Specifying and Refining a Measurement Model for a Simulation-Based Assessment

CSE Report 619

Roy Levy and Robert J. Mislevy

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)/
University of Maryland

January 2004

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

Project 3.6 Study Group on Cognitive Validity, Strand 1: Cognitively Based Models and Assessment Design
Project Director: Robert J. Mislevy, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of Maryland

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B60002, as administered by the Institute of Education Sciences, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or policies of the National Center for Education Research, the Institute of Education Sciences, or the U.S. Department of Education.

SPECIFYING AND REFINING A MEASUREMENT MODEL FOR A SIMULATION-BASED ASSESSMENT[1]

Roy Levy and Robert J. Mislevy
CRESST/University of Maryland, College Park

Abstract

The challenges of modeling students' performance in simulation-based assessments include accounting for multiple aspects of knowledge and skill that arise in different situations and the conditional dependencies among multiple aspects of performance in a complex assessment. This paper describes a Bayesian approach to modeling and estimating cognitive models in such situations, in terms of both statistical machinery and actual instrument development. The 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 (MCMC) procedures. This process is illustrated in the context 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.

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, for example, a scantron sheet with circles filled in, essays to be scored by raters, and portfolios. Though 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 needing inferential procedures to be developed individually. Nonstandard and complex tasks result in complex work products, and different combinations of knowledge and skill may be tapped in different tasks or subtasks. Drawing proper inferences in these situations requires models that accumulate and incorporate

[1] We wish to thank David Williamson, Malcolm Bauer, Russell Almond, Duanli Yan, and Margaret Redman of Educational Testing Service and John Behrens and Sarah Demark of Cisco Learning Institute for their involvement with, and their support of our participation in, the NetPASS project.

information in order to produce "scores" that are 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, and more specifically to inference in educational measurement; and moreover, that it is particularly useful for inferences from innovative and complex assessment instruments (Mislevy, 1994). In what follows, we describe such probability-based reasoning in detail, and illustrate ensuing methods in practice via an example from a complex assessment of the cognitive development of students in the Cisco Networking Academy Program (CNAP). We draw upon language and concepts of the evidence-centered assessment design methodology of Mislevy, Steinberg, and Almond (2003), referring in particular to Student Models, Evidence Models, and Task Models of the conceptual assessment framework or CAF.

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. Though the particulars of NetPASS will be described in some 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.

Bayesian Inference Networks in Assessment

A Bayesian approach to assessment starts by characterizing aspects of students' knowledge and skill in terms of a vector-valued "Student Model variable" θ, and aspects of their behavior in terms of possibly vector-valued "observable variables" X. Conditional probability distributions $P(X \mid \theta)$, obtained through theory, expert opinion, empirical data, or some combination of these, characterize how performance depends on knowledge and skill in task situations. Letting the "prior" probability distribution $P(\theta)$ denote the assessor's belief about a student's θ at a given point in time, observing X leads to an updated "posterior" probability distribution $P(\theta \mid X)$ by Bayes theorem.

Though the required calculations can be carried out in simple situations using the textbook definition of Bayes theorem, computation for larger, more complex

situations quickly becomes infeasible. Recent developments with Bayesian inference networks (BINs; Jensen, 1996, 2001) permit Bayesian updating even in very large collections of variables, when conditional independence relationships posited by theory or entailed by observational designs can be exploited. Fortunately, this is often the case in educational assessment, so BINs can serve as the statistical model for updating Student Model variables (see Martin & VanLehn, 1995, and Mislevy, 1994, on the use of BINs in assessment). The relationships among variables in a BIN constitute the reasoning structures of the network. 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 1, depicting the NetPASS Student Model, is an example) of a joint probability distribution over a set of random variables, and consists of the following elements (Jensen, 1996):

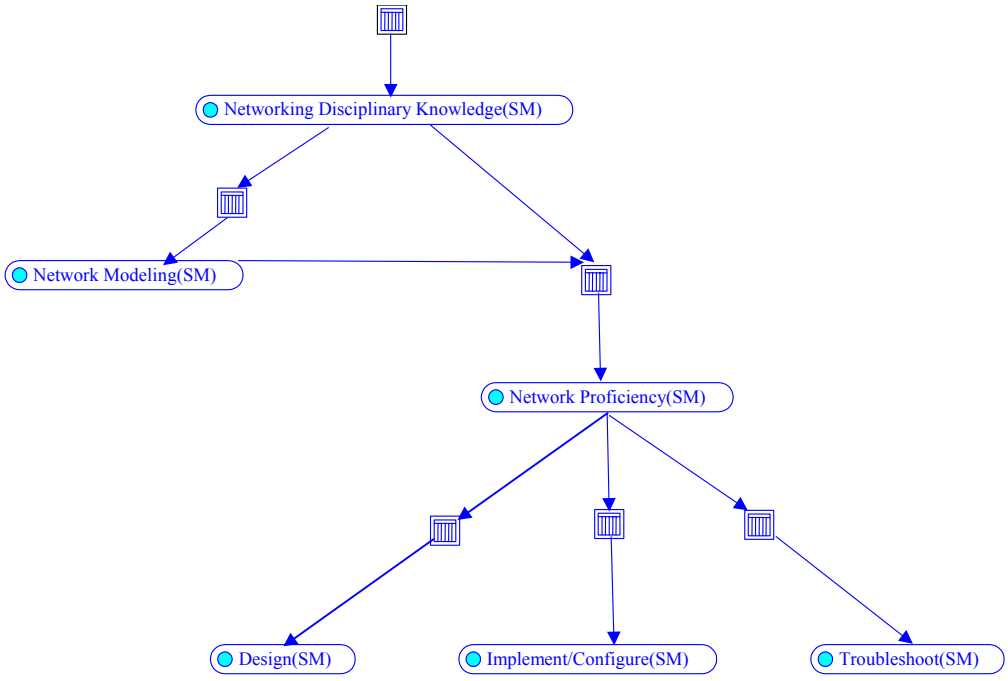

Figure 1. The NetPASS Student Model.

- 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 1, for example, *Design* is a child of *Network Proficiency*, and both *Network Disciplinary Knowledge* and *Network Modeling* are parents of *Network Proficiency*.

- 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.

- The variables and the directed edges together constitute 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 variable, there is an 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. Associated with variables having no parents, such as *Network Disciplinary Knowledge* in Figure 1, are unconditional probability distributions.

As described below, we can define fragments of BINs in terms of a BIN for Student Model variables and a BIN for conditional distributions of the observable variables of each task, or Evidence Model BINs. Characteristics of tasks can be important in determining the conditional probabilities in evidence model BIN fragments; in the sequel, we shall refer to Task Model variable Y in this connection. Before turning to the probability framework used to represent these models, we will note some recurring ways that variables in BINs for assessment relate to one another, and which we will want to build into conditional probability distributions.

Relationships Among Variables

This section sketches out a variety of evidentiary structures among the Student Model and observable variables. Though certainly not an exhaustive set of all possible structures, these structures appear repeatedly in the NetPASS assessment.

Bivariate Relationships in Modeling Skills Involved in an Assessment

Bivariate relationships concern 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: The value of one variable influences expectations for the other in the form of a probability distribution.

- Ceiling: The value of one variable not only influences the expectation of the other, but sets a maximum value that the other one can take. For example, the value of the child cannot exceed the value of the parent.

Multivariate Relationships in Modeling Performance on an Assessment

Multivariate relationships involve 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 are generalizations of the bivariate relationships discussed above. The following illustrations concern modeling performance—that is, observable variables modeled as dependent on multiple skills and abilities.

- Conjunctions: A generalization of the ceiling relationship. The minimum value of the skills defines the ceiling for performance; the absence of any of them leads to an 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 leads to an expectation of an increase in performance.

- Conditional dependence relationships: 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, & Mariano, 2002).

These basic structures represent but a few of the limitless number of ways to model relationships. For other common structures, see Mislevy et al. (2002). While estimates of these relationships can come from data, the assessment designer's

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, Carlin, Stern, and Rubin (1995, p. 3) defined 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 noted, "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 (a) targeted knowledge, skills and abilities, (b) ways in which such knowledge, skills, and abilities are demonstrated in performance, and (c) characteristics of situations that provide the opportunity to observe such performance. The Student Model, Evidence Models, and Task Models provide this information (Williamson, Bauer, Steinberg, Mislevy, & Behrens, 2003).

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

$$\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{iP}) \tag{1}$$

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

$$\mathbf{Y}_j = (Y_{j1}, \ldots, Y_{jL}), \tag{2}$$

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

$$\mathbf{X}_j = (X_{j1}, \ldots, X_{jM}) \tag{3}$$

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 on the extracting of observables from work products via the evaluation component of Evidence Models, let us assume these important procedures have been completed, providing 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; that is, 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,

$$\boldsymbol{\theta}_i \sim P(\boldsymbol{\theta}_i \mid \boldsymbol{\lambda}). \tag{4}$$

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 of responding to observable m from task j with a value of k. The collection of these, for any particular observable, is then

$$\pi_{mj} = (\pi_{mj1}, \pi_{mj2}, \ldots, \pi_{mjK}), \tag{5}$$

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, that is, the conditional probability of X_{imj} given $\boldsymbol{\theta}_i$. More formally, if

$$\pi_{mjk} = P(X_{imj} = x_{imjk} \mid \boldsymbol{\theta}_i), \tag{6}$$

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

$$X_{imj} \sim P(X_{imj} \mid \boldsymbol{\theta}_i, \pi_{mj}). \tag{7}$$

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 observables, 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

$$\pi_{mj} \sim P(\pi_{mj} \mid \eta_{mj}), \tag{8}$$

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 parameters that define those distributions is denoted $\boldsymbol{\eta}$.

The joint probability of all parameters can be expressed as

$$P(\boldsymbol{\lambda}, \boldsymbol{\eta}, \boldsymbol{\theta}, \boldsymbol{\pi}, \mathbf{X}) = P(\boldsymbol{\lambda}) \times P(\boldsymbol{\eta} \mid \boldsymbol{\lambda}) \times P(\boldsymbol{\theta} \mid \boldsymbol{\lambda}, \boldsymbol{\eta}) \times P(\boldsymbol{\pi} \mid \boldsymbol{\lambda}, \boldsymbol{\eta}, \boldsymbol{\theta}) \times P(\mathbf{X} \mid \boldsymbol{\lambda}, \boldsymbol{\eta}, \boldsymbol{\theta}, \boldsymbol{\pi}). \tag{9}$$

Taking advantage of the conditional independence relationships implied in eqs. (4)–(8), this expression can be simplified in light of additional knowledge and assumptions we bring to the assessment context, as follows:

$$P(\boldsymbol{\lambda}, \boldsymbol{\eta}, \boldsymbol{\theta}, \boldsymbol{\pi}, \mathbf{X}) = P(\boldsymbol{\lambda}) \times P(\boldsymbol{\theta} \mid \boldsymbol{\lambda}) \times P(\boldsymbol{\eta}) \times P(\boldsymbol{\pi} \mid \boldsymbol{\eta}) \times P(\mathbf{X} \mid \boldsymbol{\theta}, \boldsymbol{\pi}). \tag{10}$$

In setting up the full model, our goal then becomes to specify the forms of the various terms in eq. (10). We have already mentioned that we think of observable variables as conditional on Student Model variables. In a complex assessment that includes multiple Student Model variables that are related, such as NetPASS, there is the need to model the dependencies among the Student Model variables. Much of the discussion regarding modeling observables 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 in NetPASS, 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, Dawid, Lauritzen, & Cowell, 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. We follow Mislevy et al. (2002) in exploiting experience from 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. Samejima's Graded Response Model (GRM; 1969) can be used to model an ordinal polytomous outcome variable X_{ij}. For an observable variable X_{ij} that can take on any integral value from 1 to *K* define the probability that the response is in category *k* or above as

$$P(X_{ij} \geq k) = \text{logit}^{-1}(a_j(\theta_i - b_{jk})), \tag{11}$$

for *k*=2,…,*K*, where b_{jk} is the location parameter associated with separating the k^{th} from the k-1^{th} category. The probability of response being in the k^{th} category is

$$P(X_{ij} = k) = P(X_{ij} \geq k) - P(X_{ij} \geq k+1). \tag{12}$$

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

Note the parsimony of the model. For example, 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, only three parameters require estimation: the discrimination a_j and the two category boundaries contained in $\mathbf{b}$. Though the GRM was introduced in terms of modeling a polytomous variable as conditional on a continuous variable (Samejima, 1969), the current application follows the use of the logistic function in modeling polytomous variables as dependent on a discrete variable (see, e.g., Formann, 1985; Formann & Kohlmann, 1998).

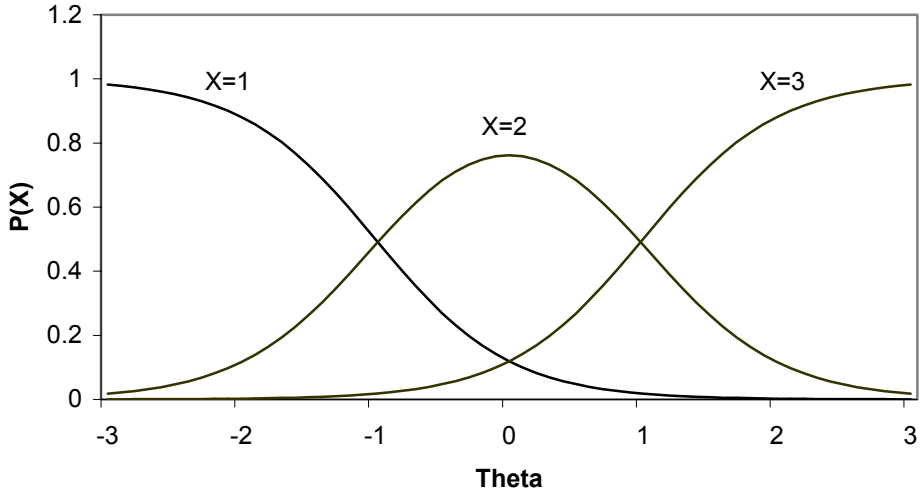

Figure 2. Response curves from the Graded Response IRT model with $a = 1$ and $\mathbf{b} = (-2, +2)$.

Applications in NetPASS

The logic of the GRM can be extended to fit child variables 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 modeling observable variables on latent variables. In NetPASS we also 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). For another example of using an ordered polytomous IRT model to model latent variables, see Patz et al. (2002).

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 (Andrich, 1982). 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: one variable, θ_i, 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) model or distill down the relationships between multiple parents and children to fit the unidimensional GRM. We proceed with 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 as to produce one variable that will serve in the unidimensional GRM; this variable is an "effective theta" denoted as θ^{**}. In IRT models, the conditional probabilities of response are determined by theta and the "item" parameters a_{mj} and $\mathbf{b}_{mj}$.[2] 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 $\mathbf{b}_{mj}$) and estimates the examinee's latent trait(s) with respect to each item. The impact of the item, in terms of both 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 et al., 2002). First, the use of paradigmatic structures to characterize relationships among variables may be comforting to subject matter experts (SMEs), who while familiar with the domain and the structure of knowledge and therefore able to provide the form of 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

[2] For ease of exposition, we will continue to discuss the effective theta method in terms of items (i.e., an observable child variable). As with the GRM, the effective theta method is not restricted to the case of observable child variables.

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 good 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

$$\pi_k = P(\theta_c = k \mid \theta_1) \tag{13}$$

where, as before, k=1,…,K are the possible values of θ_c. 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 functions and the GRM is required.

As will be described below, the relationships between all of the variables in NetPASS are positive. As illustrated in Figure 2, there is a positive monotonic relationship between theta and the response category: As theta increases, the probabilities of higher levels of response increase. 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.

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:

$$\theta^{**} \equiv g(\theta_1) = c \times \theta_1 + d\ . \tag{14}$$

Note the simplicity of the model: There are two parameters to estimate, c and d, *regardless* of the number of states of the parent or the child. 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 typical IRT models a and b parameters define the conditional probability distribution. The constant

parameter, d, is akin to b in eq. (11) and is related to the average value for the child variable. The slope parameter, c, is akin to a in eq. (11) and defines the strength of association between θ_1 and θ_c. 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.[3] 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. As generalizations of unidimensional models, two classes of multidimensional models are (a) combinations of linear mappings (e.g., compensatory relationships, discussed below) and (b) linear mappings of combinations (e.g., conjunctive relationships, discussed below). 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 et al. (2002).

The Application of the Effective Theta Method to the GRM

The effective theta method fixes the a and $\mathbf{b}$ parameters 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, we set $a = 1$ and $\mathbf{b} = (-2, +2)$. When using the effective theta method and the GRM to model values of latent variables, we set $a = 1$ 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.

---

[3] 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 & 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.

The Student Model

Properties of Student Model Variables

The NetPASS Student Model, on the whole, aims to represent the knowledge, skills, and abilities that are important for success at CNAP. The operational Student Model (Figure 1) also includes the specification of statistical relationships among variables. All the variables described in this section are 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 a student's level on that particular aspect of the domain; these values are coded as 1-5, respectively.

Quantitative Modeling of Relationships in the Student Model

In terms of the joint probability distribution (eq. (10)), 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. Where possible, the subscript identifying the variable will be abbreviated, that is, θ_{NDK} refers to *Network Disciplinary Knowledge*, θ_{NM} refers to *Network Modeling*, and θ_{NP} refers to *Network Proficiency*.

Direct dependence. With direct dependence, the value of the child is dependent on only one parent, which determines a probability distribution for the child. We thus define the effective theta as a linear function of the lone parent variable:

$$\theta_c^{**} \equiv c_{c,1} \times \theta_1 + d_{c,1} \tag{15}$$

where θ_c^{**} is the effective theta for the distribution of the child, and θ_1 is the parent.[4]

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. (15):

$$\theta_{Design}^{**} \equiv c_{Design,NP} \times \theta_{NP} + d_{Design,NP}\,. \tag{16}$$

---

[4] Though it may seem superfluous for simple equations, we will subscript the parameters (here $c_{c,1}$ and $d_{c,1}$ with the child variable followed by the parent variable.

Effective thetas calculated for all possible values of *Network Proficiency* with $c_{Design,NP}$ = 2 and $d_{Design,NP}$ = –5.8 and are given in Table 1. The values for $c_{Design,NP}$ and $d_{Design,NP}$ 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 $c_{Design,NP}$ and $d_{Design,NP}$. Because using values of 2 and –5.8 results in the conditional distribution experts expect, our prior distributions for each parameter will be based on these values.

Similarly, to obtain the effective theta for *Implement*, instantiate eq. (15):

$$\theta^{**}_{Implement} \equiv c_{Implement,NP} \times \theta_{NP} + d_{Implement,NP}. \tag{17}$$

Table 2 displays effective thetas calculated for all possible values of *Network Proficiency* with $c_{Implement,NP}$ = 2 and $d_{Implement,NP}$ = –6.2. These values represent expert expectations and will serve as the basis for the prior distributions in the calibration. The resulting conditional probabilities are also given in Table 2. Likewise, the effective theta for *Troubleshoot* is defined as:

$$\theta^{**}_{Troubleshoot} \equiv c_{Troubleshoot,NP} \times \theta_{NP} + d_{Troubleshoot,NP} \tag{18}$$

Table 3 contains the conditional probabilities obtained with $c_{Troubleshoot,NP}$ = 2 and $d_{Troubleshoot,NP}$ = –7.0. As before, these values for $c_{Troubleshoot,NP}$ and $d_{Troubleshoot,NP}$ represent expert expectations and serve as the basis for the prior distributions in the calibration.

Table 1

Conditional Probability Table for *Design*

			Pr (Design = *k*)				
Network Proficiency	θ_{NP}	θ^{**}_{Design}	Novice	Semester 1	Semester 2	Semester 3	Semester 4
Novice	1	–3.8	0.689974	0.252701	0.049162	0.007050	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	0.004740	0.033679	0.192309	0.768525

$\theta^{**}_{Design} \equiv c_{Design,NP} \times \theta_{NP} + d_{Design,NP}$ $\qquad$ $\theta^{**}_{Design} \equiv 2 \times \theta_{NP} + (-5.8)$

Table 2

Conditional Probability Table for *Implement*

Network Proficiency	θ_{NP}	$\theta^{**}_{Implement}$	Pr (Implement = *k*)				
			Novice	Semester 1	Semester 2	Semester 3	Semester 4
Novice	1	–4.2	0.768525	0.192309	0.033679	0.004740	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	0.004740	0.033679	0.192309	0.768525

$\theta^{**}_{Implement} \equiv c_{Implement,NP} \times \theta_{NP} + d_{Implement,NP}$ $\theta^{**}_{Implement} \equiv 2 \times \theta_{NP} + (-6.2)$

Table 3

Conditional Probability Table for *Troubleshoot*

Network Proficiency	θ_{NP}	$\theta^{**}_{Troubleshoot}$	Pr (Troubleshoot = *k*)				
			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	0.500000	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	0.500000

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

To illustrate how these prior estimates reflect 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] Our expectation is that the level of *Design* will be higher than the level of *Implement*, which will be higher than the level of *Troubleshoot*. But there are no mathematical constraints to force *Design* to be higher than *Implement* and *Implement* to be higher than *Troubleshoot*. 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. 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:

$$\theta_c^{**} \equiv c_{c,1} \times \theta_1 + d_{c,1}. \tag{19}$$

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. (19):

$$\theta_{NM}^{**} \equiv c_{NM,NDK} \times \theta_{NDK} + d_{NM,NDK}. \tag{20}$$

Table 4 contains the possible values for *Network Disciplinary Knowledge*, the values for the effective theta obtained with $c_{NM,NDK}$ = 2 and $d_{NM,NDK}$ = –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;

[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.

Table 4

Unstructured Conditional Probability Table for *Network Modeling*

Network Disciplinary Knowledge	θ_{NDK}	θ^{**}_{NM}	Pr (Network Modeling = *k*)				
			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

$\theta^{**}_{NM} \equiv c_{NM,NDK} \times \theta_{NDK} + d_{NM,NDK}$ $\theta^{**}_{NM} \equiv 2 \times \theta_{NDK} + (-8.0)$

that is, 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 $c_{NM,NDK}$ and $d_{NM,NDK}$ 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

Table 5

Corrected Conditional Probability Table for *Network Modeling*

Network Disciplinary Knowledge	θ_{NDK}	θ^{**}_{NM}	Pr (Network Modeling = *k*)				
			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

deferred until later; it should be sufficient for our purposes now to say that compensatory in this context refers to an additive model.

Example from NetPASS. *Network Disciplinary Knowledge* and *Network Modeling* serve as parents for *Network Proficiency* (Figure 1). 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 *Network Modeling* is a secondary factor, with *Network Disciplinary Knowledge* essentially serving as a prerequisite and *Network Modeling* serving 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

$$\theta^{*}_{NP} \equiv c_{NP,baseline} \times \theta_{NDK} + d_{NP,baseline} . \tag{21}$$

Define the effective theta as

$$\theta^{**}_{NP} \equiv \theta^{*}_{NP} + c_{NP,compensatory}[\theta_{NM} - (\theta_{NDK} - 1)] . \tag{22}$$

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 $c_{NP,compensatory}$. 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 $c_{NP,baseline}$ = 2, $d_{NP,baseline}$ = –6.0, and $c_{NP,compensatory}$ = 1 are given in Table 6. The effective theta obtained from eq. (22) 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 itself 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 $c_{NP,baseline}$, $d_{NP,baseline}$, and $c_{NP,compensatory}$ reflect expert

Table 6

Unstructured Conditional Probability Table for *Network Proficiency*

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

$$\theta^{*}_{NP} \equiv c_{NP,baseline} \times \theta_{NDK} + d_{NP,baseline} \qquad \theta^{**}_{NP} \equiv \theta^{*}_{NP} + c_{NP,compensatory}[\theta_{NM} - (\theta_{NDK} - 1)]$$

$$\theta^{*}_{NP} \equiv 2 \times \theta_{NDK} + (-6) \qquad \theta^{**}_{NP} \equiv \theta^{*}_{NP} + (1)[\theta_{NM} - (\theta_{NDK} - 1)]$$

opinions regarding the conditional probability distribution and will serve as the basis for the prior distributions.

Exogenous variable. *Network Modeling*, *Network Proficiency*, *Design*, *Implement*, and *Troubleshoot* were all 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 assess 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 operating on the level of fourth-

Table 7

Corrected Conditional Probability for *Network Proficiency*

Network Disciplinary Knowledge	θ_{NDK}	Network Modeling	θ_{NM}	θ^{**}_{NP}	Pr (Network Proficiency = *k*)				
					Novice	Semester 1	Semester 2	Semester 3	Semester 4
Novice	1	Novice	1	–3	1.0	0	0	0	0
Semester 1	2	Novice	1	–2	0.36788	0.63212	0	0	0
Semester 1	2	Semester 1	2	–1	0.23840	0.76160	0	0	0
Semester 2	3	Novice	1	–1	0.13534	0.43233	0.43233	0	0
Semester 2	3	Semester 1	2	0	0.06487	0.30301	0.63212	0	0
Semester 2	3	Semester 2	3	1	0.03597	0.20243	0.76160	0	0
Semester 3	4	Novice	1	0	0.04979	0.23254	0.48513	0.23254	0
Semester 3	4	Semester 1	2	1	0.02042	0.11491	0.43233	0.43233	0
Semester 3	4	Semester 2	3	2	0.00916	0.05572	0.30301	0.63212	0
Semester 3	4	Semester 3	4	3	0.00495	0.03103	0.20244	0.76159	0
Semester 4	5	Novice	1	1	0.01799	0.10122	0.38080	0.38080	0.11920
Semester 4	5	Semester 1	2	2	0.00669	0.04073	0.22152	0.46212	0.26894
Semester 4	5	Semester 2	3	3	0.00247	0.01551	0.10122	0.38080	0.50000
Semester 4	5	Semester 3	4	4	0.00091	0.00578	0.04073	0.22152	0.73106
Semester 4	5	Semester 4	5	5	0.00034	0.00214	0.01551	0.10122	0.88079

semester students; conversely, it is also possible that students might be quite behind, and 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

Table 8

Probability Table for *Network Disciplinary Knowledge*

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

in **e** to 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. (10), 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, whereas $P(\boldsymbol{\lambda})$ 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 $\boldsymbol{\lambda}$ consists of

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

NetPASS consists of three distinct types of Evidence Models, each corresponding to a different aspect of *Network Proficiency*: Design, Implement, and

[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 $P(\boldsymbol{\lambda})$. See note 12.

Troubleshoot. A pictorial representation of a Design Evidence Model is given in Figure 3. 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 shown as dependent on *DK and DesignE*. As noted above, conditional independence is a key concept in BINs. Achieving conditional independence is required to achieve the computational simplicity of eq. (10). Now 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*. We therefore introduce a context variable, *Design ContextE*, meant to account for this possible (construct irrelevant) dependency. Note that the distribution for *Design ContextE*, the square to the left of the node in Figure 3, 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*

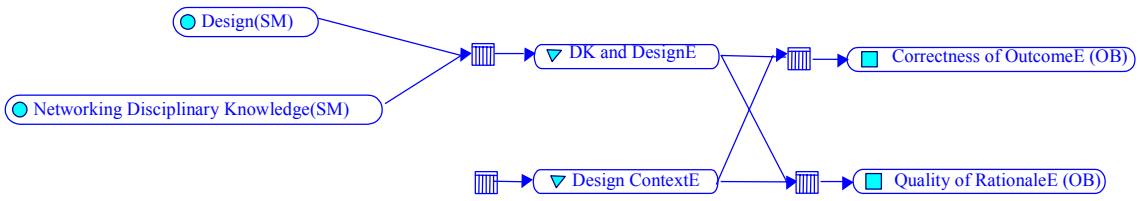

Figure 3. A Design Evidence Model.

RationaleE. The observables are conditionally independent only given both parents. Figure 3 represents a complete Design Evidence Model where the observables are both (a) modeled in relation to Student Model variables, and (b) conditionally independent given their parents. This method of modeling conditional dependencies among related observables has also been implemented in the context of IRT by Bradlow, Wainer, and Wang (1999).

An Implement Evidence Model is depicted in Figure 4. 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 the first 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 the lower part of Figure 4. *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*.

Note that all the observables have *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 5. Its interpretation is analogous to the Implement Evidence Model.

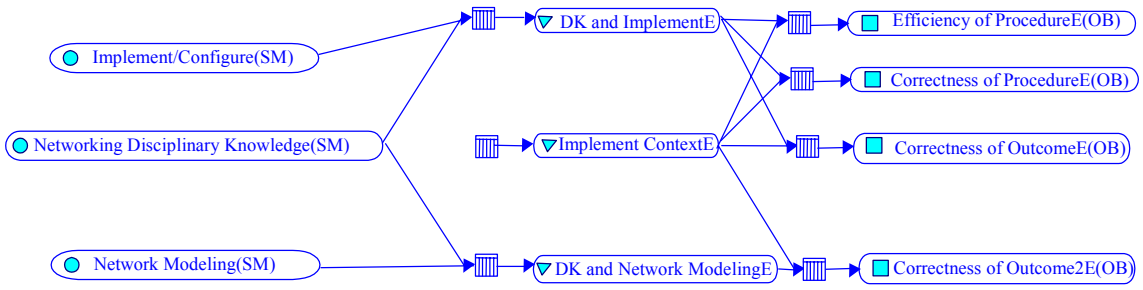

Figure 4. An Implement Evidence Model.

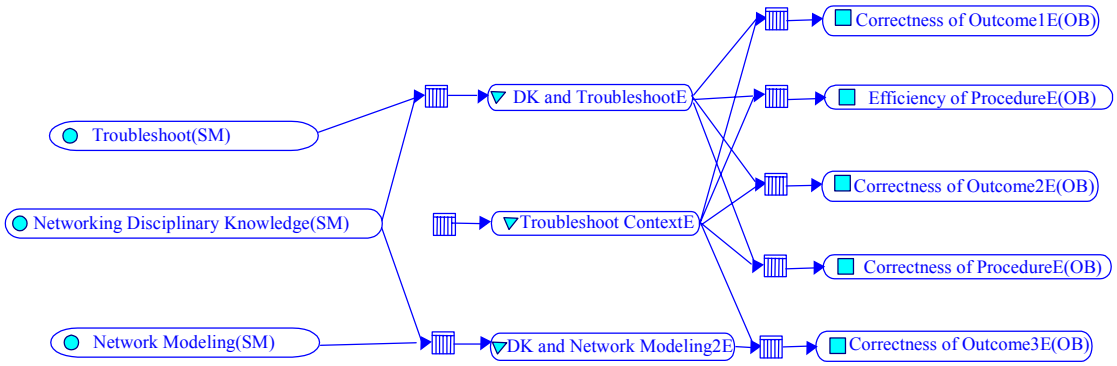

Figure 5. A Troubleshoot Evidence Model.

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. 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. These expectations are effected in the prior distributions for the *c* and *d* parameters associated with these tasks, but evidence in the form of student performances will be able to alter, even reverse, these orderings if warranted.

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 variable) and the appropriate context variable, each localized to the particular instance of the particular Evidence Model.[7]

[7] The names of all of the "instrumental" variables, context variables, and observables in Figures 3, 4, and 5 ended with "E," indicating that these instantiations were the Design Easy, Implement Easy, and Troubleshoot Easy instantiations, respectively.

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 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, that is, 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

$$\theta_c^* \equiv \min(\theta_1, \theta_2) \tag{23}$$

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

$$\theta_c^{**} \equiv u_c(\theta_c^*) = c_{c,\theta_c^*} \times \theta_c^* + d_{c,\theta_c^*} \tag{24}$$

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. (23). 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:

$$\theta_c^{**} \equiv [c_{c,\theta_c^*} \times \theta_c^* + d_{c,\theta_c^*}] + [c_{c,\theta_1} \times (\theta_1 - \theta_c^*)] + [c_{c,\theta_2} \times (\theta_2 - \theta_c^*)], \tag{25}$$

where the contents of the first set of brackets is just that defined in eq. (24), the contents of the second set of brackets captures the impact of how high above the

[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.

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, that is, those above θ_c^*. To fix this, we force the probabilities for the values above θ_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 3. *DK and DesignE* is formed by a leaky conjunction of *Network Disciplinary Knowledge* and *Design*. Thus to calculate the effective theta first instantiate equation (23):

$$\theta^*_{DKandDesign} \equiv \min\left(\theta_{NDK}, \theta_{Design}\right). \tag{26}$$

Next instantiate eq. (25) to calculate the effective theta:

$$\begin{aligned}\theta^{**}_{DKandDesignE} \equiv & [c_{DKandDesignE,\theta^*_{DKandDesign}} \times \theta^*_{DKandDesign} + d_{DKandDesignE,\theta^*_{DKandDesign}}] \\ & + [c_{DKandDesignE,NDK} \times (\theta_{NDK} - \theta^*_{DKandDesign})] \\ & + [c_{DKandDesignE,Design} \times (\theta_{Design} - \theta^*_{DKandDesign})]\end{aligned} \tag{27}$$

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 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. (27) with $c_{DKandDesignE,\theta^*_{DKandDesign}} = 2$, $d_{DKandDesignE,\theta^*_{DKandDesign}} = -6.0$, $c_{DKandDesignE,NDK} = .2$, and $c_{DKandDesignE,Design} = .4$ 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

[9] Let us suppose that $\theta_1 < \theta_2$. In that case, θ_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 θ_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.

modeled by having the value of $c_{DKandDesignE,Design}$ be greater than $c_{DKandDesignE,NDK}$.[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,

$$\begin{aligned}\theta^{**}_{DKandDesignM} \equiv & [c_{DKandDesignM,\theta^{*}_{DKandDesign}} \times \theta^{*}_{DKandDesign} + d_{DKandDesignM,\theta^{*}_{DKandDesign}}] \\ & + [c_{DKandDesignM,NDK} \times (\theta_{NDK} - \theta^{*}_{DKandDesign})] \\ & + [c_{DKandDesignM,Design} \times (\theta_{Design} - \theta^{*}_{DKandDesign})]\end{aligned} \tag{28}$$

and

$$\begin{aligned}\theta^{**}_{DKandDesignH} \equiv & [c_{DKandDesignH,\theta^{*}_{DKandDesign}} \times \theta^{*}_{DKandDesign} + d_{DKandDesignH,\theta^{*}_{DKandDesign}}] \\ & + [c_{DKandDesignH,NDK} \times (\theta_{NDK} - \theta^{*}_{DKandDesign})] \\ & + [c_{DKandDesignH,Design} \times (\theta_{Design} - \theta^{*}_{DKandDesign})]\end{aligned} \tag{29}$$

are the effective thetas for *DK and DesignM* and *DK and DesignH*, respectively.[11] By construction, 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.

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

[11] Note that we need not compute counterparts of eq. (23) 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.

Table 9

Conditional Probability Table for *Network Disciplinary Knowledge and DesignE*

Network Disciplinary Knowledge	Design	Pr (Network Disciplinary Knowledge and DesignE = *k*)				
		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	0.03980	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

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

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

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

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. (23):

$$\theta^{*}_{DKandImplement} \equiv \min\left(\theta_{NDK}, \theta_{Implement}\right) \tag{30}$$

$$\theta^{*}_{DKandNM} \equiv \min\left(\theta_{NDK}, \theta_{NM}\right). \tag{31}$$

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

$$\begin{aligned}\theta^{**}_{DKandImplementE} \equiv & [c_{DKandImplementE,\theta^{*}_{DKandImplement}} \times \theta^{*}_{DKandImplement} + d_{DKandImplementE,\theta^{*}_{DKandImplement}}] \\ & + [c_{DKandImplementE,NDK} \times (\theta_{NDK} - \theta^{*}_{DKandImplement})] \\ & + [c_{DKandImplementE,Implement} \times (\theta_{Implement} - \theta^{*}_{DKandImplement})]\end{aligned} \tag{32}$$

and

$$\begin{aligned}\theta^{**}_{DKandNME} \equiv & [c_{DKandNME,\theta^{*}_{DKandImplement}} \times \theta^{*}_{DKandNM} + d_{DKandNME,\theta^{*}_{DKandNM}}] \\ & + [c_{DKandNME,NDK} \times (\theta_{NDK} - \theta^{*}_{DKandNM})] \\ & + [c_{DKandNME,NM} \times (\theta_{NM} - \theta^{*}_{DKandNM})]\end{aligned} . \tag{33}$$

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

$$\begin{aligned}\theta^{**}_{DKandImplementM} \equiv & [c_{DKandImplementM,\theta^{*}_{DKandImplement}} \times \theta^{*}_{DKandImplement} + d_{DKandImplementM,\theta^{*}_{DKandImplement}}] \\ & + [c_{DKandImplementM,NDK} \times (\theta_{NDK} - \theta^{*}_{DKandImplement})] \\ & + [c_{DKandImplementM,Implement} \times (\theta_{Implement} - \theta^{*}_{DKandImplement})]\end{aligned} \tag{34}$$

and

$$\begin{aligned}\theta^{**}_{DKandNMM} \equiv & [c_{DKandNMM,\theta^{*}_{DKandImplement}} \times \theta^{*}_{DKandNM} + d_{DKandNMM,\theta^{*}_{DKandNM}}] \\ & + [c_{DKandNMM,NDK} \times (\theta_{NDK} - \theta^{*}_{DKandNM})] \\ & + [c_{DKandNMM,NM} \times (\theta_{NM} - \theta^{*}_{DKandNM})]\end{aligned} . \tag{35}$$

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

$$\begin{aligned}\theta^{**}_{DKandImplementH} \equiv & [c_{DKandImplementH,\theta^{*}_{DKandImplement}} \times \theta^{*}_{DKandImplement} + d_{DKandImplementH,\theta^{*}_{DKandImplement}}] \\ & + [c_{DKandImplementH,NDK} \times (\theta_{NDK} - \theta^{*}_{DKandImplement})] \\ & + [c_{DKandImplementH,Implement} \times (\theta_{Implement} - \theta^{*}_{DKandImplement})]\end{aligned} \quad (36)$$

and

$$\begin{aligned}\theta^{**}_{DKandNMM} \equiv & [c_{DKandNMH,\theta^{*}_{DKandImplement}} \times \theta^{*}_{DKandNM} + d_{DKandNMH,\theta^{*}_{DKandNM}}] \\ & + [c_{DKandNMH,NDK} \times (\theta_{NDK} - \theta^{*}_{DKandNM})] \\ & + [c_{DKandNMH,NM} \times (\theta_{NM} - \theta^{*}_{DKandNM})]\end{aligned} \quad . \quad (37)$$

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 equations above 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.

Modeling the *DK and Troubleshoot* and *DK and NM2* variables for the three instantiations of the *Troubleshoot* evidence model follows exactly that of modeling *DK and Implement* and *DK and NM* and hence will not be discussed further. As before, the expected conditional probabilities for these instrumental variables in the *Troubleshoot* evidence models are 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 l^{th} parent variable θ_l is the linear mapping function

$$\theta^{*}_{c,l} \equiv g_{c,l}(\theta_l) = c_{c,l} \times (\theta_l) + d_{c,l} \quad (38)$$

then the combination all L linear mapping functions would be

$$\theta^{**}_{t} \equiv h_t(\theta^{*}_{c,1}, \ldots, \theta^{*}_{c,L}) = \sum_{l=1}^{L} \theta^{*}_{c,l} \; . \quad (39)$$

The particular advantage of this strategy is that the relevance of each of the requisite skills can be assessed (Mislevy et al., 2002). 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, as the sum of the intercepts, but not their individual values, is identified (Mislevy et al., 2002). 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, as discussed below.

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 to the t^{th} observable is

$$\theta^*_{t,1} \equiv g_{t,1}(\theta_1) = c_{t,1} \times (\theta_1) + d_{t,1} \tag{40}$$

and the marginal contribution of θ_2 is given as

$$\theta^*_{t,2} \equiv g_{t,2}(\theta_2) = c_{t,2} \times (\theta_2) + d_{t,2} = c_{t,2} \times (\theta_2)\,. \tag{41}$$

Note that $d_{t,2}$ has been dropped on the right side of eq. (41). 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), $c_{t,2}$ captures all the information and $d_{t,2}$ is unnecessary. To specify the expression for the effective theta, instantiate eq. (39):

$$\theta^{**}_t \equiv h_t(\theta^*_{tl}) = c_{t,1} \times (\theta_1) + c_{t,2} \times (\theta_2) + d_{t,1}\,. \tag{42}$$

We can think of the compensatory relationship that involves a context variable as simply the sum of the marginal values $\theta^*_{t,1}$ and $\theta^*_{t,2}$, 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 et al. (2002).

Examples from NetPASS. Each instance of a Design Evidence Model contains two observables obtained from work products produced in response to a common

[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.

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 t^{th} observable in the Design Easy instance, instantiate eq. (42)

$$\theta_t^{**} = c_{t,DKandDesignE} \times (\theta_{DKandDesignE}) + c_{t,DesignContextE} \times (\theta_{DesignContextE}) + d_{t,DKandDesignE}. \quad (43)$$

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. (43) with $c_{t,DKandDesignE} = 2$, $d_{t,DKandDesignE} = -5.0$, $c_{t,DesignContextE} = .4$, and reflect the opinions and expectations of the SMEs; these values serve to define the prior distributions for the calibration of the model.

Table 10

Conditional Probability Table for the Observables in the Design Easy Evidence Model

					Pr (X = *k*)		
DKandDesignE	$\theta_{DKandDesignE}$	Design ContextE	$\theta_{DesignContextE}$	θ_t^{**}	Low	Medium	High
Novice	1	Low	–1	–1.7	0.802184	0.193320	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	0.193320	0.802184
Semester 4	5	Low	–1	2.3	0.001359	0.067780	0.930862
Semester 4	5	High	1	2.7	0.000611	0.031685	0.967705

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

[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.

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 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 t^{th} observable in the Design Medium instance to be

$$\theta_t^{**} = c_{t,DKandDesignM} \times (\theta_{DKandDesignM}) + c_{t,DesignContextM} \times (\theta_{DesignContextM}) + d_{t,DKandDesignM}. \quad (44)$$

Define the effective theta for the t^{th} observable in the Design Hard instance to be

$$\theta_t^{**} = c_{t,DKandDesignH} \times (\theta_{DKandDesignH}) + c_{t,DesignContextH} \times (\theta_{DesignContextH}) + d_{t,DKandDesignH}. \quad (45)$$

The expected difference in difficulty between the scenarios is captured in the expectation in the intercept terms: for the Design Easy instance, $d_{t,DKandDesignE} = -5.0$; for the Design Medium instance, $d_{t,DKandDesignM} = -6.0$; for the Design Hard instance, $d_{t,DKandDesignH} = -7.0$.[14] The expected strength of association between the observables and (both of) the parent variables remains unchanged; that is, 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 4. 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

[14] For an explanation, see note 4 and the discussion it concerns.

Evidence Model are hypothesized to be equal to those in the corresponding instances of the Design Evidence Model. The same can be said for modeling the observables in the Troubleshoot Evidence Models. For the first three observables in the Implement Easy instance, we define the effective theta as

$$\theta_t^{**} = c_{t,DKandImplementE} \times (\theta_{DKandImplementE}) + c_{t,ImplementContextE} \times (\theta_{ImplementContextE}) + d_{t,DKandImplementE} \quad . \tag{46}$$

For the last observable in the Implement Easy instance, we define the effective theta as

$$\theta_t^{**} = c_{t,DKandNME} \times (\theta_{DKandNME}) + c_{t,ImplementContextE} \times (\theta_{ImplementContextE}) + d_{t,DKandNME} \tag{47}$$

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.

Table 11

Conditional Probability Table for the Observables in the Design Medium Evidence Model

					Pr (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

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

Table 12

Conditional Probability Table for the Observables in the Design Hard Evidence Model

DKandDesignH	$\theta_{DKandDesignH}$	Design ContextH	$\theta_{DesignContextH}$	θ_t^{**}	Pr (X = *k*) Low	Medium	High
Novice	1	Low	–1	–2.7	0.967705	0.031685	0.000611
Novice	1	High	1	–2.3	0.930862	0.067780	0.001359
Semester 1	2	Low	–1	–1.7	0.802184	0.193320	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	0.193320	0.802184

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

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:

$$\theta_t^{**} = c_{t,DKandImplementM} \times (\theta_{DKandImplementM}) + c_{t,ImplementContextM} \times (\theta_{ImplementContextM}) + d_{t,DKandImplementM} \quad (48)$$

$$\theta_t^{**} = c_{t,DKandNMM} \times (\theta_{DKandNMM}) + c_{t,ImplementContextM} \times (\theta_{ImplementContextM}) + d_{t,DKandNMM} \quad (49)$$

and

$$\theta_t^{**} = c_{t,DKandImplementH} \times (\theta_{DKandImplementH}) + c_{t,ImplementContextH} \times (\theta_{ImplementContextH}) + d_{t,DKandImplementH} \quad (50)$$

$$\theta_t^{**} = c_{t,DKandNMH} \times (\theta_{DKandNMH}) + c_{t,ImplementContextH} \times (\theta_{ImplementContextH}) + d_{t,DKandNMH} \quad (51)$$

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. 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. (41)).

Summary

In the preceding sections the variables in the three instances of the three Evidence Models have been quantitatively specified. In terms of the joint probability distribution in eq. (10), 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. (10) have been fully specified except $P(\boldsymbol{\lambda})$ and $P(\boldsymbol{\eta})$. $P(\boldsymbol{\lambda})$ 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.

$P(\boldsymbol{\eta})$ 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. 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 d to be distributed normally with mean defined by expert expectation and variance of 1. Similarly, we define the priors for all coefficients c 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.

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. All terms on the right side of eq. (10) 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, condition on 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 sensible (at least more sensible 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:

$$P(\boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\eta}, \boldsymbol{\lambda} \mid \mathbf{X}) \propto P(\mathbf{X} \mid \boldsymbol{\theta}, \boldsymbol{\pi}) \times P(\boldsymbol{\theta} \mid \boldsymbol{\lambda}) \times P(\boldsymbol{\lambda}) \times P(\boldsymbol{\pi} \mid \boldsymbol{\eta}) \times P(\boldsymbol{\eta}). \quad (42)$$

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 ($\boldsymbol{\pi}$), and the task parameters ($\boldsymbol{\eta}$, which define the conditional probabilities of the observables).[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, Richardson, & Spiegelhalter, 1996a). 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 popular method for constructing such a chain is via the Metropolis sampler (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953). For a complete discussion of this and other MCMC techniques, see Brooks (1998) and Gilks, Richardson, and Spiegelhalter (1996b).

Empirical Analysis

The data set consisted of 216 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 over 28 values for each of the observables. The computer

[15] Note the similarity between eq. (42), the posterior distribution, and eq. (10), the joint distribution. The difference is that in the joint distribution, $\mathbf{X}$ is a random variable, whereas in the posterior distributions for the parameters, $\mathbf{X}$ is fixed at the values that are actually observed.

program WinBUGS 1.4 (Spiegelhalter, Thomas, Best, & Lunn, 2003) 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; WinBUGS' convergence diagnostics (Brooks & Gelman, 1998; Gelman & Rubin, 1992) were computed from these multiple chains to determine chain length and number of "burn-in" cycles. Analysis of convergence consisted of monitoring the overestimate and the underestimate of the true posterior variance as 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 were 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 posterior). Second, though there exist autocorrelations among the values *within* each chain, there is no correlation among the values *between* parallel chains; that is, the chains are independent. Pooling the values from parallel multiple chains serves to mitigate the impact of serial dependence (Gelman, 1996). Finally, the use of multiple chains with overdispersed starting points 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 and Discussion

General results. A question of immediate interest concerns the impact of the data on the posterior distributions for the parameters that define the conditional probability distributions. The average posterior standard deviation was 0.73 with a standard deviation of 0.17. Figure 6 displays the distribution of posterior standard deviations. 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. Recall that the prior standard deviation for all parameters is 1.0, and thus the percent increase in precision is merely a 1-1 transformation of the posterior standard deviation. As it is a general metric for summarizing the impact of data in a Bayesian model, we will continue to discuss the results in terms of percent increase in precision. The distribution of percent increase in precision for most of the parameters (three parameters were excluded from this analysis, as discussed below) is displayed in Figure 7. The average increase in precision is 118.15 with a standard deviation of 111.81. Select parameters will be discussed below in further detail; overall, most parameters showed reasonable increases in precision. The average percent increase in precision for the parameters as listed by the portion of the model is given in Table 13.

For the most part, there were mild increases in precision for the variables that define the conditional distributions of the latent variables, that is, the variables in the Student Model, and the instrumental variables in the various instantiations of the Evidence Models. Larger increases in posterior precision were observed in the

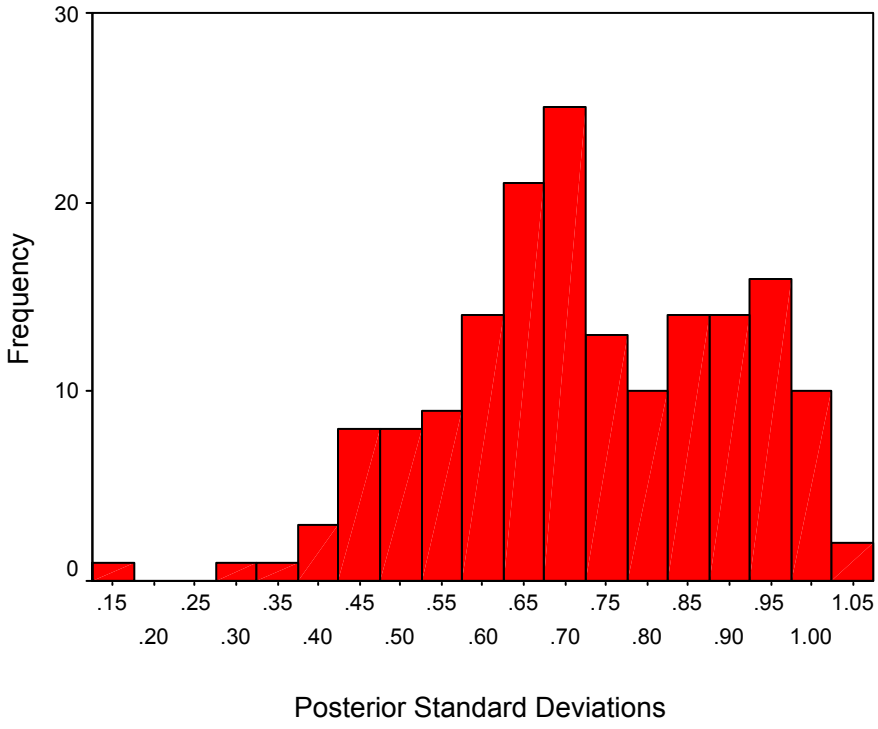

Figure 6. Histogram of posterior standard deviations.

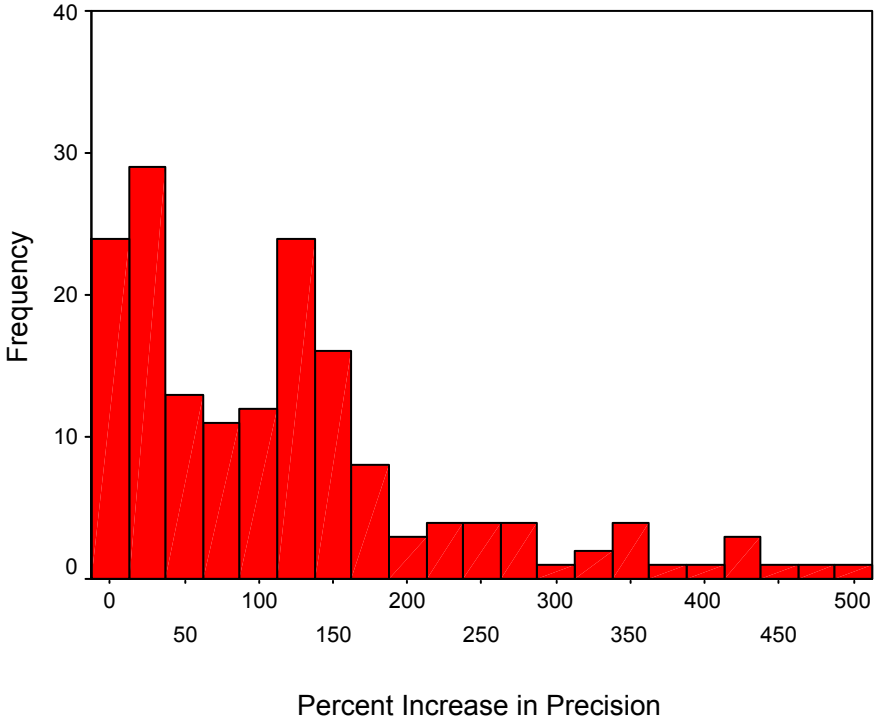

Figure 7. Histogram of percent increase in precision.

Table 13

Average Percent Increase in Precision for Parameters That Define Conditional Distributions by Model Portion

Model fragment	Average increase in precision
Student Model	54.30
Latent variables in Design Evidence Models	82.98
Observable variables in Design Evidence Models	216.05
Latent variables in Implement Evidence Models	85.04
Observable variables in Implement Evidence Models	254.75
Latent variables in Troubleshoot Evidence Models	85.99
Observable variables in Troubleshoot Evidence Models	147.43

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) informs 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. The variables that define the conditional distributions in the Student Model are most removed from the observables, and therefore, overall, show the smallest increase in precision.

Selected parameters. Two parameters, the intercepts in the effective theta equations for *NDKandNMM* and *NDKandNMH,* showed *decreases* in precision (–12.65 and –9.12, respectively). It appears as though two factors at work here. First, the data do not inform on intercepts as well as coefficients (mean percent increase in precision for intercepts is 21.860530; mean percent increase in precision for coefficients, excepting the three highest, is 162.910746). In addition, recall that only one observable in each Implement Evidence Model instantiation informs on the *NDKandNM* variable; thus, it is not surprising that parameters associated with these variables are not as well estimated. Similarly, intercept parameters for other instrumental variables on which only one observable is dependent showed small increases in precision. The three parameters excluded from Figure 7 are those with the largest increases in precision. These parameters were the coefficient for *Implement ContextM* for the third observable in the Implement Medium Evidence Model, the coefficient for *DK and TroubleshootM* for the fourth variable in the Troubleshoot Medium Evidence Model, and the coefficient for *Implement ContextE* for the third observable in the Implement Easy Evidence Model. The values for the percent increase in precision are 814.94, 1020.05, and 3820.94, respectively. Whether these values are appropriately due to greater-than-average amounts of information in the data or are artifacts of parameterization or estimation cannot be stated (although convergence indices and posterior distributions did not indicate abnormalities). These parameters were therefore excluded from the previous analysis, and will be the focus of future work with larger samples.

Table 14 contains the prior means and summaries of the posterior distributions for the parameters in the effective theta equations for the first and third observables in the Troubleshoot Medium Evidence Model. Note that because the observables come from the same Evidence Model instantiation, their priors were identical. The

Table 14

Posterior Results of Selected Parameters

Parameter	Prior mean	Posterior mean	Posterior *SD*	% Increase in precision
$c_{1,DKandTroubleshootM}$	2	2.029	0.440	415.82
$d_{1,DKandTroubleshootM}$	–6	–6.778	0.793	59.18
$c_{1,TroubleshootContextM}$	0.4	0.932	0.610	168.92
$c_{3,DKandTroubleshootM}$	2	1.923	0.450	393.17
$d_{3,DKandTroubleshootM}$	–6	–4.865	0.855	36.86
$c_{3,TroubleshootContextM}$	0.4	0.759	0.536	248.07

prior distributions for $c_{1,DKandTroubleshootM}$ and $c_{3,DKandTroubleshootM}$ were centered at 2. Both posteriors have means close to 2 and the large increases in precision indicate the data (a) conform to SME expectation and (b) inform on the parameters considerably. Similarly, the posterior distributions for $c_{1,TroubleshootContextM}$ and $c_{3,TroubleshootContextM}$ also indicate that, for both observables, there was a stronger context effect involved than anticipated. On the other hand, though the priors for the intercept parameters were the same, the posteriors are greatly different. The intercept for the first observable is lower than the mean of the prior. Conversely, the intercept for the second observable is higher than the mean of the prior (and the mean of the posterior for the first intercept). The interpretation of this result is that, though they were expected to be of equal difficulty, the first observable is considerably more challenging than the second as can be seen by the number of Low, Medium, and High responses to these variables (Figure 8).

These sets of parameters define the conditional probability distributions for the two observable variables considered here. The prior conditional probability distribution is contained in Table 11. The posterior conditional distributions (based on the parameters' posterior means) for the two observables are given in Tables 15 and 16, where it is clearly seen that examinees are more likely to perform well on the third observable, as compared to the first.

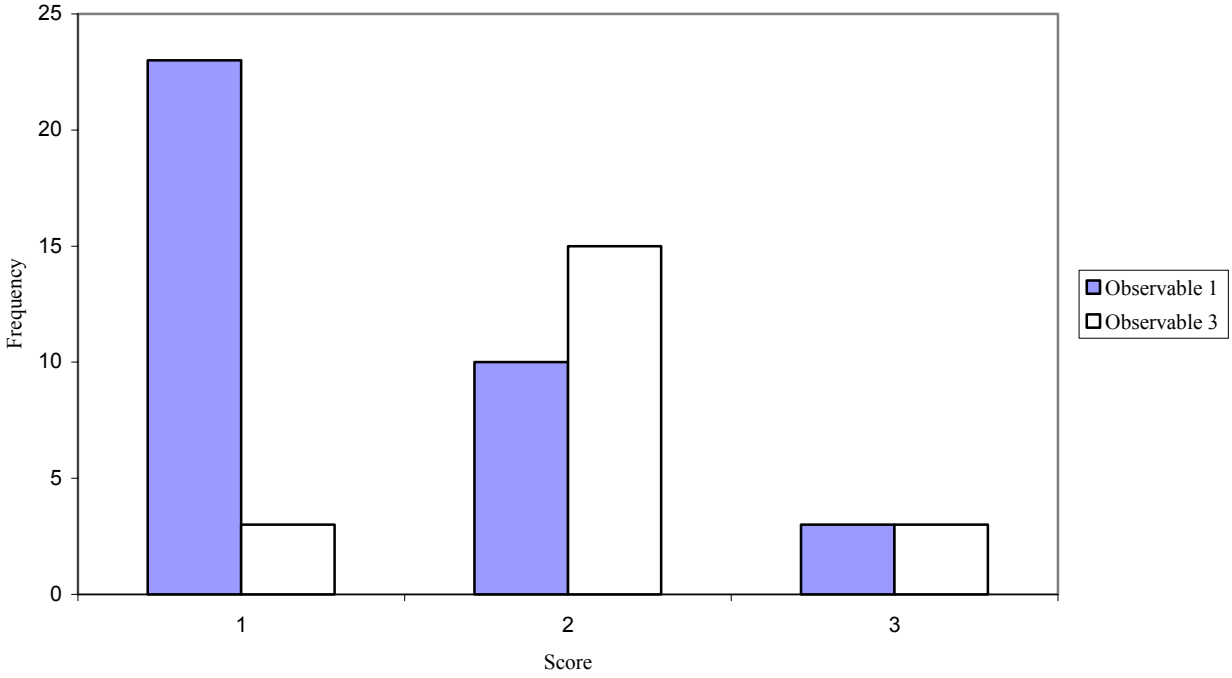

Figure 8. Frequencies of low, medium, and high responses to two observables from the Troubleshoot Medium Evidence Model.

Table 15

Posterior Conditional Probability Table for the First Observable in the Troubleshoot Medium Evidence Model

		Pr (X = *k*)		
DKandTrbM	Trb ContextM	Low	Medium	High
Novice	Low	0.975	0.024	0.000
Novice	High	0.860	0.137	0.003
Semester 1	Low	0.839	0.157	0.003
Semester 1	High	0.448	0.531	0.022
Semester 2	Low	0.407	0.567	0.026
Semester 2	High	0.096	0.757	0.147
Semester 3	Low	0.082	0.748	0.169
Semester 3	High	0.014	0.419	0.567
Semester 4	Low	0.011	0.381	0.607
Semester 4	High	0.002	0.089	0.909

Table 16

Posterior Conditional Probability Table for the Third Observable in the Troubleshoot Medium Evidence Model

		Pr (X = *k*)		
DKandTrbM	Trb ContextM	Low	Medium	High
Novice	Low	0.846	0.151	0.003
Novice	High	0.546	0.439	0.015
Semester 1	Low	0.445	0.533	0.022
Semester 1	High	0.149	0.756	0.094
Semester 2	Low	0.105	0.760	0.135
Semester 2	High	0.025	0.558	0.417
Semester 3	Low	0.017	0.466	0.517
Semester 3	High	0.004	0.166	0.830
Semester 4	Low	0.002	0.118	0.880
Semester 4	High	0.001	0.029	0.971

This example encapsulates the estimation of the conditional probability tables: The conditional distributions are parsimoniously parameterized, and prior beliefs regarding the psychometric properties of the observable variables based on expert expectations are revised in light of the information that pilot data bring to bear.

Of particular interest are the parameters associated with the adjustments to the conjunctions (e.g., $c_{DKandDesignE,NDK}$ and $c_{DKandDesignE,Design}$ in eq. (27)). If the posterior distributions indicate that these parameters are small (recall they are bounded below at zero), the inference is made that such adjustments to the conjunction may constitute overfitting and may be dropped from the model without great loss. However, the average posterior mean for these parameters is .928 (the minimum value was .744) indicating that such adjustments contribute to the model. More general strategies for assessing model fit will be discussed below.

Examinee parameters. The preceding discussion has focused exclusively on the parameters that define the conditional probability distributions in the Student Model and the Evidence Models. In addition, the Student Model variables themselves were monitored for all examinees. Two research questions surrounding examinee parameters are (a) the possibility that there is more information in the data regarding examinee parameters than the parameters that define the conditional

distributions, as has been observed in other calibration studies (Mislevy et al., 2002), and (b) whether calibration studies can support inferences about examinees.

Though discrete, the impact of the data on the Student Model variables may still be discussed in terms of percent increase in precision. An assumption of exchangeability implies the prior distributions for all examinees are identical. Prior standard deviations and average percent increase in precision for the Student Model variables and the observed percent increase in precision for selected examinees are given in Table 17.

The average percent increase in precision indicates that the data inform on the Student Model parameters (Table 17) less than they do on the parameters that define the conditional probability distributions in either the Student Model or the Evidence Models (Table 13). It is not surprising that there is the least amount of information regarding *Network Proficiency* as it is most removed from the data. Though *Network Disciplinary Knowledge and Network Modeling* are parents of *Network Proficiency* (Figure 1) and seemingly more removed from the observables, they appear in the Evidence Models (Figures 3–5). Indeed, we might expect to see large increases in precision for *Network Disciplinary Knowledge* and *Network Modeling* for this reason. This is partially borne out in the case of *Network Modeling*. The low average percent increase in precision for *Network Disciplinary Knowledge* seems to indicate that there is not a lot of information in the data about *Network Disciplinary Knowledge*, relative to the other Student Model parameters. However, the posterior standard deviation

Table 17

Summary of Prior and Posterior Results for Student Model Variables and Results for Selected Examinees

Variable	Prior *SD*	Average posterior *SD*	Average % increase in precision	% Increase for selected examinees		
				A	B	C
Network Disciplinary Knowledge	0.887	0.833	25.534	882.944	77.196	–10.418
Network Modeling	1.094	0.951	48.147	951.463	107.410	10.732
Network Proficiency	1.077	1.059	10.269	54.338	20.918	–20.648
Design	1.258	1.126	34.590	980.018	146.476	0.270
Implement	1.262	1.041	52.377	99.964	139.759	63.728
Troubleshoot	1.241	1.048	47.535	380.843	53.932	16.801

for *Network Disciplinary Knowledge* is smaller than that of the other variables.[16] A large average increase in precision is not observed because the *prior* standard deviation for *Network Disciplinary Knowledge* is also considerably smaller than that of the other variables; we do not observe a large increase in precision for *Network Disciplinary Knowledge* because the expert expectation regarding the variability of *Network Disciplinary Knowledge* was closer to what the data suggest, compared to the other Student Model variables.

Turning to the individual examinees, the data clearly inform most on examinee A and least on examinee C. This is not a surprising result, as examinee A completed 7 of the 9 tasks resulting in 28 observed data points, whereas examinee B completed 6 tasks resulting in 19 data points, and examinee C completed only 1 task, resulting in 4 data points. The lone task that examinee C completed was the Implement Easy task and therefore the largest increase in precision for examinee C is for *Implement*. Regarding the plausibility of inferences about examinees, we caution against interpreting the results for examinees who have completed only a few tasks, particularly regarding variables for which little or no evidence is observed (i.e., *Design* and *Troubleshoot* for examinee C). However, considerable increases in precision were observed for examinee A, and inferences regarding such an examinee's proficiency would be more warranted. For example, Table 18 gives the prior and posterior probabilities of *Design* for examinees A, B, and C.[17] These probabilities are graphed in Figure 9. The posterior for examinee C is much closer to the prior than the posteriors of the other examinees, reflecting the relative lack of knowledge regarding examinee C. The posterior distribution for examinee B indicates a high concentration in Semester 2 and Semester 3 relative to the prior; that is, the posterior probability for examinee B is higher than the prior for Semester 2 and Semester 3 (and lower for Novice, Semester 1, and Semester 4). As expected, the posterior distribution for examinee A is much narrower than either of the other posteriors, indicating a higher level of precision regarding examinee A. Thus while we can say very little about examinee C, more can be said about examinee B, and even more can be said about examinee A.

---

[16] Similarly, since *Network Modeling* appears in six of the Evidence Model instantiations, its posterior standard deviation is lower than those of the other Student Model variables (except *Network Disciplinary Knowledge*).

[17] The prior was calculated by compiling the distribution with all conditional probability parameters set to the values defined by expert expectation.

Table 18

Prior and Posterior Density Functions of *Design* for Examinees A, B, and C

	Prior	A	B	C
Novice	0.114	0	0.002	0.211
Semester 1	0.187	0	0.061	0.281
Semester 2	0.271	0.005	0.500	0.251
Semester 3	0.245	0.152	0.309	0.154
Semester 4	0.184	0.843	0.127	0.104

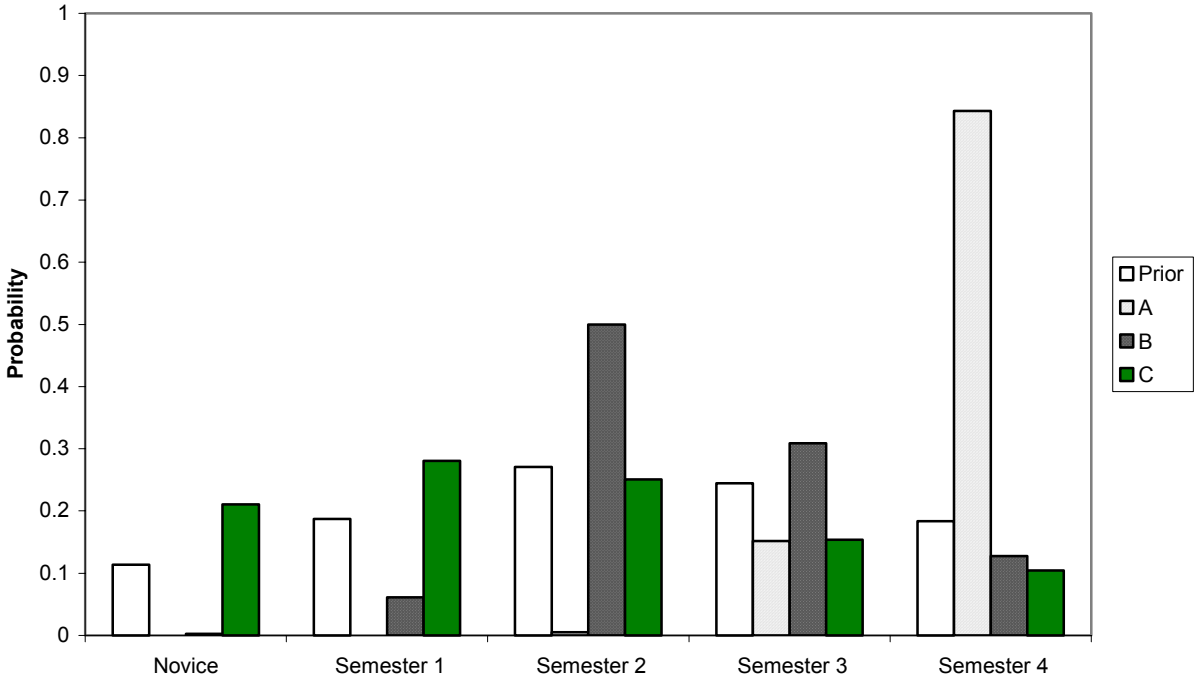

Figure 9. Prior and posterior probability densities of *Design* for Examinees A, B, and C.

The association between the number of data points and percent increase in precision observed among examinees A, B, and C bears out throughout the data. Table 19 gives the correlations between the number of observed values and the percent increase in precision for each of the Student Model variables; for five of the six Student Model variables, there is a statistically significant positive correlation

Table 19

Correlations Between Number of Data Points and % Increase in Precision for Student Model Variables

Correlation between number of data points and % increase in precision					
Network Disciplinary Knowledge	Network Modeling	Network Proficiency	Design	Implement	Troubleshoot
.619*	.596*	–.009	.451*	.326*	.489*

* Significant at the .01 level.

between the number of data points and percent increase in precision. This implies that, provided the examinees engage in many, if not all, of the tasks, reporting results for individual students may be justified, especially for low-stakes purposes, even without large calibration samples. Though results would lean heavily on the expert-posited structure and initial estimates, changes from the prior to the posterior distributions can reflect the relative difficulty of the tasks and the contribution of Student Model variables.

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), Gilks et al. (1996b), and Spiegelhalter, Best, Carlin, and van der Linde. (2002). Many promising techniques involve the use of replicated data distributions (e.g., Mislevy et al., 2002). 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. Likewise, our interests lie in comparing the existing model to those that reduce the number of instrumental parameters or exclude adjustments to conjunctions. 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

Andrich, D. (1982). An extension of the Rasch model for ratings providing both location and dispersion parameters. *Psychometrika, 47*, 105-113.

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

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64,* 153-168.

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

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

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

Formann, A. K. (1985). Constrained latent class models: Theory and applications. *The British Journal of Mathematical and Statistical Psychology, 38,* 87-111.

Formann, A. K., & Kohlmann, T. (1998). Structural latent class models. *Sociological Methods and Research, 26,* 530-565.

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., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis.* London: Chapman and Hall.

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

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.). (1996b). *Markov chain Monte Carlo in practice*. London: Chapman and Hall.

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

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.

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.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equations of states calculations for fast computing machines. *Journal of Chemical Physics, 21,* 1087-1091.

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

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

Mislevy, R. J., Senturk, D., Almond, R. G., Dibello, L. V., Jenkins, F., Steinberg, L. S., et al. (2002). *Modeling conditional probabilities in complex educational assessments* (CSE Tech. Rep. No. 580). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). *On the structure of educational assessments* (CSE Tech. Rep. No. 597). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Patz, R. J., Junker, B. W., Johnson, M. S., & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large scale educational assessment data. *Journal of Educational and Behavioral Statistics, 27,* 341-384.

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

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

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph No. 17*, *34*(No. 4, Part 2).

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

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B (Statistical Methodology), 64,* 583-639.

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

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Lunn, D. (2003). *WinBUGS version 1.4: User manual*. Cambridge Medical Research Council Biostatistics Unit. Available 3 November 2003 from http://www.mrc-bsu.cam.ac.uk/bugs/

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

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