

In such a case as been illustrated here, in order to model the 15-cell conditional probability table of a child variable that has three levels conditional on a parent that has five levels, we need only estimate three parameters, the discrimination a_j and the two category boundaries contained in $\mathbf{b}$.

Applications in NetPASS

Though the above illustration depicts a variable with three categories, the logic of this model can be extended to fit observables with any number of categories. When the GRM is employed to model observed responses in the Evidence Models, we will use a model with three categories, as there are three possible values (Low, Medium, High) for the observed variables. Nothing in the GRM restricts its use to when one variable (i.e., the child variable) is observed. Indeed, we employ the GRM to model latent variables as conditional on other latent variables in the Student Model and in the Evidence Models. In these cases, we will use a model with five categories, as latent variables can take on any of five possible values (novice, semester 1, semester 2, semester 3, semester 4). The only change from the preceding discussion is that the child variable can take on values from 1 to 5 (rather than 1, 2, or 3) and that there needs to be 4 (rather than 2) location parameters (to partition the continuum into 5 rather than 3 levels). Though the presentation of the GRM in the preceding section has been couched in terms of observed responses (to items), it may more generally be said to model conditional probabilities, where the conditional probabilities of an observed response is just one special case. In the case where the child variable is latent, we might call the conditional probabilities defined in the expressions above "probabilities of an examinee having a certain value for the child variable" rather than "probabilities of an examinee giving a certain response to the item." For another example of using an ordered polytomous IRT model to model latent variables, see Patz, Junker, & Johnson (2000).

In all the instances in NetPASS, we will assume the category boundaries are equally spaced apart. In this case, we need not estimate K–1 category boundaries, but just one location parameter creating an even more parsimonious representation. Future work may include releasing this additional constraint to allow for unequally spaced category boundaries.

The Effective Theta Method

The GRM, like most IRT models, is unidimensional: there is one variable, θ_i, that serves as the parent for the observables. Complex assessments such as NetPASS involve many variables and, more importantly, conceptualize observables as being dependent on more than one variable. Thus, we must either implement a multivariate IRT (e.g. Reckase, 1985; Sympson, 1978) model or distill down the relationships between multiple parents and children to fit the unidimensional GRM. We proceed by adopting the latter strategy and take the following steps. First, we adopt a set of parameters that will remain constant throughout, a_{mj} and $\mathbf{b}_{mj}$. Next we seek to combine the parent variables in such a manner to produce one variable that will serve in the unidimensional GRM; this variable is an "effective theta" denoted as θ_i^{**}. In IRT models, the conditional probabilities of response are determined by theta and the "item" parameters a_{mj} and $\mathbf{b}_{mj}$[3]. In fixing these parameters the conditional probabilities are then a function of the effective theta, which itself is a function of the parent variables. Coefficients and intercepts in the calculation of the effective theta are akin to scale and location parameters in usual IRT formulations. In essence, this is simply a shift in the estimation. Typical IRT models posit an examinee's latent trait(s) as being constant and estimate the items (in terms of a_{mj} and $\mathbf{b}_{mj}$) accordingly. Instead, the effective theta method holds the scale constant (by fixing a_{mj} and $\mathbf{b}_{mj}$),

[3] For ease of exposition, we will continue to discuss the effective theta method in terms of items (i.e., an observable child variable). Like the GRM, the effective theta method is by no means restricted to case of observable child variables.

and estimates the examinee's latent trait(s) with respect to each item. The impact of the item, both in terms of overall difficulty and association to examinee proficiencies, is part of the calculation of the effective theta.

The effective theta method brings two distinct advantages (Mislevy, Senturk, et al., in press). First, this may be more comforting to SMEs, who while familiar with the domain and the structure of knowledge and able to provide the form of the relationships (e.g., "familiarity with either procedure A or B is sufficient," or "once an examinee has skill A performance becomes mainly a function of skill B", etc.) may not feel comfortable specifying a complete conditional probability table. Second, unidimensional IRT models are quite popular in the psychometric community and now the problem is on a scale familiar to experts in educational measurement. Thus, they may feel more comfortable with capturing and modeling knowledge elicited from the SMEs. For example, if experts believe that an item is easier than most or is very closely related to proficiency, we have a solid idea about just what the values of the parameters should be. Of course, these values are by no means fixed. Our approach is to elicit initial opinions from SMEs, quantify them by assigning numerical priors, and then refine the values based on pretest data and pilot testing.

Unidimensional Models

In the case where a variable, θ_c, has one parent, θ_1, define the conditional probabilities as

$$\pi_k = P(\theta_c = k \mid \theta_1) \tag{26}$$

where, as before, k=1,...,K are the possible values of θ_2. We model the conditional probabilities, π_k for k=1,...,K, via a projection, or mapping, function $g(\theta_1)$ which we then enter into the GRM. A note about each of the mapping function and the GRM is required.

As will be described below, the relationships between all of the variables in NetPASS are positive. When constructing an effective theta from parent variables, the mapping function from the parent variable(s) to the effective theta should therefore be monotonic and positive. Figure 3 offers a visual depiction of this concept. There is a positive monotonic relationship between theta and the response category: as theta increases, the probabilities of higher levels of response increase.

Assuming the levels of θ_1 are roughly equally spaced apart, we code the values of θ_1 accordingly and define the effective theta via a linear function, $g(\theta_1)$, as the map:

$$\theta^{**} \equiv g(\theta_1) = c \times \theta_1 + d \tag{27}$$

Note the simplicity of the model: there are two parameters to estimate, c and d, *regardless* of the number of states of the parent, θ_1, or the child, θ_c. The effective theta can be thought of intuitively as the combination of the parent variable θ_1 and the features of the conditional distribution, represented by c and d.

We have specified the structure of $P(\theta_c \mid \theta_1, \pi_k)$ where the conditional probabilities, π_k, are defined by the parameters c and d. In IRT models *a* and *b* parameters define the conditional probability distribution. Thus it should come as no surprise that the two sets of parameters are related. The constant parameter, d, is akin to *b* in eq. (17) and is related to the average value for the child variable. The slope parameter, c, is akin to *a* in eq. (17) and defines the strength of association between θ_1 and θ_2. Higher values of the slope parameter indicate a stronger association between the parent and child. Higher values of the intercept parameter indicate that, on average, the value of the child is higher.[4] The slope and intercept parameters capture the conditional distribution; estimation of the conditional probability distribution thus becomes the estimation of these parameters.

Multidimensional Relationships

Consider the case where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_L)$; we must now build a mapping function, $f_t(\boldsymbol{\theta})$, to project a vector of variables onto an effective theta. In this section we will describe general categories of mapping functions for multidimensional.

- Combinations of linear mapping functions. For each parameter $\theta_l, l = 1, \ldots, L$, first define a (unidimensional) linear mapping function

  $$\theta^*_{tl} \equiv g_{tl}(\theta_l) = c_t \times (\theta_l) + d_t \qquad (28)$$

  to specify the marginal influence of θ_l on the value of child variable *t*. Then define a function

  $$\theta^{**}_t \equiv h_t(\theta^*_{t1}, \ldots, \theta^*_{tL}) \qquad (29)$$

  that defines how the aspects of proficiency interact to define proficiency for this particular outcome. This variable, θ^{**}_t, is the effective theta to be entered into the GRM. Compensatory relationships are of this form, examples of which will be provided below.
- Linear mapping functions of combinations. First define a function

[4] This marks a departure from more common formulations of IRT models where higher values of the intercept term indicate *lower* probabilities of the child taking on higher values; e.g., in more common binary IRT models (Hambleton and Swaminathan, 1985), higher *b* values indicate a more difficult item, with lower probabilities of correct response. The notation used here is consistent with that of Bock's (1972) slope-intercept form.

$$r_t(\theta_1,\ldots\theta_L) \equiv \theta_t^* \tag{30}$$

of the L Student Model variables that describes the form of their interaction, e.g., a ceiling. Next define a linear mapping function

$$\theta_t^{**} \equiv u_t(\theta_t^*) = c_t \times \theta_t^* + d_t \tag{31}$$

that adjusts for the sensitivity of the observable to the proficiency (via the slope) and average difficulty (via the intercept). Again, θ_t^{**} is the effective theta to be entered into the GRM. Conjunctive relationships are of this form, examples of which will be provided below. Again, only two parameters are required for these cases.

- Everything else. Many other structures and procedures for mapping a multidimensional space to a unidimensional effective theta can be constructed.

Examples discussed will be restricted to those relationships that appear in NetPASS. The reader interested in the quantification procedures for a number of other relationships is referred to Mislevy, Senturk, et al. (in press).

The Application of the Effective Theta Method to the GRM

The effective theta method fixes the parameters a_j and $\mathbf{b}$ in the GRM and models an effective theta as a function of the examinee proficiency variables and parameters (the slopes and intercept) that define the conditional distribution. When using the effective theta method and the GRM to model observed responses, $a = 1$ and $\mathbf{b} = (-2, +2)$. When using the effective theta method and the GRM to model values of latent variables, $a = 1$ and $\mathbf{b} = (-3, -1, +1, +3)$. The conditional distributions are captured in the coefficients and intercepts of the equation for the effective theta. The accurate modeling of the relationships in the Student Model and the Evidence Models and the estimation of these parameters constitute the calibration of the

NetPASS assessment. When the specific relationships in NetPASS are presented in the following sections, they will be illustrated with specific values for these parameters.

The Student Model

Properties of Student Model Variables

The Student Model, on the whole, aims to represent the knowledge, skills, and abilities that are important for success at CNAP. Figure 2 shows part of the NetPASS Student Model including all Student Model variables that are the target of inference in the assessment. The complete Student Model also includes the specification of statistical relationships among variables and other variables to facilitate statistical modeling, which will be addressed below.

All the variables described in this section are considered to be discrete, and can take on any of five values, couched in terms of CNAP's four semester courses: complete novice, semester 1, semester 2, semester 3, and semester 4, where the level indicates one's cognitive level on that particular aspect of the domain.

Quantitative Modeling of Relationships in the Student Model

In terms of the joint probability distribution (eq. (15)), the quantitative modeling of the relationships in the Student Model amounts to the specification of $P(\boldsymbol{\theta} \mid \boldsymbol{\lambda})$. Several relationships will be discussed, each followed by examples as they appear in NetPASS

Direct Dependence

With direct dependence, the value of the child is dependent on only one parent, which determines a probability distribution for the child.

Basic formulas.

Define the effective theta as a linear function of the lone parent variable:

$$\theta_c^{**} \equiv c_c \times \theta_1 + d_c \tag{32}$$

where θ_c^{**} is the effective theta to be used for the distribution of the child, and θ_1 is the parent variable

Examples from NetPASS.

Discussions with SMEs revealed that the relationships between *Design* and *Network Proficiency*, *Implement* and *Network Proficiency*, and *Troubleshoot* and *Network Proficiency* may be modeled as direct dependence relationships. To obtain the effective theta for *Design*, instantiate eq. (32):

$$\theta_{Design}^{**} \equiv c_{Design} \times \theta_{NetworkProficiency} + d_{Design} \tag{33}$$

Effective thetas calculated for all possible values of *Network Proficiency* with c_{Design} = 2 and d_{Design} = -5.8 and are given in Table 1. The values for c_{Design} and d_{Design} were chosen because when the resulting effective thetas are entered into the GRM to produce a conditional probability distribution (Table 1), the resulting distribution approximately matched the opinions and expectations of SMEs. We will eventually estimate the value of c_{Design} and d_{Design}; because values of 2 and –5.8, respectively, result in the conditional distribution experts expect, our prior distributions for each parameter will be based on these values.

To obtain the effective theta for *Implement*, again instantiate eq. (32):

$$\theta_{Implement}^{**} \equiv c_{Implement} \times \theta_{NetworkProficiency} + d_{Implement} \tag{34}$$

Effective thetas calculated for all possible values of *Network Proficiency* with $c_{Implement}$ = 2 and $d_{Implement}$ = -6.2 and are given in Table 2. The values for $c_{Implement}$ and $d_{Implement}$ were chosen because when the resulting effective thetas are entered into the GRM to produce a conditional probability distribution (Table 2), the resulting distribution approximately matched the opinions and expectations of SMEs; again, these values for $c_{Implement}$ and $d_{Implement}$ represent expert expectations and will serve as the basis for the prior distributions in the calibration.

Likewise, the effective theta for *Troubleshoot* is defined as:

$$\theta^{**}_{Troubleshoot} \equiv c_{Troubleshoot} \times \theta_{NetworkProficiency} + d_{Troubleshoot} \qquad (35)$$

Effective thetas calculated for all possible values of *Network Proficiency* with $c_{Troubleshoot}$ = 2 and $d_{Troubleshoot}$ = -7.0 and are given in Table 3, as is the resulting conditional probability distribution. As before, these values for $c_{Troubleshoot}$ and $d_{Troubleshoot}$ represent expert expectations and will serve as the basis for the prior distributions in the calibration.

To illustrate how these prior estimates mimic expert expectations, compare the values in Table 3 to the values in Tables 1 and 2; for all values of *Network Proficiency*, the effective theta for *Troubleshoot* is always lower than the effective theta for *Implement* which is always lower than the effective theta for *Design*. As a result, for all values of *Network Proficiency*, the probability of high levels is lower for *Troubleshoot* than for *Implement*, which is lower than for *Design*. This reflects SME expectation that *Design* is the easiest aspect of *Network Proficiency* to master, followed by *Implement*, followed by *Troubleshoot*.[5] Though our expectation is that the level of *Design* will be higher than the level of *Implement*, which will be higher than the level

[5] The expected difference in the ability to acquire the cognitive skills of *Design*, *Implement*, and *Troubleshoot* is entirely captured by the change in the expected intercept parameter, as the coefficient used in compiling Tables 1-3 is unchanged.

of *Troubleshoot* the model is not constrained so that this will always be true. There are no mathematical constraints to force *Design* to be higher than *Implement* and *Implement* to be higher than *Troubleshoot*. Indeed, should empirical evidence indicate otherwise, it is possible for this property of the conditional distributions to change.

Ceiling Relationships

Ceiling relationships are not unlike direct dependence relationships: in both cases, one parent determines the probability distribution for the child variable. The parent variable, or some transformation of it, sets the ceiling value for the child, which can take on any value at or below the ceiling.

Basic formulas.

The quantification of ceiling relationships is quite similar to that of direct dependence relationships. Define the effective theta as a linear function of the lone parent variable:

$$\theta_c^{**} \equiv c_c \times \theta_1 + d_c \qquad (36)$$

This effective theta is then entered into the GRM to produce a probability distribution for the values of the child. These values do not represent the correct probability distribution of the child, for the GRM allows for the child to take on values higher than the ceiling. We thus impose the ceiling structure and adjust the probability distribution accordingly by setting the probabilities for levels above the ceiling to 0 and renormalizing the remaining probabilities.

Examples from NetPASS.

Discussions with SMEs revealed that *Network Modeling* cannot be higher than *Network Disciplinary Knowledge*. To obtain the effective theta for *Network Modeling*, instantiate eq. (36):

$$\theta^{**}_{NetworkModeling} \equiv c_{NetworkModeling} \times \theta_{NetworkDisciplinaryKnowledge} + d_{NetworkModeling} \tag{37}$$

Table 4 contains the possible values for *Network Disciplinary Knowledge*, the values for the effective theta obtained with $c_{NetworkModeling} = 2$ and $d_{NetworkModeling} = -8.0$, and the probabilities that result from the GRM. This distribution does not reflect the ceiling structure hypothesized by the SMEs. This structure is imposed on the distribution by forcing probabilities for levels of *Network Modeling* above the level of *Network Disciplinary Knowledge* to 0 and renormalizing such that the conditional distributions, i.e., the rows in the table, sum to 1. These corrected probabilities are given in Table 5. Again, the values of the parameters in the model were selected to mimic expert expectation and will serve as the basis for the prior distribution for $c_{NetworkModeling}$ and $d_{NetworkModeling}$ in the calibration of the model.

Baseline-Ceiling Relationships

Define a relationship that involves two parents: one parent sets a baseline value and the other serves in a compensatory relationship with the first parent to define the effective theta. In addition, the first parent variable imposes a ceiling relationship on the resulting probabilities. The procedures for defining baseline relationships and implementing ceiling relationships have already been presented. A more complete explanation of compensatory relationships is deferred until later; it should be sufficient for our purposes now to say that compensatory in this context refers to an additive model.

Example in NetPASS.

Network Disciplinary Knowledge and *Network Modeling* serve as parents for *Network Proficiency* (Figure 2). Discussions with SMEs revealed that *Network Proficiency* cannot be higher than *Network Disciplinary Knowledge* and that *Network Proficiency* is expected to be higher then *Network Modeling*, though it is possible for the latter to be higher than the former. Furthermore, *Network Disciplinary Knowledge* is the primary contributing factor to *Network Proficiency* and that *Network Modeling* is a secondary factor; with *Network Disciplinary Knowledge* essentially serving as a prerequisite, *Network Modeling* serves as an additional compensatory variable. Therefore, a baseline based on *Network Disciplinary Knowledge* is used and then adjusted based on the value of *Network Modeling*.

Define the baseline theta as a linear transformation of *Network Disciplinary Knowledge* as

$$\theta^{*}_{NetworkProficiency} \equiv c_{baseline} \times \theta_{NetworkDisciplinaryKnowledge} + d_{baseline} \tag{38}$$

where $\theta_{NetworkDisciplinaryKnowledge}$ is the value of *Network Disciplinary Knowledge*. Define the effective theta as

$$\theta^{**}_{NetworkProficiency} \equiv \theta^{*}_{NetworkProficiency} + c_{compensatory}[\theta_{NetworkModeling} - (\theta_{NetworkDisciplinaryKnowledge} - 1)] \tag{39}$$

where in addition to those variables in eq. (38), $\theta_{NetworkModeling}$ is the value of *Network Modeling*. Consider the term in the brackets. *Network Modeling* can never be higher than *Network Disciplinary Knowledge*, thus the term in the brackets represents how much *Network Modeling* contributes above *Network Disciplinary Knowledge*. When *Network Modeling* is one level below *Network Disciplinary Knowledge* (as it is expected to be, as shown in Table 5), the contribution is 0. When *Network Modeling* is equal to *Network Disciplinary Knowledge*, the contribution is equal to the value of $c_{compensatory}$. When *Network Modeling* is two or more levels below *Network Disciplinary Knowledge*, the contribution is negative. The possible combinations of *Network Disciplinary Knowledge* and *Network Modeling* and the resulting effective thetas with $c_{baseline} = 2$, $d_{baseline} = -6.0$, and $c_{compensatory} = 1$ are given in Table 6. The effective theta obtained from eq. (39) is then entered into the GRM to obtain the conditional probability distribution for *Network Proficiency*, also given in Table 6. As with the previous ceiling relationship, the GRM does not retain the ceiling structure; the ceiling is imposed by setting all probabilities for levels of the child greater than the level of *Network Disciplinary Knowledge* to 0 and renormalizing the probabilities. The corrected probability distributions are given in Table 7. Again, the values of $c_{baseline}$, $d_{baseline}$, and $c_{compensatory}$ used here reflects expert opinions regarding the conditional probability distribution and will serve as the basis for the prior distributions.

Exogenous Variable

In specifying the distributions for *Network Modeling*, *Network Proficiency*, *Design*, *Implement*, and *Troubleshoot*, each of these variables was modeled as conditional on some other parent variable(s). To complete the specification of the Student Model, the lone exogenous variable, *Network Disciplinary Knowledge*, must also be specified. As NetPASS is intended to for the assessment of third semester students in the CNAP sequence, experts posited that the majority of examinees would be on the level of third semester students. Slightly fewer would be on the level of second semester students. Since it is possible for examinees to be ahead of pace, there might be some that are operating on the level of fourth semester students; conversely, it is also possible that students might be quite behind, it is even possible that some might be operating at the level of a first semester student or even that of a complete novice. Using an effective theta value of .6 results in an appropriate distribution, which is given in Table 8. Since this variable is not posited to be conditional on any other in the model, it was modeled using a Dirichlet distribution in the manner described by Spiegelhalter et al. (1993). To model a variable in this way, a vector, **e**, is defined with pseudocounts of examinees. For example, with **e** containing the values .1477, .8498, 3.5042, 4.0798, and 1.4185, define *Network Disciplinary Knowledge* to be distributed as a Dirichlet distribution with parameters contained in **e**. In essence, the values in **e** serve as pseudocounts of examinees; the distribution for *Network Disciplinary Knowledge* is one that would be empirically obtained if we observed examinees in the relative frequencies defined in Table 8. Since we desire to have vague prior distributions, we define the pseudocounts accordingly. Operationally, this is accomplished by setting the values in **e** sum to 10. Thus, we have modeled the prior distribution for *Network Disciplinary Knowledge* as if we observed the relative frequencies in Table 8 but on a sample of size 10 (Spiegelhalter et al., 1993).

Summary

In the preceding sections section we have quantitatively specified the variables in the Student Model. In terms of the joint probability distribution in eq. (15), we have specified most of the $P(\boldsymbol{\theta} \mid \boldsymbol{\lambda})$ and hinted at the $P(\boldsymbol{\lambda})$ terms.[6] $P(\boldsymbol{\theta} \mid \boldsymbol{\lambda})$ refers to the distribution of the Student Model variables, while $P(\boldsymbol{\lambda})$ refers to the distribution of the parameters that define the distribution of the Student Model variables. In terms of the effective theta method, $\boldsymbol{\theta}$ are the Student Model variables themselves and $\boldsymbol{\lambda}$ consist of :

- The various c, and d parameters used to define the distributions of *Network Modeling*, *Network Proficiency*, *Design*, *Implement*, and *Troubleshoot*

[6] When we further elaborate on the Evidence Models, we will see that there will be several more variables that might be thought of as being components of the $P(\boldsymbol{\theta} \mid \boldsymbol{\lambda})$ and the $P(\boldsymbol{\lambda})$.

- *e* parameters used to define the distribution of *Network Disciplinary Knowledge*

In order to enact a fully Bayesian model, distributions the various *c* and *d* parameters will need to be specified. This discussion is deferred until after the description of the modeling of the relationships in the Evidence Models.

Evidence Models

Qualitative Description of the Evidence Models

Evidence Models consist of two components, the first being the evaluation component that defines how features of the work product will amount to observables to be used as evidence for inferences about examinees. As the main purpose of this paper is to examine how variables, be they latent or observable, relate to each other the discussion of Evidence Models will presume the evaluation component has been adequately defined and will consist of a discussion of the second component, the statistical model. The statistical model defines how observables should influence belief about the Student Model variables. Thus at one end of the chain there are Student Model variables, on which we would like the assessment to inform us, and at the other end of the chain there are observable variables, the data extracted from the work products. The statistical model defines how the observable data relates to and informs on the Student Model variables. It is the middle link in the chain, serving to take observable data and make it evidence in support of a claim about the Student Model variables.

In NetPASS, there are three distinct types of Evidence Models, each corresponding to a different aspect of *Network Proficiency*: Design, Implement, and Troubleshoot. A pictorial representation of part of a Design Evidence Model is given in Figure 4. The *Network Disciplinary Knowledge* and *Design* variables are those defined in the Student Model; definitions of the others follow. *DK and DesignE* represents the combination of the two Student Model variables involved in this Evidence Model. *DK and DesignE* is not itself of inferential interest; it serves to link the Student Model variables to the observable; such an "instrumental" variable is defined for convenience during modeling. *Correctness of OutcomeE* and *Quality of RationaleE* are the two observable variables in this Evidence Model. The two observables are dependent on *DK and DesignE*, as shown by the directed edges from *DK and DesignE* to each of them. As noted above, conditional independence is a key concept in BINs. Achieving conditional independence is required to achieve the computational simplicity of eq. (15) and more generally for IRT models to apply (Mokken, 1997); the errors associated with modeling variables as conditionally independent when in fact they are not has been documented (Mislevy & Patz, 1995). As they have been modeled so far, the observable variables are not conditionally independent. Their dependence is in part due to their mutual dependence on *DK and DesignE*; however they may be dependent in another way. Both of these variables were formed from the same task: *one* task was presented to an examinee, who in turn responded to this task with a work product, which was then submitted to the evaluation component of the Evidence Model to form the two observables we now see in the model. Since both observables come from the work product to a common task, there may be a dependency between the variables due to the *task*, not due to the parent variable *DK and DesignE*. Such a concern for this type of dependency is analogous to the presence of a method factor in factor analysis. We therefore introduce a context variable, *Design ContextE*, meant to account for this possible (construct irrelevant) dependency. Figure 5 represents a complete Design Evidence Model. Included in this model is *Design ContextE*, which with directed edges to both observables represents another parent to the observables. Note that the distribution for *Design ContextE*, the square to the left of the node in

Figure 5, has no directed edges flowing into it meaning that the distribution of *Design ContextE* is not a conditional distribution; *Design ContextE* is an exogenous variable. The two parents, *DK and DesignE* and *Design ContextE*, represent distinct and independent portions of the dependency between *Correctness of OutcomeE* and *Quality of RationaleE*. Alone, neither can account for the dependency between the observables; the observables are conditionally dependent given one (either) parent, and are conditionally independent given both parents. Figure 5 represents a complete Design Evidence Model such that the observables are both (1) modeled in relation to Student Model variables, and (2) conditionally independent given their parents.

Part of an Implement Evidence Model is depicted in Figure 6. The definitions of these variables are analogous to their counterparts defined above for the Design Evidence Model. In addition to the data used to form these three observables, the work products examinees produce in response to the task contain information regarding other Student Model variables. More specifically, the work products examinees produce in response to this task lead to another observable dependent on *Network Disciplinary Knowledge* and *Network Modeling*. This portion of the Implement Evidence Model is depicted in Figure 7. *Network Disciplinary Knowledge* and *Network Modeling* combine to yield *DK and Network ModelingE*, which is the parent of an observable*, Correctness of Outcome 2E. DK and Network ModelingE* is structured in exactly the same way as *DK and ImplementE*, except *Network Modeling* joins *Network Disciplinary Knowledge* as a parent, replacing *Implement*.

We combine this portion of the Evidence Model with the one in Figure 6. The complete Evidence Model is depicted in Figure 8. Note that all the observables have (the same) *Implement ContextE* as one parent. Again, this is because all the observables are formed from the same work product from *one* task, and therefore might have dependencies among them above and beyond that which can be attributable to either *DK and ImplementE* or *DK and Network ModelingE*. A Troubleshoot Evidence Model is depicted in Figure 9. Its interpretation is analogous to the Implement Evidence Model in Figure 8.

We have so far mentioned the different *types* of Evidence Models: Design, Implement, and Troubleshoot. There are three different *instantiations* of each type, corresponding to the expected difficulty of the task presented to the examinee. For instance there are Design Easy, Design Medium, and Design Hard instantiations, which use observables extracted from Design Easy, Design Medium, and Design Hard tasks, respectively. Naturally, it is a bit premature to refer to a task as easier or more difficult than any other. After all, the goal is to calibrate the model and gain information on the difficulties of the tasks. The terms "Easy," "Medium," and "Hard" capture expert expectation, as the tasks were constructed to be of different difficulties.

For each instantiation of each type of Evidence Model there will be the appropriate "instrumental" variable (i.e., the combination of *Network Disciplinary Knowledge* and another Student Model) and the appropriate context variable, each localized to the particular instance of the particular Evidence Model.[7]

Quantitative Modeling of Specific Relationships in the Evidence Models

Conjunctive Relationships

Conjunctive relationships are those in which multiple skills are required for performance. In terms of BINs, this amounts to modeling the relationship as such: for a child to reach certain values, all of its parents must have (at least) that value. Mathematically, this is a minimum

[7] The names of all of the "instrumental" variables, context variables, and observables in Figures 5, 8, and 9 ended with 'E', indicating that these instantiations were the Design Easy, Implement Easy, and Troubleshoot Easy instantiations, respectively.

function; the minimum value of the parents sets the value for the child. When using a formal conjunction (i.e., minimum) function to define the effective theta, using the GRM will yield a probability distribution for all the possible values. These values do not represent the probability distribution of the child, for, as in the ceiling relationships, in using the GRM the structure of the conjunction is lost; the GRM allows for the child to take on values higher than the minimum of the parents. The conjunctive structure i.e., the ceiling value, is thus subsequently imposed the probability distribution is adjusted accordingly.

Basic formulas.

Let θ_1 and θ_2 be parent variables for a child variable θ_c; furthermore, let θ_1, θ_2, and θ_c take on any of five possible states. Define

$$\theta^{*}_{\theta_c} \equiv \min(\theta_1, \theta_2), \tag{40}$$

Define a linear transformation of $\theta^{*}_{\theta_c}$:

$$\theta^{**}_{\theta_c} \equiv u_{\theta_c}\left(\theta^{*}_{\theta_c}\right) = c_{\theta_c} \times \theta^{*}_{\theta_c} + d_{\theta_c}. \tag{41}$$

Entering this value into the GRM would lead to a probability distribution for the possible values of θ_c which would then be adjusted so that the value of θ_c could not exceed the ceiling, defined in eq. (40). This would be a model of a "leaky" conjunction.[8] However, it may be the case in a leaky conjunction that the expected value of the child is not merely a function of the minimum value of the parents, but may also depend on *which* parent sets the minimum and what the value of *the other parent* is. Thus, a more complete definition of the effective theta would be:

$$\theta^{**}_{\theta_c} \equiv [c_{\theta_c} \times \theta^{*}_{\theta_c} + d_{\theta_c}] + [c_{\theta_1} \times (\theta_1 - \theta^{*}_{\theta_c})] + [c_{\theta_2} \times (\theta_2 - \theta^{*}_{\theta_c})] \tag{42}$$

[8] The term "leaky" is used to indicate that though the value of the child has a ceiling at the minimum of its parents, probabilities "leak" below the ceiling, meaning that it is possible for the child to take on a value below the ceiling.

where the contents of the first set of brackets is just that defined in eq. (41), the contents of the second set of brackets captures the impact of how high above the minimum θ_1 is, and the contents of the third set of brackets captures the impact of how high above the minimum θ_2 is.[9]

Once the effective theta is obtained, it is entered into the GRM to obtain a probability distribution for the value of the child. The GRM will return probabilities for all possible values, even those outlawed by the leaky conjunction, i.e., those above $\theta^*_{\theta_c}$. To fix this, we will force the probabilities for the values above $\theta^*_{\theta_c}$ to be 0 and renormalize the others. Let us illustrate this by turning to NetPASS.

Examples from NetPASS.

Consider again the Design Easy Evidence Model, depicted in Figure 5. *DK and DesignE* is formed by a leaky conjunction of *Network Disciplinary Knowledge* and *Design*. Thus to calculate the effective theta first instantiate equation (40):

$$\theta^*_{DKandDesign} \equiv \min(\theta_{NDK}, \theta_{Design}) \tag{43}$$

where θ_{NDK} is the value of *Network Disciplinary Knowledge*, and θ_{Design} is the value of *Design*. Next instantiate eq. (42) to calculate the effective theta:

$$\begin{aligned}\theta^{**}_{DKandDesignE} &\equiv [c_{DKandDesignE} \times \theta^*_{DKandDesign} + d_{DKandDesignE}] \\ &\quad + [c_{NDKE} \times (\theta_{NDK} - \theta^*_{DKandDesign})] + [c_{DesignE} \times (\theta_{Design} - \theta^*_{DKandDesign})]\end{aligned} \tag{44}$$

These effective thetas are entered into the GRM to produce probabilities for the child, *DK and DesignE*. Again, using the GRM as such will result in possible values for the child above the

[9] Let us suppose that $\theta_1 < \theta_2$. In that case, $\theta^*_{\theta_c}$ would be θ_1 and the value in the second set of brackets would be 0. However, the third set of brackets would contribute to the value of $\theta^{**}_{\theta_c}$. If $\theta_2 < \theta_1$ the situation would be reversed. In the case where the values of the parents are equal (and hence, both parents equal the minimum) the contribution of both brackets would be 0.

minimum of the parents. These probabilities must be set to zero and the rest of the probabilities in each case (i.e., each row in the table) must be renormalized. Table 9 illustrates the correct structure of the probabilities.

The values listed in Table 9 were calculated using eq. (44) with $c_{DKandDesignE} = 2$, $d_{DKandDesignE} = -6.0$, $c_{NDKE} = .2$, and $c_{DesignE} = .4$ to reflect the opinions and expectations of SMEs. SMEs hypothesized that the impact of *Design* was greater than that of *Network Disciplinary Knowledge*. This is modeled by having the value of $c_{DesignE}$ be greater than c_{NDKE}.[10] As with the parameters in the Student Model, no mathematical constraints have been placed on the values; SME expectations serve as the basis for our prior distributions for the parameter to be refined by the information in the data.

The *DK and DesignE* variable in the Design Easy instance is not of inferential interest; it serves the purpose of capturing the structure of the relationship between the Student Model variables and the observables in the Evidence Model. This "instrumental" variable is modeled in the Design Medium and Design Hard instances in exactly the same way. That is,

$$\begin{aligned}\theta^{**}_{DKandDesignM} \equiv {} & [c_{DKandDesignM} \times \theta^{*}_{DKandDesign} + d_{DKandDesignM}] \\ & + [c_{NDKM} \times (\theta_{NDK} - \theta^{*}_{DKandDesign})] + [c_{DesignM} \times (\theta_{Design} - \theta^{*}_{DKandDesign})]\end{aligned} \tag{45}$$

and

$$\begin{aligned}\theta^{**}_{DKandDesignH} \equiv {} & [c_{DKandDesignH} \times \theta^{*}_{DKandDesign} + d_{DKandDesignH}] \\ & + [c_{NDKH} \times (\theta_{NDK} - \theta^{*}_{DKandDesign})] + [c_{DesignH} \times (\theta_{Design} - \theta^{*}_{DKandDesign})]\end{aligned} \tag{46}$$

[10] This can be illustrated in much the same way as was the expected difference between *Design*, *Implement*, and *Troubleshoot*.

are the effective thetas for *DK and DesignM* and *DK and DesignH*, respectively.[11] Naturally, SME expectations for the parameters in these equations match those defined in the effective theta equation for *DK and DesignE*; the expected conditional probabilities for *DK and DesignM* and *DK and DesignH* are therefore just those given in Table 9.

Turning to the Implement Evidence Models, the specification of *DK and ImplementE* and *DK and Network ModelingE* in the Implement Easy instance, *DK and ImplementM* and *DK and Network ModelingM* in the Implement Medium instance, and *DK and ImplementH* and *DK and Network ModelingH* in the Implement Hard instance mirrors that of their counterparts in the Design Evidence Models, save for which variables are the parents. That is, to obtain the effective thetas first instantiate eq. (40):

$$\theta^{*}_{DKandImplement} \equiv \min\left(\theta_{NDK}, \theta_{Implement}\right) \tag{47}$$

$$\theta^{*}_{DKandNM} \equiv \min\left(\theta_{NDK}, \theta_{NM}\right) \tag{48}$$

where $\theta_{Implement}$ is the value of *Implement* and θ_{NM} is the value of *Network Modeling*. The effective thetas for the Implement Easy instance are defined as:

$$\begin{aligned}\theta^{**}_{DKandImplementE} \equiv {} & [c_{DKandImplementE} \times \theta^{*}_{DKandImplement} + d_{DKandImplementE}] \\ & + [c_{NDKE} \times (\theta_{NDK} - \theta^{*}_{DKandImplement})] + [c_{ImplementE} \times (\theta_{Implement} - \theta^{*}_{DKandDesign})]\end{aligned} \tag{49}$$

and

$$\begin{aligned}\theta^{**}_{DKandNME} \equiv {} & [c_{DKandNME} \times \theta^{*}_{DKandNM} + d_{DKandNME}] \\ & + [c_{NDKE} \times (\theta_{DK} - \theta^{*}_{DKandNM})] + [c_{NME} \times (\theta_{NM} - \theta^{*}_{DKandDesig})]\end{aligned} \tag{50}$$

The effective thetas for the Implement Medium instance are defined as:

[11] Note that we need not compute counterparts of eq. (43) for the Design Medium and Design Hard instances; as the minimum of the Student Model variables, θ_{DK} and θ_{Design}, does not change from instance to instance.

$$\begin{aligned}\theta^{**}_{DKandImplementM} \equiv {} & [c_{DKandImplementM} \times \theta^{*}_{DKandImplement} + d_{DKandImplementM}] \\ & + [c_{NDKM} \times (\theta_{NDK} - \theta^{*}_{DKandImplement})] + [c_{ImplementM} \times (\theta_{Implement} - \theta^{*}_{DKandDesign})]\end{aligned} \tag{51}$$

and

$$\begin{aligned}\theta^{**}_{DKandNMM} \equiv {} & [c_{DKandNMM} \times \theta^{*}_{DKandNM} + d_{DKandNMM}] \\ & + [c_{NDKM} \times (\theta_{NDK} - \theta^{*}_{DKandNM})] + [c_{NMM} \times (\theta_{NM} - \theta^{*}_{DKandDesign})]\end{aligned} \tag{52}$$

The effective thetas for the Implement Hard instance are defined as:

$$\begin{aligned}\theta^{**}_{DKandImplementH} \equiv {} & [c_{DKandImplementH} \times \theta^{*}_{DKandImplement} + d_{DKandImplementH}] \\ & + [c_{NDKH} \times (\theta_{NDK} - \theta^{*}_{DKandImplement})] + [c_{ImplementH} \times (\theta_{Implement} - \theta^{*}_{DKandDesign})]\end{aligned} \tag{53}$$

and

$$\begin{aligned}\theta^{**}_{DKandNMH} \equiv {} & [c_{DKandNMH} \times \theta^{*}_{DKandNM} + d_{DKandNMH}] \\ & + [c_{NDKH} \times (\theta_{NDK} - \theta^{*}_{DKandNM})] + [c_{NMH} \times (\theta_{NM} - \theta^{*}_{DKandDesign})]\end{aligned} \tag{54}$$

As in the Design Evidence Models, these effective thetas must be entered into the GRM, impossible states must be zeroed out and the remaining probabilities must be renormalized. Discussions with SMEs indicated that the values of the parameters that define the effective thetas in the above equations are expected to be the same as their counterparts in the Design Evidence Model instances; the conditional probabilities based on this expectation are therefore those given in Table 9.

Regarding the Troubleshoot Evidence Models, the specification of the instrumental variables in the Easy, Medium, and Hard instances mirrors that of their counterparts described above. That is, to obtain the effective thetas first instantiate eq. (40):

$$\theta^{*}_{DKandITroubleshoot} \equiv \min(\theta_{NDK}, \theta_{Troubleshoot}) \tag{55}$$

$$\theta^{*}_{DKandNM2} \equiv \min(\theta_{NDK}, \theta_{NM}) \tag{56}$$

where in addition to variables previously defined, $\theta_{Troubleshoot}$ is the value of *Troubleshoot.* The effective thetas for the Troubleshoot Easy instance are defined as:

$$\begin{aligned}\theta^{**}_{DKandTroubleshootE} \equiv & [c_{DKandTroubleshootE} \times \theta^{*}_{DKandTroubleshoot} + d_{DKandTroubleshootE}] \\ & + [c_{NDKE} \times (\theta_{NDK} - \theta^{*}_{DKandTroubleshoot})] \\ & + [c_{TroubleshootE} \times (\theta_{Troubleshoot} - \theta^{*}_{DKandDesign})]\end{aligned} \tag{57}$$

and

$$\begin{aligned}\theta^{**}_{DKandNM2E} \equiv & [c_{DKandNM2E} \times \theta^{*}_{DKandNM} + d_{DKandNM2E}] \\ & + [c_{NDK2E} \times (\theta_{NDK} - \theta^{*}_{DKandNM})] + [c_{NM2E} \times (\theta_{NM} - \theta^{*}_{DKandDesign})]\end{aligned} \tag{58}$$

The effective thetas for the Troubleshoot Medium instance are defined as:

$$\begin{aligned}\theta^{**}_{DKandTroubleshootM} \equiv & [c_{DKandTroubleshootM} \times \theta^{*}_{DKandTroubleshoot} + d_{DKandTroubleshootM}] \\ & + [c_{NDKM} \times (\theta_{NDK} - \theta^{*}_{DKandTroubleshoot})] \\ & + [c_{TroubleshootM} \times (\theta_{Troubleshoot} - \theta^{*}_{DKandDesign})]\end{aligned} \tag{59}$$

and

$$\begin{aligned}\theta^{**}_{DKandNM2M} \equiv & [c_{DKandNM2M} \times \theta^{*}_{DKandNM} + d_{DKandNM2M}] \\ & + [c_{NDK2M} \times (\theta_{NDK} - \theta^{*}_{DKandNM})] + [c_{NM2M} \times (\theta_{NM} - \theta^{*}_{DKandDesign})]\end{aligned} \tag{60}$$

The effective thetas for the Implement Hard instance are defined as:

$$\begin{aligned}\theta^{**}_{DKandTroubleshootH} \equiv & [c_{DKandTroubleshootH} \times \theta^{*}_{DKandTroubleshoot} + d_{DKandTroubleshootH}] \\ & + [c_{NDKH} \times (\theta_{NDK} - \theta^{*}_{DKandTroubleshoot})] \\ & + [c_{TroubleshootH} \times (\theta_{Troubleshoot} - \theta^{*}_{DKandDesign})]\end{aligned} \tag{61}$$

and

$$\begin{aligned}\theta^{**}_{DKandNM2H} \equiv & [c_{DKandNM2H} \times \theta^{*}_{DKandNM} + d_{DKandNM2H}] \\ & + [c_{NDK2H} \times (\theta_{NDK} - \theta^{*}_{DKandNM})] + [c_{NM2H} \times (\theta_{NM} - \theta^{*}_{DKandDesign})]\end{aligned} \tag{62}$$

As in the Design and Implement Evidence Models, these effective thetas must be entered into the GRM, impossible states must be zeroed out and the remaining probabilities must be renormalized. Discussions with SMEs indicated that the values of the parameters that define the effective thetas in the above equations are expected to be the same as their counterparts in the Design and Implement Evidence Model instances; the conditional probabilities based on this expectation are therefore those given in Table 9.

Compensatory Relationships

A common method for modeling compensatory relationships is weighted sums or averages, as in multiple factor analysis (Thurstone, 1947). When modeling a compensatory relationship, one's first inclination may be to simply sum up the linear mappings for each parent variable to the child. More formally, if the marginal contribution of parent variable θ_l is the linear mapping function

$$\theta^{*}_{tl} \equiv g_{tl}(\theta_l) = c_{tl} \times (\theta_l) + d_{tl} \qquad (63)$$

then the combination all L linear mapping functions would be

$$\theta^{**}_{t} \equiv h_t(\theta^{*}_{t1}, \ldots, \theta^{*}_{tL}) = \sum_{l=1}^{L} \theta^{*}_{tl} \qquad (64)$$

The particular advantage of this strategy is that the relevance of each of the requisite skills can be assessed (Mislevy, Senturk, et al., in press). This feature, which is advantageous when information regarding each of the separate skills is available from either experts and/or features of the tasks, is also problematic in that, given response data alone, the model is usually underdetermined, due to the sum of the intercepts (Mislevy, Senturk, et al., in press). However, in the case of NetPASS, all of the compensatory relationships in NetPASS involve a context variable, the impact of which can be modeled without encountering problems of underdetermination.

Basic formulas.

Let θ_1 be a parent variable for T observables $X_1,\ldots,X_T$;[12] furthermore, let θ_1 be one of the instrumental variables defined above and take on any of five states. Let θ_2 be a context variable that will also serve as a parent variable for the T observables $X_1,\ldots,X_T$; let this context variable take on any of two states, corresponding to values of High and Low. Following the discussion of the previous section, the marginal contribution of θ_1 is given as

$$\theta_{t1}^{*} \equiv g_{t1}(\theta_1) = c_{t1} \times (\theta_1) + d_{t1} \qquad (65)$$

and the marginal contribution of θ_2 is given as

$$\theta_{t2}^{*} \equiv g_{t2}(\theta_2) = c_{t2} \times (\theta_2) + d_{t2} = c_{t2} \times (\theta_2). \qquad (66)$$

Note that d_{t2} has been dropped on the right side of eq. (66). This occurs because if the two-level context variable is centered around 0, e.g., with Low coded as –1 and High coded as +1, c_{t2} captures all the information and d_{t2} is unnecessary. To specify the expression for the effective theta, instantiate eq. (64):

$$\theta_t^{**} \equiv f_t\left(\theta_{tl}^{*}\right) = c_{t1} \times (\theta_1) + c_{t2} \times (\theta_2) + d_{t1}. \qquad (67)$$

We can think of the compensatory relationship that involves a context variable as simply the sum of the marginal values θ_{t1}^{*} and θ_{t2}^{*}, the impact of θ_1 followed by the additional impact of the context variable, θ_2. For a slightly different approach to developing a compensatory relationship, from the perspective of moving from a conditionally dependent model to a conditionally independent model, see Mislevy, Senturk, et al. (in press).

[12] As compensatory relationships only appear in NetPASS in the modeling of observables, we refer to the child variables as observables; naturally, there is nothing about compensatory relationships that requires the child variables be observable.

Examples from NetPASS.

Each instance of a Design Evidence Model contains two observables obtained from work products produced in response to a common task. The *DK and Design* variable in each instance can take on any of five values corresponding to novice, semester 1, semester 2, semester 3, and semester 4, coded as 1-5. The *Design Context* variable in each instance can take on either of two values, Low or High, which are coded as –1 and +1, respectively.[13] To obtain the effective theta for the observables in the Design Easy instance, instantiate eq. (67)

$$\theta_t^{**} = c_{tDKandDesignE} \times (\theta_{DKandDesignE}) + c_{tDesignContextE} \times (\theta_{DesignContextE}) + d_{tDKandDesignE} \qquad (68)$$

Table 10 is a table of initial conditional probability distributions for the observables in the Design Easy Evidence Model. These were calculated by evaluating eq. (68) with $c_{tDKandDesignE}$ = 2, $d_{tDKandDesignE}$ = -5.0, $c_{tDesignContextE}$ = .4, and reflect the opinions and expectations of the SMEs; these values serve to define the prior distributions for the calibration of the model.

The compensatory relationship appears repeatedly in the NetPASS model. We have so far mentioned the Design Easy instance. The Design Medium and Design Hard instances will, of course, have the same structure, though we have the ability to quantitatively define the expected difference in difficulty by a change in the intercept parameter. Define the effective theta for the observables in the Design Medium instance and the observables in the Design Hard instance to be

$$\theta_t^{**} = c_{tDKandDesignM} \times (\theta_{DKandDesignM}) + c_{tDesignContextM} \times (\theta_{DesignContextM}) + d_{tDKandDesignM} \qquad (69)$$

and

[13] Though they are being specified as part of the Evidence Models, the instrumental variables representing the combination of two Student Model variables and the Context variables are all indexed by examinees (and appear as parent variables in the calculation of the effective thetas for observables). As such they may be thought of as Student Model variables (i.e., latent variables modeled as being part of examinees), though the procedure adopted here is equivalent.

$$\theta_t^{**} = c_{tDKandDesignH} \times (\theta_{DKandDesignH}) + c_{tDesignContextH} \times (\theta_{DesignContextH}) + d_{tDKandDesignH}, \qquad (70)$$

respectively. The expected difference in difficulty between the scenarios is captured in the expectation in the intercept terms: for the Design Easy instance, $d_{tDKandDesignE}$ = -5.0; for the Design Medium instance, $d_{tDKandDesignM}$ = -6.0; for the Design Hard instance, $d_{tDKandDesignH}$ = -7.0.[14] The expected strength of association between the observables, and (both of) the parent variables remains unchanged, i.e., the coefficients in the Design Medium and Design Hard scenarios are expected to be equal to their counterparts in the Design Easy scenario. Tables 11 and 12 give the conditional probabilities of response for the Design Medium and Design Hard instances, respectively. Again, the values used to calculate the expert expectations will serve as the basis for the priors in estimating the parameters in the model.

Consider now the Implement Evidence Model given in Figure 8. Like the Design Evidence Model, there are three instantiations of the Implement Evidence Model: Easy, Medium, and Hard. With more observables and more parent variables, the Implement Evidence Models are slightly different than the Design Evidence Models. Fundamentally however, they are the same; for each observable there are two parents: one is the combination of two Student Model variables (that can take on any of five values) and the other is a context variable (that can take on either of two values) designed to account for the common origin of the observables and induce conditional independence. Calculating the conditional probabilities for an Implement Evidence Model consists of simply repeating the procedure for setting up a Design Evidence Model twice; we calculate two effective thetas instead of one. Furthermore, the anticipated values for the coefficients and intercepts in the calculation of both effective thetas in the various instances of the Implement Evidence Model are hypothesized to be equal to those in the corresponding

[14] See note 5

instances of the Design Evidence Model. For the Implement Easy instance, we define the effective thetas as

$$\begin{aligned}\theta_{t1}^{**} &= c_{tDKandImplementE} \times (\theta_{DKandImplementE}) \\ &\quad + c_{tImplementContextE} \times (\theta_{ImplementContextE}) + d_{tDKandImplementE}\end{aligned} \tag{71}$$

and

$$\theta_{t2}^{**} = c_{tDKandNME} \times (\theta_{DKandNME}) + c_{tImplementContextE} \times (\theta_{ImplementContextE}) + d_{tDKandNME} \tag{72}$$

where the coefficients and the intercepts in the expressions above are expected to take on the same values as those listed for the observables in the Design Easy instance above. The expected conditional probabilities for the observables in the Implement Easy instance are just those given in Table 10.

To calculate the expected conditional probabilities for the observables in the Implement Medium instance and the Implement Hard instance the procedure just described is repeated. The effective thetas for the Implement Medium and Implement Hard instances are:

$$\begin{aligned}\theta_{t1}^{**} &= c_{tDKandImplementM} \times (\theta_{DKandImplementM}) \\ &\quad + c_{tImplementContextM} \times (\theta_{ImplementContextM}) + d_{tDKandImplementM}\end{aligned} \tag{73}$$

$$\theta_{t2}^{**} = c_{tDKandNMM} \times (\theta_{DKandNMM}) + c_{tImplementContextM} \times (\theta_{ImplementContextM}) + d_{tDKandNMM} \tag{74}$$

and

$$\begin{aligned}\theta_{t1}^{**} &= c_{tDKandImplementH} \times (\theta_{DKandImplementH}) \\ &\quad + c_{tImplementContextH} \times (\theta_{ImplementContextH}) + d_{tDKandImplementH}\end{aligned} \tag{75}$$

$$\theta_{t2}^{**} = c_{tDKandNMH} \times (\theta_{DKandNMH}) + c_{tImplementContextH} \times (\theta_{ImplementContextH}) + d_{tDKandNMH} \tag{76}$$

where the coefficients and the intercepts in the expressions above are expected to take on the same values as those listed for the observables in the Design Medium instance and the Design

Hard instance. The distributions corresponding to SME expectation for the Implement Medium and Implement Hard instances are therefore those given in Tables 11 and 12, respectively.

With these procedures, the quantification of the instances of the Troubleshoot Evidence Model is straightforward. As with the Implement Evidence Model instances, we calculate two effective thetas instead of one. And again, the expert expectations for the values for the coefficients and intercepts in the calculation of both effective thetas in the instances of the Troubleshoot Evidence Model are hypothesized to be equal to those in the Design and Implement Evidence Models. For the Troubleshoot Easy instance, we define the effective thetas as

$$\begin{aligned}\theta_{t1}^{**} &= c_{tDKandTroubleshootE} \times (\theta_{DKandTroubleshootE}) \\ &\quad + c_{tTroubleshootContextE} \times (\theta_{TroubleshootContextE}) + d_{tDKandTroubleshootE}\end{aligned} \tag{77}$$

and

$$\begin{aligned}\theta_{t2}^{**} &= c_{tDKandNM2E} \times (\theta_{DKandNM2E}) \\ &\quad + c_{tTroubleshootContextE} \times (\theta_{TroubleshootContextE}) + d_{tDKandNetworkModeling2E}\end{aligned} \tag{78}$$

The effective thetas for the Troubleshoot Medium and Troubleshoot Hard instances are

$$\begin{aligned}\theta_{t1}^{**} &= c_{tDKandTroubleshootM} \times (\theta_{DKandTroubleshootM}) \\ &\quad + c_{tTroubleshootContextM} \times (\theta_{TroubleshootContextM}) + d_{tDKandTroubleshootM}\end{aligned} \tag{79}$$

$$\begin{aligned}\theta_{t2}^{**} &= c_{tDKandNM2M} \times (\theta_{DKandNM2M}) \\ &\quad + c_{tTroubleshootContextM} \times (\theta_{TroubleshootContextM}) + d_{tDKandNetworkModeling2M}\end{aligned} \tag{80}$$

and

$$\begin{aligned}\theta_{t1}^{**} &= c_{tDKandTroubleshootH} \times (\theta_{DKandTroubleshootH}) \\ &\quad + c_{tTroubleshootContextH} \times (\theta_{TroubleshootContextH}) + d_{tDKandTroubleshootH}\end{aligned} \tag{81}$$

$$\begin{aligned}\theta_{t2}^{**} &= c_{tDKandNM2H} \times (\theta_{DKandNM2H}) \\ &\quad + c_{tTroubleshootContextH} \times (\theta_{TroubleshootContextH}) + d_{tDKandNetworkModeling2H}\end{aligned} \tag{82}$$

As before, the expected condition distributions for the Easy, Medium, and Hard instances are those given in Tables 10, 11, and 12, respectively.

Exogenous Variable

In the Evidence Models, only the Context variables are exogenous. They are modeled as taking on values of -1 and $+1$, each with probability .5. Modeling the values they can take on as symmetric around zero allows for their incorporation in the effective theta for observables without an intercept term (eq. 66)).

Summary

In the preceding sections section the variables in the three instances of the three Evidence Models have been quantitatively specified. In terms of the joint probability distribution in eq. (15), we have specified $P(\mathbf{X} \mid \boldsymbol{\theta}, \boldsymbol{\pi})$ and hinted at the $P(\boldsymbol{\eta})$ terms. $P(\mathbf{X} \mid \boldsymbol{\theta}, \boldsymbol{\pi})$ refers to the distribution of the observable variables conditional on the Student Model variables, $\boldsymbol{\theta}$, and the conditional probabilities, $\boldsymbol{\pi}$. In terms of the effective theta method, $\mathbf{X}$ are the observable variables, $\boldsymbol{\pi}$ are the conditional probabilities themselves, and $\boldsymbol{\eta}$ consist of the various *c* and *d* parameters used to define the conditional distributions. Note that we need not *specify* the conditional probabilities given the parameters that govern them (i.e., the $P(\mathbf{X} \mid \boldsymbol{\theta}, \boldsymbol{\pi})$ terms), because the conditional probabilities are a *function* of the *c* and *d* parameters. In utilizing the GRM, we define the conditional probabilities as a mathematical function of the *c* and *d* parameters. Given the *c* and *d* parameters, we *calculate* the conditional probabilities. In other words, given the *c* and *d* parameters, the conditional probabilities are known with certainty.

Specification of the Priors

So far, all the terms in eq. (15) have been fully specified except $P(\lambda)$ and $P(\eta)$. $P(\lambda)$ refers to the distribution of the parameters that define the distributions of examinee proficiencies, the various *c*, and *d*, and *e* parameters in the Student Model. In detailing the expectations of SMEs, we have already described some aspect of the distribution, namely, the value that corresponds to modeling particular expectations. To enable Bayesian estimation, parameters must not be fixed, but modeled as random variables. Leaning on intuition and past experience in IRT, we define the priors for all intercepts to be distributed normally with mean defined by expert expectation and variance of 1. Similarly, we define the priors for all coefficients to be distributed normally with mean defined by expert expectation and variance of 1, truncated at 0 to force all the coefficients to be positive. $P(\eta)$ refers to the distribution of the parameters that define the conditional probability distributions, the various *c* and *d* parameters in the calculation of the effective thetas in the Evidence Models. As before, we define the priors for all intercepts to be normally distributed with mean defined by expert expectation and variance of 1 and the priors for all coefficients to be normally distributed with mean defined by expert expectation and variance of 1, truncated at 0.

Markov Chain Monte Carlo (MCMC) Estimation

The Full Bayesian Model

We have devoted some time to setting up the Bayesian model for the NetPASS assessment. To do so, we have qualitatively defined relationships among the various variables in the NetPASS model to determine the structure of the probability distributions and then quantitatively specified the relationships, filling in the contents of the probability distributions.

Indeed, all terms on the right side of eq. (15) have been specified. Of course, all of the conditional probability distributions were based on the opinions of SMEs. If we were certain the conditional probability distributions were correct, we could proceed by administering the NetPASS assessment to examinees and input their values for the observables and draw inferences about their values on Student Model variables. However, while we expect the views of the SMEs to be correct (at least more correct than those of anyone else), we seek to augment the information gathered from discussions with experts with actual data. That is, the model as we have so far specified it represents our *prior* beliefs about the relationships of several variables and the characteristics of the tasks presented to examinees; we will collect data to *update* our beliefs regarding the relationships and the task characteristics. As with all Bayesian models, our updated beliefs will come in the form of *posterior* distributions.

With a model as complex as the NetPASS model straightforward application of Bayes' Theorem is computationally intractable. What's more, our current aim is refine our beliefs about the parameters that govern the relationships among variables. We are therefore interested in the posterior distributions for these parameters, which will represent the incorporation of information from the data to our prior beliefs based on expert opinion. We seek to condition on observed data and refine our beliefs about the parameters, which for all unobserved parameters will be (following Bayes' Theorem) proportional to the prior for that parameter multiplied by the conditional probability of the observed variables given the unobserved parameters. Expressed mathematically we aim to arrive at:

$$P(\boldsymbol{\theta},\boldsymbol{\pi},\boldsymbol{\eta},\boldsymbol{\lambda} \mid \mathbf{X}) \propto P(\mathbf{X} \mid \boldsymbol{\theta},\boldsymbol{\pi}) \times P(\boldsymbol{\theta} \mid \boldsymbol{\lambda}) \times P(\boldsymbol{\lambda}) \times P(\boldsymbol{\pi} \mid \boldsymbol{\eta}) \times P(\boldsymbol{\eta}) \qquad (83)$$

Here, $P(\boldsymbol{\theta},\boldsymbol{\pi},\boldsymbol{\eta},\boldsymbol{\lambda} \mid \mathbf{X})$ is the posterior distribution of all the unobservable parameters: examinee parameters ($\boldsymbol{\theta}$, the Student Model variables), examinee hyperparameters ($\boldsymbol{\lambda}$, those parameters which define the distributions of the Student Model variables), the conditional probabilities (π), and the task parameters (η, which define the conditional probabilities).[15]

An analytic solution for the posteriors for this model is computationally intractable and may very well be impossible. Instead, we pursue an empirical approximation via Markov chain Monte Carlo (MCMC) estimation. MCMC estimation provides an adequate and appropriate framework for computation in Bayesian analyses (Gelman et al., 1995). A complete treatment and description of MCMC estimation is beyond the scope and intent of this work; suffice it to say that for our current purposes, MCMC estimation consists of drawing from a series of distributions that is in the limit equal to drawing from the true posterior distribution (Gilks et al., 1996). That is, to empirically sample from the posterior distribution, it is sufficient to construct a Markov chain that has the posterior distribution as its stationary distribution. One general method for constructing such a chain is presented in the next section. For a more complete discussion of MCMC techniques, see Brooks, (1998) and Gilks et al. (1996).

Metropolis-Hastings Sampling

One method for constructing a Markov chain with the distribution of interest as the stationary distribution is due to a generalization (Hastings, 1970) of a method originally presented by Metropolis et al., (1953). In Metropolis-Hastings sampling (Hastings, 1970), a proposal distribution is selected to aid in the sampling scheme. Given the current value of the chain, a value is drawn from the proposal distribution; this value is then considered as a candidate for the next value in the chain.

[15] Note the similarity between eq. (82), the posterior distribution, and eq. (15), the joint distribution.

Let $\pi(\circ)$ be the target distribution of interest[16] and $q(\circ)$ be the proposal distribution. It can be shown that, under regularity conditions (see Roberts, 1996 and Tierney, 1996), the distribution of values of iterations of a Markov chain asymptotically converges to its stationary distribution if the value drawn from the proposal distribution, y, is accepted with probability

$$\alpha(x^z, y) = \min \left\{1, \left(\frac{\pi(y)\, q(x^z \mid y)}{\pi(x^z)\, q(y \mid x^z)}\right)\right\} \tag{84}$$

where x^z is the current value in the chain (see Chib and Greenberg, 1995 for a non-technical derivation). Interestingly enough, the procedure requires no particular parametric form of the proposal distribution. Note that the proposal distribution may be written as a conditional distribution (as it is in eq. (84)). Thus, $q(y \mid x^z)$ is the probability of selecting y from the proposal distribution, given the current value is x^z. Similarly, $q(x^z \mid y)$ is the probability of selecting x^z from the distribution, given the value is y.

Metropolis Sampling

In this section, the Metropolis sampler (Metropolis et al., 1953) is presented as a special case of the Metropolis-Hastings sampler. Observe that if we select $q(\circ)$ that is symmetric with respect to its arguments, i.e., $q(y \mid x^z) = q(x^z \mid y)$, the proposal distribution drops out of the equation for α, the probability of accepting y as the z+1th value in the chain. A convenient choice for a distribution symmetric about its arguments is the normal distribution with mean

[16] Note that $\pi(\circ)$ refers to the distribution of interest, i.e., the joint posterior distribution, not the conditional probabilities in the model, $\boldsymbol{\pi}$.

defined by the current value.[17] Choosing $q(\circ)$ to be a distribution that is symmetric reduces the computation in eq. (84); the acceptance probability then becomes

$$\alpha(x^z, y) = \min \left\{1, \left(\frac{\pi(y)}{\pi(x^z)}\right)\right\} \tag{85}$$

Note also that $\pi(\circ)$ appears in both the numerator and denominator, which means that we are only required to know $\pi(\circ)$ up to a constant of proportionality. This is the key feature of MCMC techniques that allows for estimation of complex Bayesian models, where the calculation of a posterior distribution (eq. (2)) is often intractable, but the posterior can be defined up to a constant of proportionality. In the case of NetPASS, the posterior distribution (eq. (83)) is specified only up to a constant of proportionality.

The Algorithm

Let ϑ be the set of parameters that define the joint posterior distribution, i.e., the parameters that define the model. Initialize the variables in the vector as ϑ^0. For the *z*+1th iteration:

Draw ϑ^y from the proposal distribution, $q(\vartheta \mid \vartheta^z)$;

Accept ϑ^y as the value for the *z*+1th iteration with probability:

$$\alpha(\vartheta^z, \vartheta^y) = \min \left\{1, \left(\frac{\pi(\vartheta^y)}{\pi(\vartheta^z)}\right)\right\};$$

Otherwise, retain ϑ^z as the value in the z+1th iteration.

[17] In this case, $q(y \mid x^z)$ would be a normal distribution with a mean of x; $q(x^z \mid y)$ would be a normal distribution with a mean of y.

Asymptotically, the distribution of draws converges to the stationary distribution, $\pi(\circ)$, the true posterior distribution. Thereafter, draws from the distribution represent sampling from the posterior. The empirical distribution of many draws from this distribution for any parameter approximates its marginal distribution.

MCMC theory states that, asymptotically, the chain will converge to its stationary distribution. Empirical Monte Carlo distributions are representative of the posterior only after the chain has converged to its stationary distribution. Values from iterations prior to convergence will be discarded as "burn-in" values. There exist a number of procedures and techniques for determining the number of iterations necessary for convergence (e.g., Brooks & Gelman, 1998; Gelman & Rubin, 1992; Geweke, 1992; Heidelberger & Welch, 1983; Raftery & Lewis, 1992; see Brooks & Roberts, 1996, and Carlin, 1996 for reviews). We employ the procedure similar to that due to Gelman and Rubin (1992) based on classical analysis of variance; specifically we employ the extension of this procedure given by Brooks and Gelman (1998). This convergence assessment procedure calls for multiple chains to be run from overdispersed starting values. Two estimates of the variation inherent in the chain are computed. First, the average "within-chain" variation, the average of the individual chains' central 80% intervals is computed as an underestimate of the true variation. Next, the values from all chains are pooled together and the central 80% interval of this "pooled" chain is computed as an overestimate of the true variation. Upon convergence, the underestimate and the overestimate will converge to same value and the ratio of the overestimate to the underestimate should approach 1 (from above). There exist multivariate generalizations of this technique (Brooks &

Gelman, 1998), though we opt to monitor each parameter in a univariate manner so as to locate those parameters seemingly responsible for slow (or lack of) convergence.[18]

Empirical Analysis

The computer program WinBUGS 1.3 (Spiegelhalter et al., 1995) was used to obtain a Metropolis sampling solution to the model. Three chains were run in parallel for 100,000 iterations, each beginning with quite different starting values. The data set consisted of 195 examinees; taking between one and seven of the nine scenarios (typically, each scenario requires an hour and a half to complete), on average there were just over 25 values for each of the observables. Analysis of convergence consisted of monitoring the overestimate and the underestimate of the true posterior variance as described above and detailed in Brooks and Gelman (1998). Consideration of these convergence diagnostics indicated that as many as 36,000 iterations are necessary to achieve convergence. This slow convergence is in part due to the slow "mixing" of each individual chain due to considerably high autocorrelations, which in some cases was as high as .50, even for correlations of lag 40. In these cases, the individual chains mix quite slowly, thus chains starting from overdispersed starting values require a great number of iterations to converge.

Prior to data analysis, the first 40,000 iterations of each chain were discarded as "burn-in values" leaving 60,000 iterations per chain. These remaining iterations were pooled in the analysis of the final data for several reasons. First, all these iterations are empirical representations of the true posterior (i.e., values occur with the relative frequencies of the true

[18] At best, we can only determine which parameters for which it is doubtful that they have reached their true (marginal) posterior, which is of interest. We cannot determine with certainty if such parameters are "responsible" for slow convergence. Removing or fixing parameters that are seemingly slow to converge does not guarantee increased speed of convergence; conversely, such parameters may converge rapidly if other, seemingly nonproblematic parameters are removed or fixed.

posterior). Second, though there exists autocorrelations among the values *within* each chain, there is no correlation among the values *between* parallel chains i.e., the chains are independent. Pooling the values from parallel multiple chains serves to eliminate any serial dependence (Gelman, 1996). Finally, the use of multiple chains with overdispersed starting not only serves to detect lack of convergence, but also ensures that all chief regions of the posterior distribution are accounted for in the analysis (Gelman, 1996).

Empirical Results

A question of immediate interest concerns the impact of the data on the posteriors. A metric for summarizing the impact is the percent increase in precision, given as $100 \times \frac{(posterior\ SD)^{-2} - (prior\ SD)^{-2}}{(prior\ SD)^{-2}}$; a value of 0 indicates no new information is gained by incorporating the data while a value of 100 indicates that there is twice as much information regarding a parameter after incorporating the data. For the most part, there were mild increases in precision for the parameters in the Student Model. Likewise, there were mainly mild increases in the precision for the parameters that define the conditional distributions of the instrumental variables in each of the instantiations of the Evidence Models. Several large increases in posterior precision were observed in the parameters that define the conditional distributions of the observables. This is not a surprising result, as the evidence contained in the data, i.e., known values for certain observables, inform directly on the conditional distributions of observables, but only indirectly (via the propagation throughout the BIN) on the parameters that define the conditional distributions that are somewhat removed from the observables.

Conclusion and Pointers to the Future

One step in the immediate future is the assessment of model fit. Strategies for fit assessment include those detailed by Gelman et al., (1995) and Gilks, Richardson, and Spiegelhalter (1996). Many promising techniques involve the use of replicated data distributions (e.g., Mislevy, Senturk, et al., in press). Avenues for investigating model fit include (among others) analysis of the structural representations of the model. For instance, in the Student Model, *Networking Disciplinary Knowledge* served as a ceiling for *Network Modeling*; one alternative is to remove this constraint and investigate the impact. Other potential routes include relaxing the assumption of roughly spaced intervals of the variables or testing the necessity of the context variables in the Evidence Models. Other areas of future work concerning NetPASS include the collection of more data and the construction and investigation of new tasks.

An effort has been put forth to document the processes involved in the quantitative specification of the expected relationships between latent and observed variables and the subsequent estimation of the model via MCMC procedures. It has been emphasized that the procedures and techniques detailed and illustrated above have quite broad applicability for modeling in general and for modeling educational assessments in particular. That is, the use of Bayesian Inference Networks as a means of propagating information in assessment contexts is consistent with the role of assessment as an evidentiary argument regarding examinees. To that end, the construction and estimation of such networks is of the utmost importance. It is our hope that this work will lead to further research in the area of constructing and estimating similar measurement models used in complex assessments.

References

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37(1),* 29-51.

Brooks, S. P. (1998). Markov chain Monte Carlo method and its application. *The Statistician*, 47, 69-100.

Brooks, S. P., and Gelman, A. (1998). Alternative methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7, 434-455.

Brooks S. P., and Roberts, G. O. (1996). Discussion on Convergence of Markov chain Monte Carlo algorithms (by N. G. Polson). In J. M. Bernardo, A. F. M. Smith, A. P. Dawid, & J. O. Berger (Eds.), *Bayesian Statistics 5*. Oxford: Oxford University Press.

Casella, G., and George, E. I. (1992). Explaining the Gibbs sampler. The American Statistician 46: 167-174.

Chib, S., & Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. The American Statistician 49: 327-335.

Cowles, M. K., and Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. *J. Am. Statist. Ass.*, 91, 883-904.

Edwards, W. (1998). Hailfinder: Tools for and experiences with Bayesian normative modeling. *American Psychologist, 53,* 416-428.

Frederiksen, N., Mislevy, R.J., & Bejar, I.I. (Eds.). (1993). Test theory for a new generation of tests. Hillsdale, NJ: Erlbaum.

Gelman, A. (1996). Inference and monitoring convergence. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice* (pp. 131-143). London: Chapman and Hall.

Gelman, A. and Rubin, D. (1992). Inference from iterative sampling using multiple sequences. *Statistical Science,* 7, 457-511.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In J. M. Bernardo, A. F. M. Smith, A. P. Dawid, & J. O. Berger (Eds.), *Bayesian Statistics 4*. Oxford: Oxford University Press.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996a). Introducing Markov Chain Monte Carlo. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice* (pp. 1-19). London: Chapman and Hall.

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

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston: Kluwer.

Heidelberger, P. and Welch, P. D. (1983). Simulation run length control in the presence of an initial transient. *Ops Res.*, 31, 1109-1144.

Jensen, F.V. (1996). An introduction to Bayesian networks. New York: Springer-Verlag.

Jensen, F. V. (2001). Bayesian networks and decision graphs. New York: Springer-Verlag.

Lauritzen, S.L., & Spiegelhalter, D.J. (1988). Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *Journal of the Royal Statistical Society, Series B, 50,* 157-224.

Martin, J.D. & VanLehn, K. (1995). A Bayesian approach to cognitive assessment. In P. Nichols, S. Chipman, & R. Brennan (Eds.), Cognitively diagnostic assessment (pp. 141-165). Hillsdale, NJ: Erlbaum.

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

Mislevy, R.J., Almond, R.G., & Steinberg, L.S. (1998). A note on knowledge-based model construction in educational assessment. CSE Technical Report 480, Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Mislevy, R. J., & Patz, R. J. (1995) *On the consequences of ignoring certain conditional dependencies in cognitive diagnosis.* Presented at the Annual Meeting of the American Statistical Association, Orlando, FL, August, 1995.

Mislevy, R.J., Senturk, D., Almond, R.G., Dibello, L.V., Jenkins, F., Steinberg, L.S., & Yan, D. (in press). Modeling conditional probabilities in complex educational assessments. CSE Technical Report, Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (in press). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Discussion.*

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.

Mokken, R. J. (1997). Nonparametric models for dichotomous responses. In W. J. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (Chapter 20, pp. 351-367). New York: Springer.

Patz, R. J., Junker, B. W., & Johnson, M. S. (2000). *The hierarchical rater model for rated test items and its application to large scale educational assessment data*. Department of Statistics Technical Report 712. Pittsburgh: Carnegie Mellon University. (Available from http://www.stat.cmu.edu/cmu-stats/tr/tr712/tr712.html)

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference.* San Mateo, CA: Kaufmann.

Raftery, A. E., and Lewis, S. M. (1992). How many iterations in the Gibbs sampler? In J. M. Bernardo, A. F. M. Smith, A. P. Dawid, & J. O. Berger (Eds.), *Bayesian Statistics 4*, 763-774. Oxford: Oxford University Press.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9*, 401-412.

Roberts, G. O. (1996). Markov chain concepts related to sampling algorithms. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice* (pp. 45-57). London: Chapman and Hall.

Schum, D.A. (1987). *Evidence and inference for the intelligence analyst.* Lanham, Md.: University Press of America.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph No. 17.

Spiegelhalter, D.J., Dawid, A.P., Lauritzen, S.L., & Cowell, R.G. (1993). Bayesian analysis in expert systems. *Statistical Science, 8,* 219-247.

Sympson, J. B. (1978). A model for testing multidimensional items. In D. J. Weiss (Ed.), *Proceedings of the 1977 Computerized Adaptive Testing Conference* (pp. 82-98).

Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago: University of Chicago Press.

Tierney, L. (1996). Introduction to general state-space Markov chain theory. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice* (pp. 59-74). London: Chapman and Hall.

Williamson, D. M., Bauer, M., Steinberg, L. S., Mislevy, R. J., & Behrens, J. T. (2003). Creating a complex measurement model using evidence centered design. Paper presented at the annual meeting of the National Council on Measurement in Education, April, 2003.

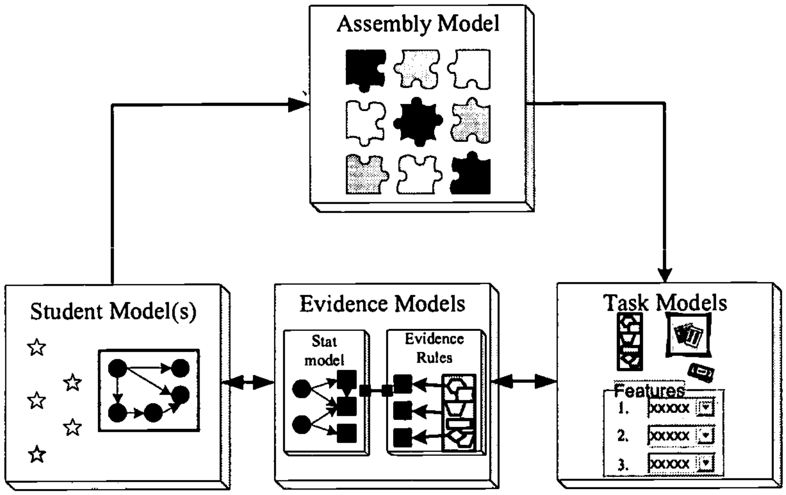

Figure 1: The Conceptual Assessment Framework

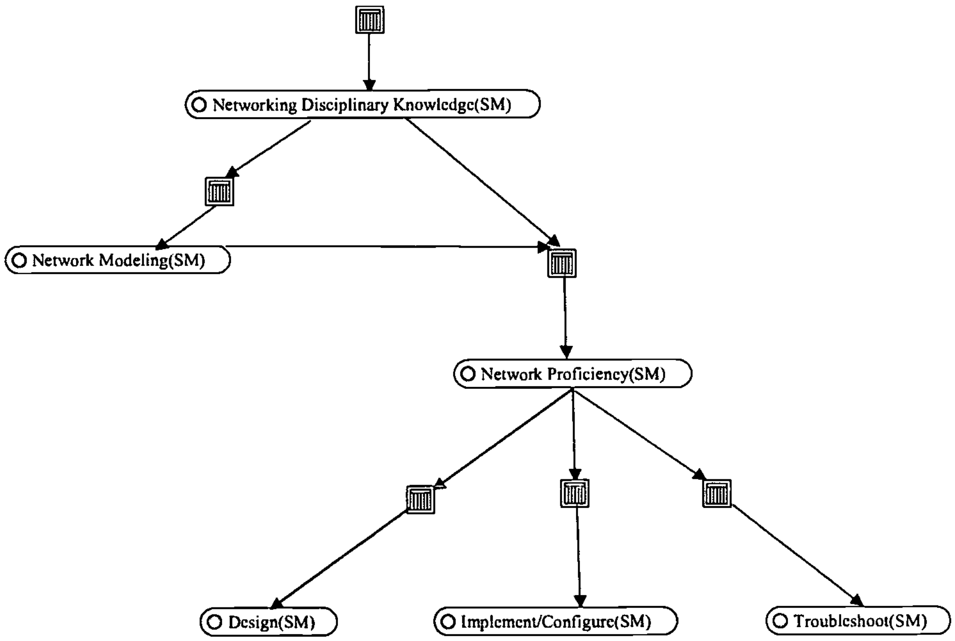

Figure 2: The NetPASS Student Model

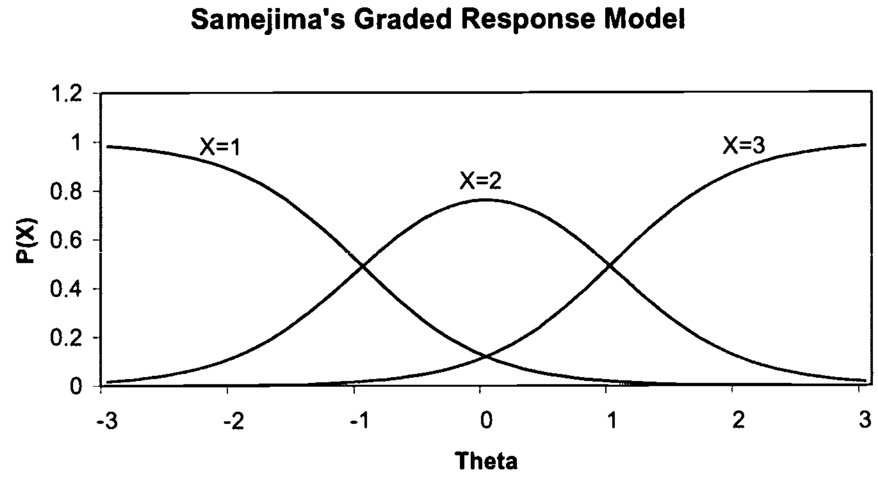

Figure 3:Response curves from the Graded Response IRT model with $a = 1$ and $\mathbf{b} = (-2, +2)$

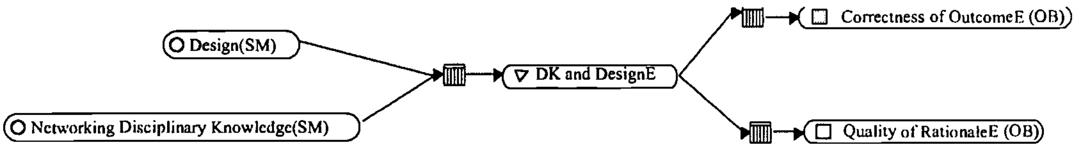
Design(SM)
Networking Disciplinary Knowledge(SM)
DK and DesignE
Correctness of OutcomeE (OB)
Quality of RationaleE (OB)

Figure 4: Part of a Design Evidence Model

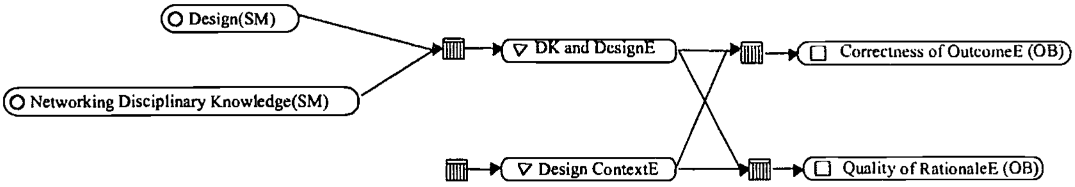
Design(SM)
Networking Disciplinary Knowledge(SM)
DK and DesignE
Design ContextE
Correctness of OutcomeE (OB)
Quality of RationaleE (OB)

Figure 5: A Complete Design Evidence Model

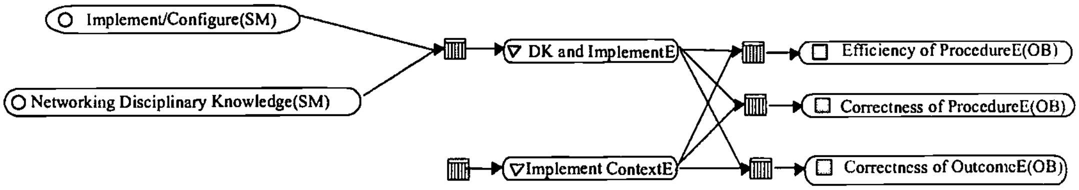
○ Implement/Configure(SM)
○ Networking Disciplinary Knowledge(SM)
▽ DK and ImplementE
▽Implement ContextE
□ Efficiency of ProcedureE(OB)
□ Correctness of ProcedureE(OB)
□ Correctness of OutcomeE(OB)

Figure 6: Part of an Implement Evidence Model

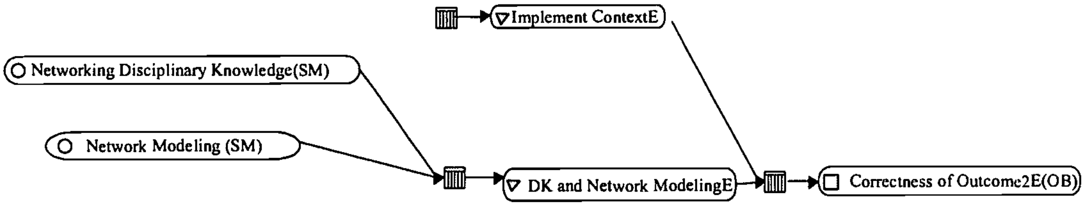
▽Implement ContextE
○ Networking Disciplinary Knowledge(SM)
○ Network Modeling (SM)
▽ DK and Network ModelingE
□ Correctness of Outcome2E(OB)

Figure 7: Part of an Implement Evidence Model

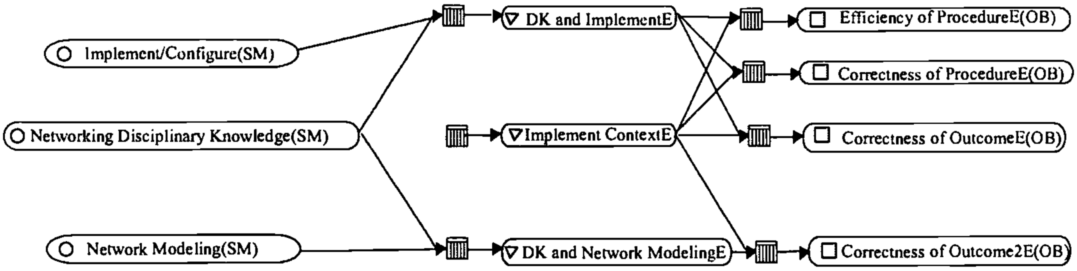
○ Implement/Configure(SM)
○ Networking Disciplinary Knowledge(SM)
○ Network Modeling(SM)
▽ DK and ImplementE
▽Implement ContextE
▽DK and Network ModelingE
□ Efficiency of ProcedureE(OB)
□ Correctness of ProcedureE(OB)
□ Correctness of OutcomeE(OB)
□ Correctness of Outcome2E(OB)

Figure 8: A Complete Implement Evidence Model

BEST COPY AVAILABLE

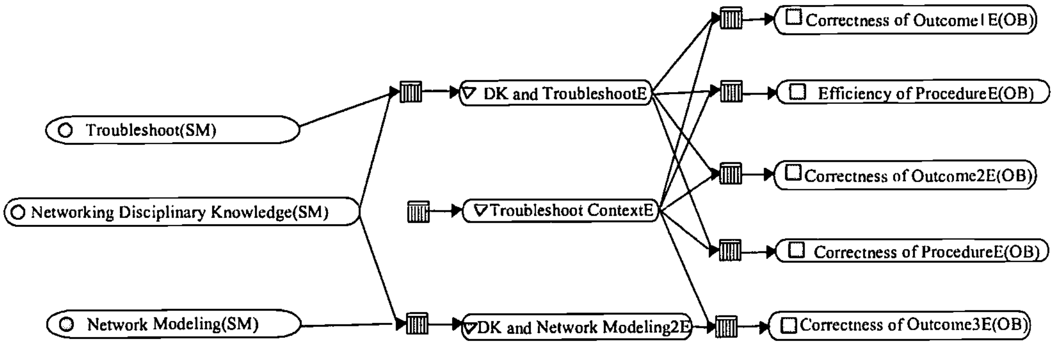
○ Troubleshoot(SM)
○ Networking Disciplinary Knowledge(SM)
○ Network Modeling(SM)
▽ DK and TroubleshootE
▽Troubleshoot ContextE
▽DK and Network Modeling2E
□Correctness of Outcome1E(OB)
□ Efficiency of ProcedureE(OB)
□Correctness of Outcome2E(OB)
□ Correctness of ProcedureE(OB)
□Correctness of Outcome3E(OB)

Figure 9: A Complete Troubleshoot Evidence Model

			Pr (Design = *k*)				
Network Proficiency	$\theta_{NetworkProficiency}$	θ^{**}_{Design}	Novice	Semester 1	Semester 2	Semester 3	Semester 4
Novice	1	-3.8	0.689974	0.252701	0.049162	0.00705	0.001113
Semester 1	2	-1.8	0.231475	0.458499	0.252701	0.049162	0.008163
Semester 2	3	0.2	0.039166	0.192309	0.458499	0.252701	0.057324
Semester 3	4	2.2	0.005486	0.033679	0.192309	0.458499	0.310026
Semester 4	5	4.2	0.000746	0.00474	0.033679	0.192309	0.768525

$$\theta^{**}_{Design} \equiv c_{Design} \times \theta_{NetworkProficiency} + d_{Design}$$

$$\theta^{**}_{Design} \equiv 2 \times \theta_{NetworkProficiency} + (-5.8)$$

Table 1: Conditional Probability Table for *Design*

			Pr (Implement = *k*)				
Network Proficiency	$\theta_{NetworkProficiency}$	$\theta^{**}_{Implement}$	Novice	Semester 1	Semester 2	Semester 3	Semester 4
Novice	1	-4.2	0.768525	0.192309	0.033679	0.00474	0.000746
Semester 1	2	-2.2	0.310026	0.458499	0.192309	0.033679	0.005486
Semester 2	3	-0.2	0.057324	0.252701	0.458499	0.192309	0.039166
Semester 3	4	1.8	0.008163	0.049162	0.252701	0.458499	0.231475
Semester 4	5	3.8	0.000746	0.00474	0.033679	0.192309	0.768525

$$\theta^{**}_{Implement} \equiv c_{Implement} \times \theta_{NetworkProficiency} + d_{Implement}$$

$$\theta^{**}_{Implement} \equiv 2 \times \theta_{NetworkProficiency} + (-6.2)$$

Table 2: Conditional Probability Table for *Implement*

			Pr (Troubleshoot = k)				
Network Proficiency	$\theta_{NetworkProficiency}$	$\theta^{**}_{Troubleshoot}$	Novice	Semester 1	Semester 2	Semester 3	Semester 4
Novice	1	-5.0	0.880797	0.101217	0.015514	0.002137	0.000335
Semester 1	2	-3.0	0.5	0.380797	0.101217	0.015514	0.002473
Semester 2	3	-1.0	0.119203	0.380797	0.380797	0.101217	0.017986
Semester 3	4	1.0	0.017986	0.101217	0.380797	0.380797	0.119203
Semester 4	5	3.0	0.002473	0.015514	0.101217	0.380797	0.5

$$\theta^{**}_{Troubleshoot} \equiv c_{Troubleshoot} \times \theta_{NetworkProficiency} + d_{Troubleshoot}$$
$$\theta^{**}_{Troubleshoot} \equiv 2 \times \theta_{NetworkProficiency} + (-7.0)$$

Table 3: Conditional Probability Table for *Troubleshoot*

			Pr (Network Modeling = k)				
Network Disciplinary Knowledge	$\theta_{NetworkDisciplinaryKnowledge}$	$\theta^{**}_{NetworkModeling}$	Novice	Semester 1	Semester 2	Semester 3	Semester 4
Novice	1	-6.0	0.952574	0.040733	0.005782	0.000788	0.000123
Semester 1	2	-4.0	0.731059	0.221516	0.040733	0.005782	0.000911
Semester 2	3	-2.0	0.268941	0.462117	0.221516	0.040733	0.006693
Semester 3	4	0.0	0.047426	0.221516	0.462117	0.221516	0.047426
Semester 4	5	2.0	0.006693	0.040733	0.221516	0.462117	0.268941

$$\theta^{**}_{NetworkModeling} \equiv c_{NetworkModeling} \times \theta_{NetworkDisciplinaryKnowledge} + d_{NetworkModeling}$$
$$\theta^{**}_{NetworkModeling} \equiv 2 \times \theta_{NetworkDisciplinaryKnowledge} + (-8.0)$$

Table 4: Unstructured Conditional Probability Table for *Network Modeling*

			Pr (Network Modeling = *k*)				
Network Disciplinary Knowledge	$\theta_{NetworkDisciplinaryKnowledge}$	$\theta^{**}_{NetworkModeling}$	Novice	Semester 1	Semester 2	Semester 3	Semester 4
Novice	1	-6.0	1.0	0	0	0	0
Semester 1	2	-4.0	0.767456	0.232544	0	0	0
Semester 2	3	-2.0	0.282331	0.485125	0.232544	0	0
Semester 3	4	0.0	0.049787	0.232544	0.485125	0.232544	0
Semester 4	5	2.0	0.006693	0.040733	0.221516	0.462117	0.268941

Table 5: Corrected Conditional Probability Table for *Network Modeling*

					P(NP=k)				
NDK	θ_{NDK}	Network Modeling	θ_{NM}	θ^{**}_{NP}	Novice	Sem 1	Sem 2	Sem 3	Sem 4
Novice	1	Novice	1	-3	0.5	0.380797	0.101217	0.015514	0.002473
Sem 1	2	Novice	1	-2	0.268941	0.462117	0.221516	0.040733	0.006693
Sem 1	2	Sem 1	2	-1	0.119203	0.380797	0.380797	0.101217	0.017986
Sem 2	3	Novice	1	-1	0.119203	0.380797	0.380797	0.101217	0.017986
Sem 2	3	Sem 1	2	0	0.047426	0.221516	0.462117	0.221516	0.047426
Sem 2	3	Sem 2	3	1	0.017986	0.101217	0.380797	0.380797	0.119203
Sem 3	4	Novice	1	0	0.047426	0.221516	0.462117	0.221516	0.047426
Sem 3	4	Sem 1	2	1	0.017986	0.101217	0.380797	0.380797	0.119203
Sem 3	4	Sem 2	3	2	0.006693	0.040733	0.221516	0.462117	0.268941
Sem 3	4	Sem 3	4	3	0.002473	0.015514	0.101217	0.380797	0.500000
Sem 4	5	Novice	1	1	0.017986	0.101217	0.380797	0.380797	0.119203
Sem 4	5	Sem 1	2	2	0.006693	0.040733	0.221516	0.462117	0.268941
Sem 4	5	Sem 2	3	3	0.002473	0.015514	0.101217	0.380797	0.500000
Sem 4	5	Sem 3	4	4	0.000911	0.005782	0.040733	0.221516	0.731059
Sem 4	5	Sem 4	5	5	0.000335	0.002137	0.015514	0.101217	0.880797

Table 6: Unstructured Conditional Probability Table for *Network Proficiency*

					P(NP=k)				
NDK	θ_{NDK}	Network Modeling	θ_{NM}	θ^{**}_{NP}	Novice	Sem 1	Sem 2	Sem 3	Sem 4
Novice	1	Novice	1	-3	1	0	0	0	0
Sem 1	2	Novice	1	-2	0.36788	0.63212	0	0	0
Scm 1	2	Sem 1	2	-1	0.23840	0.76160	0	0	0
Sem 2	3	Novice	1	-1	0.13534	0.43233	0.43233	0	0
Sem 2	3	Sem 1	2	0	0.06487	0.30301	0.63212	0	0
Sem 2	3	Sem 2	3	1	0.03597	0.20243	0.76160	0	0
Sem 3	4	Novice	1	0	0.04979	0.23254	0.48513	0.23254	0
Sem 3	4	Sem 1	2	1	0.02042	0.11491	0.43233	0.43233	0
Sem 3	4	Sem 2	3	2	0.00916	0.05572	0.30301	0.63212	0
Sem 3	4	Sem 3	4	3	0.00495	0.03103	0.20244	0.76159	0
Sem 4	5	Novice	1	1	0.01799	0.10122	0.38080	0.38080	0.11920
Sem 4	5	Sem 1	2	2	0.00669	0.04073	0.22152	0.46212	0.26894
Sem 4	5	Sem 2	3	3	0.00247	0.01551	0.10122	0.369997	0.50000
Sem 4	5	Sem 3	4	4	0.00091	0.00578	0.04073	0.22152	0.73106
Sem 4	5	Sem 4	5	5	0.00034	0.00214	0.01551	0.10122	0.88079

Table 7: Corrected Conditional Probability Table for *Network Proficiency*

Pr (Network Disciplinary Knowledge = *k*)				
Novice	Semester 1	Semester 2	Semester 3	Semester 4
.01477	.08498	.35042	.40798	.14185

Table 8: Probability Table for *Network Disciplinary Knowledge*

		Pr (NDK and DesignE = *k*)				
Network Disciplinary Knowledge	Design	Novice	Semester 1	Semester 2	Semester 3	Semester 4
Novice	Novice	1.0	0	0	0	0
Novice	Semester 1	1.0	0	0	0	0
Novice	Semester 2	1.0	0	0	0	0
Novice	Semester 3	1.0	0	0	0	0
Novice	Semester 4	1.0	0	0	0	0
Semester 1	Novice	1.0	0	0	0	0
Semester 1	Semester 1	0.36788	0.63212	0	0	0
Semester 1	Semester 2	0.30638	0.69362	0	0	0
Semester 1	Semester 3	0.25799	0.74201	0	0	0
Semester 1	Semester 4	0.22159	0.77841	0	0	0
Semester 2	Novice	1.0	0	0	0	0
Semester 2	Semester 1	0.33548	0.66452	0	0	0
Semester 2	Semester 2	0.06487	0.30301	0.63212	0	0
Semester 2	Semester 3	0.05002	0.25636	0.69362	0	0
Semester 2	Semester 4	0.0398	0.21819	0.74201	0	0
Semester 3	Novice	1.0	0	0	0	0
Semester 3	Semester 1	0.30638	0.69362	0	0	0
Semester 3	Semester 2	0.05676	0.27872	0.66452	0	0
Semester 3	Semester 3	0.00916	0.05572	0.30301	0.63212	0
Semester 3	Semester 4	0.00696	0.04306	0.25636	0.69362	0
Semester 4	Novice	1.0	0	0	0	0
Semester 4	Semester 1	0.28058	0.71942	0	0	0
Semester 4	Semester 2	0.05002	0.25636	0.69362	0	0
Semester 4	Semester 3	0.00795	0.04881	0.27872	0.66452	0
Semester 4	Semester 4	0.00091	0.00578	0.04073	0.22152	0.73106

$$\theta^{**}_{DKandDesignE} \equiv [c_{DKandDesignE} \times \theta^{*}_{DKandDesign} + d_{DKandDesignE}] + [c_{NDKE} \times (\theta_{NDK} - \theta^{*}_{DKandDesign})] + [c_{DesignE} \times (\theta_{Design} - \theta^{*}_{DKandDesign})]$$

$$\theta^{**}_{DKandDesignE} \equiv [2 \times \theta^{*}_{DKandDesign} + -6.0] + [.2 \times (\theta_{NDK} - \theta^{*}_{DKandDesign})] + [.4 \times (\theta_{Design} - \theta^{*}_{DKandDesign})]$$

Table 9: Conditional Probability Table for *NDK and DesignE*

					P(X=k)		
DKandDesignE	$\theta_{DKandDesignE}$	Design ContextE	$\theta_{DesignContextE}$	θ_t^{**}	Low	Medium	High
Novice	1	Low	-1	-1.7	0.802184	0.19332	0.004496
Novice	1	High	1	-1.3	0.645656	0.344392	0.009952
Semester 1	2	Low	-1	-0.7	0.354344	0.613361	0.032295
Semester 1	2	High	1	-0.3	0.197816	0.733045	0.069138
Semester 2	3	Low	-1	0.3	0.069138	0.733045	0.197816
Semester 2	3	High	1	0.7	0.032295	0.613361	0.354344
Semester 3	4	Low	-1	1.3	0.009952	0.344392	0.645656
Semester 3	4	High	1	1.7	0.004496	0.19332	0.802184
Semester 4	5	Low	-1	2.3	0.001359	0.06778	0.930862
Semester 4	5	High	1	2.7	0.000611	0.031685	0.967705

$$\theta_t^{**} = c_{tDKandDesignE} \times (\theta_{DKandDesignE}) + c_{tDesignContextE} \times (\theta_{DesignContextE}) + d_{tDKandDesignE}$$

$$\theta_t^{**} = 2 \times (\theta_{DKandDesignE}) + .4 \times (\theta_{DesignContextE}) + (-5.0)$$

Table 10: Conditional Probability Table for the observables in the Design Easy Evidence Model

					P(X=k)		
DKandDesignM	$\theta_{DKandDesignM}$	Design ContextM	$\theta_{DesignContextM}$	θ_t^{**}	Low	Medium	High
Novice	1	Low	-1	-2.2	0.916827	0.081514	0.001659
Novice	1	High	1	-1.8	0.832018	0.164297	0.003684
Semester 1	2	Low	-1	-1.2	0.598688	0.389184	0.012128
Semester 1	2	High	1	-0.8	0.401312	0.572091	0.026597
Semester 2	3	Low	-1	-0.2	0.167982	0.748846	0.083173
Semester 2	3	High	1	0.2	0.083173	0.748846	0.167982
Semester 3	4	Low	-1	0.8	0.026597	0.572091	0.401312
Semester 3	4	High	1	1.2	0.012128	0.389184	0.598688
Semester 4	5	Low	-1	1.8	0.003684	0.164297	0.832018
Semester 4	5	High	1	2.2	0.001659	0.081514	0.916827

$$\theta_t^{**} = c_{tDKandDesignM} \times (\theta_{DKandDesignM}) + c_{tDesignContextM} \times (\theta_{DesignContextM}) + d_{tDKandDesignM}$$

$$\theta_t^{**} = 2 \times (\theta_{DKandDesignM}) + .4 \times (\theta_{DesignContextM}) + (-6.0)$$

Table 11: Conditional Probability Table for the observables in the Design Medium Evidence Model

					P(X=k)		
DKandDesignH	$\theta_{DKandDesignH}$	Design ContextH	$\theta_{DesignContextH}$	θ_t^{**}	Low	Medium	High
Novice	1	Low	-1	-2.7	0.967705	0.031685	0.000611
Novice	1	High	1	-2.3	0.930862	0.06778	0.001359
Semester 1	2	Low	-1	-1.7	0.802184	0.19332	0.004496
Semester 1	2	High	1	-1.3	0.645656	0.344392	0.009952
Semester 2	3	Low	-1	-0.7	0.354344	0.613361	0.032295
Semester 2	3	High	1	-0.3	0.197816	0.733045	0.069138
Semester 3	4	Low	-1	0.3	0.069138	0.733045	0.197816
Semester 3	4	High	1	0.7	0.032295	0.613361	0.354344
Semester 4	5	Low	-1	1.3	0.009952	0.344392	0.645656
Semester 4	5	High	1	1.7	0.004496	0.19332	0.802184

$$\theta_t^{**} = c_{tDKandDesignH} \times (\theta_{DKandDesignH}) + c_{tDesignContextH} \times (\theta_{DesignContextH}) + d_{tDKandDesignH}$$

$$\theta_t^{**} = 2 \times (\theta_{DKandDesignH}) + .4 \times (\theta_{DesignContextH}) + (-7.0)$$

Table 12: Conditional Probability table for the observables in the Design Hard Evidence Model

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

ERIC
Educational Resources Information Center

REPRODUCTION RELEASE

(Specific Document)

TM034955

I. DOCUMENT IDENTIFICATION:

Title: Specifying and Refining a Complex Measurement Model

Author(s): Roy Levy and Robert J. Mislevy

Corporate Source: University of Maryland

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

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☑

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

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

Level 2A

□

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

□

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:

Printed Name/Position/Title: Roy Levy

Organization/Address: Department of Measurement, Statistics and Evaluation
1230 Benjamin Building, U of MD College Park MD 20742

Telephone: 301 405-3623

FAX:

E-Mail Address: levyr@wam.umd.edu

Date: 5/7/03

(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: University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Lab, Bldg 075
College Park, MD 20742
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:

University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Lab, Bldg 075
College Park, MD 20742
Attn: Acquisitions