

Running head: SPECIFYING AND REFINING A MEASUREMENT MODEL

Specifying and Refining a Measurement Model for a Computer Based Interactive Assessment

Roy Levy and Robert J. Mislevy

University of Maryland, College Park

Either author may be contacted at

1230 Benjamin Building
University of Maryland
College Park, MD, USA 20742

To appear in *The International Journal of Testing*

Abstract

The challenges of modeling students' performance in computer based interactive assessments include accounting for multiple aspects of knowledge and skill that arise in different situations and the conditional dependencies among multiple aspects of performance. This paper describes a Bayesian approach to modeling and estimating cognitive models in such situations, both in terms of 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 procedures. This process is illustrated in the context of NetPASS, a computer based interactive assessment in the domain of computer networking. We describe a parameterization of the relationships in NetPASS via an ordered polytomous item response model and detail the updating of the model with observed data via Bayesian statistical procedures ultimately being provided by Markov chain Monte Carlo estimation.

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

Specifying and Refining a Measurement Model for a Computer Based Interactive Assessment

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 such as a scan-tron sheet with circles filled in, essays to be scored by raters, and portfolios. While methods for drawing inferences from examinees' work products to their knowledge, skills, and abilities exist for the more popular assessment instruments, new and innovative assessment instruments are often left to develop inferential procedures 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 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 to 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 including inference in educational measurement; moreover, probability based reasoning 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 Design (ECD; Mislevy, Steinberg, & Almond, 2003), referring in particular to student, evidence, and task models of the conceptual assessment framework or CAF (for an overview and the development of the CAF with regard to NetPASS, see Williamson et al., in submission).

Specifically, we will discuss the development of the probabilistic model for NetPASS, a measurement device to be utilized in assessing cognitive development primarily of students in the third semester of Cisco Networking Academy Program's sequence of courses on computer networking. While the particulars of the NetPASS implementation will be described in some detail, the process of instrument and model development can be reinstated 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 $P(X | \theta)$, 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 $P(\theta | X)$ by Bayes theorem.

While 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; Brooks, 1998; Jensen, 1996, 2001; Pearl, 1988) 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 Almond & Mislevy, 1999; 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 representation (of which Figure 1, depicting the NetPASS student model, is an example) of a joint probability distribution over a set of random variables. The variables are represented by ellipses and referred to as nodes. The directed edges (arrows) indicate 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. The absence of an edge between two variables indicates a conditional independence between them, given variables on the path(s) between them. For each variable, there is a 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 prior probability distributions.

[[Insert Figure 1 About Here]]

We can define fragments of BINs in terms of a BIN for student model variables and a BIN for the 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 variables Y in this connection.

The Probability Framework

Gelman, Carlin, Stern, and Rubin (1995) define the first step in conducting a Bayesian analysis as setting up a full probability model, specifically, a joint distribution of all quantities, observable and unobservable. Furthermore they note “the model should be consistent with knowledge about the underlying scientific problem and the data collection process” (p. 3). In assessment, this “knowledge” is knowledge about the domain of interest, specifying the (1) targeted knowledge, skills and abilities, (2) ways in which such knowledge, skills, and abilities are demonstrated in performance, and (3) characteristics of situations that provide the opportunity to observe such performance. The student model, evidence models, and task models provide this information (Williamson et al., in submission).

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

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

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

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

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

$$\mathbf{X}_j = (X_{j1}, \dots, X_{jM}) \quad (3)$$

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 (DeMark & Behrens, in submission), 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 (Lindley & Smith, 1972) entails a common prior distribution, (i.e., before any responses to tasks are observed the student model is in the same state for all examinees). Beliefs about the expected values and associations among the student model variables are expressed through the structure of the model and higher level hyperparameters $\boldsymbol{\lambda}$. Thus, for all examinees,

$$\boldsymbol{\theta}_i \sim P(\boldsymbol{\theta}_i | \boldsymbol{\lambda}). \quad (4)$$

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 $P(\boldsymbol{\theta}_i | \boldsymbol{\lambda})$ 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, θ_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

$$\pi_{mj} = (\pi_{mj1}, \pi_{mj2}, \dots, \pi_{mjK}), \quad (5)$$

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 , the conditional probability of X_{imj} given θ_i . More formally, if

$$\pi_{mjk} = P(X_{imj} = x_{imjk} | \theta_i), \quad (6)$$

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

$$X_{imj} \sim P(X_{imj} | \theta_i, \pi_{mj}). \quad (7)$$

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

$$\pi_{mj} \sim P(\pi_{mj} | \eta_{mj}), \quad (8)$$

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 π ; the complete set of parameters that define those distributions is denoted η .

The joint probability of all parameters can be expressed as

$$P(\lambda, \eta, \theta, \pi, \mathbf{X}) = P(\lambda) \times P(\eta | \lambda) \times P(\theta | \lambda, \eta) \times P(\pi | \lambda, \eta, \theta) \times P(\mathbf{X} | \lambda, \eta, \theta, \pi). \quad (9)$$

This expression can be simplified in light of additional knowledge and assumptions we bring to the assessment context by taking advantage of the conditional independence relationships implied in eqs. (4) – (8), yielding:

$$P(\lambda, \eta, \theta, \pi, \mathbf{X}) = P(\lambda) \times P(\theta | \lambda) \times P(\eta) \times P(\pi | \eta) \times P(\mathbf{X} | \theta, \pi). \quad (10)$$

In setting up the full model, the 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, the need arises 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 et al., 1993). This procedure

quickly becomes unwieldy as the number of levels of the parent(s) or child increases. We therefore seek a more efficient way to model the conditional probabilities. 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 one 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 $k=2, \dots, K$, where b_{jk} is the location parameter associated with separating the k^{th} from the $k-1^{\text{th}}$ category and θ_i is the latent trait for examinee i . The probability of response being in the k^{th} category is

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

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 an ordered polytomous variable as conditional on a continuous variable (Samejima, 1969), the current application follows the use of the logistic function in modeling ordered polytomous variables as dependent on a discrete variable (see e.g., Formann, 1985; Formann & Kohlmann, 1998).

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 corresponding to the proficiency level in terms of the CNAP sequence of courses (Novice, Semester 1, Semester 2, Semester 3, Semester 4).

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 to produce one variable that will serve in the unidimensional GRM; this

variable is an *effective theta*, denoted as θ^{**} . For example, suppose a child variable is modeled as conditional on a single parent variable, θ_1 . Assuming the levels of θ_1 are roughly equally spaced apart, we code the values of θ_1 accordingly and define the effective theta for the child variable via a linear function

$$\theta^{**} \equiv c \times \theta_1 + d. \quad (13)$$

In IRT models (e.g., eq. (11); Hambleton & Swaminathan, 1985), the conditional probabilities of response are determined by theta and the item parameters a_{mj} and \mathbf{b}_{mj} .¹ In fixing these parameters the conditional probabilities are then a function of the effective theta, which itself is a function of the parent variable(s) and the c and d parameters. Coefficients (c) and intercepts (d) in the calculation of the effective theta are akin to scale (a) and location (b) 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 (Hambleton & Swaminathan, 1985). 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. Intuitively, the effective theta may be thought of as the combination of the parent variable θ_1 and the features of the conditional distribution, represented by c and d . 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 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). 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.), SMEs may not feel comfortable specifying a complete conditional probability table. Second, unidimensional IRT models are quite popular in the psychometric community and now the problem is on a scale familiar to experts in educational measurement. Thus, they may feel more comfortable with capturing and modeling knowledge elicited from the SMEs. For example, if SMEs 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.

The Application of the Effective Theta Method to the GRM

When using the effective theta method and the GRM to model observed responses (which can take on any of three values), we set $a = 1$ and $\mathbf{b} = (-2, +2)$. When using the effective theta method and the GRM to model values of latent variables (which can take on any of five values), 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.

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. Recall that 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.

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 $P(\boldsymbol{\theta} | \boldsymbol{\lambda})$. Several relationships will be discussed, each followed by examples as they appear in NetPASS. Where possible, the subscript identifying the variable will be abbreviated, (i.e., θ_{NDK} refers to *Network Disciplinary Knowledge*, θ_{NM} refers to *Network Modeling*, and θ_{NP} refers to *Network Proficiency*).

Direct Dependence

In direct dependence relationships, 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:

$$\theta_c^{**} \equiv c_{c,1} \times \theta_1 + d_{c,1} \tag{14}$$

where θ_c^{**} is the effective theta to be used for the distribution of the child, and θ_1 is the parent variable.²

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*, *Implement*, and *Troubleshoot* instantiate eq. (14):

$$\theta_{Design}^{**} \equiv c_{Design,NP} \times \theta_{NP} + d_{Design,NP}, \tag{15}$$

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

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

Effective thetas for *Design* are 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, as are the resulting conditional distributions for *Design* (in regular typeface). Table 1 also displays effective thetas for *Implement* (in italic typeface), as well as the resulting conditional distributions, calculated for all possible values of *Network Proficiency* with $c_{Implement,NP} = 2$ and $d_{Implement,NP} = -6.2$. Table 1 also contains the effective thetas for *Troubleshoot* and conditional probabilities obtained with $c_{Troubleshoot,NP} = 2$ and $d_{Troubleshoot,NP} = -7.0$ (in bold typeface). The values for the c and d parameters were chosen because when the resulting effective thetas are entered into the GRM to produce the conditional probability distributions, the resulting distributions approximately matched the opinions and expectations of SMEs for *Design*, *Implement*, and *Troubleshoot*. Values of the c and d parameters will eventually be estimated. Because values used to produce the distributions in Table 1 result in the conditional distributions experts expect, prior distributions for these parameters will be based on these values.

[[Insert Table 1 About Here]]

To illustrate how these prior estimates reflect expert expectations, compare the values for *Design*, *Implement*, and *Troubleshoot*, in Table 1; 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*.³ The 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:

$$\theta_c^{**} \equiv c_{c,1} \times \theta_1 + d_{c,1}. \quad (18)$$

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 zero 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. (18):

$$\theta_{NM}^{**} \equiv c_{NM,NDK} \times \theta_{NDK} + d_{NM,NDK} . \tag{19}$$

The probabilities that result from the GRM do 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 zero and renormalizing such that the conditional distributions (i.e., the rows in the table), sum to one.

These corrected probabilities, based on parameter values of $c_{NM,NDK} = 2$ and $d_{NM,NDK} = -8.0$ are given in Table 2. 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.

[[Insert Table 2 About Here]]

Baseline-Ceiling Relationships

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

Example in NetPASS. *Network Disciplinary Knowledge* and *Network Modeling* serve as parents for *Network Proficiency* (Figure 1). Discussions with SMEs revealed that *Network*

Proficiency cannot be higher than *Network Disciplinary Knowledge* and that *Network Proficiency* is expected to be higher than *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*, essentially serving as a prerequisite, and that *Network Modeling* is a secondary factor, 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

$$\theta_{NP}^* \equiv c_{NP,baseline} \times \theta_{NDK} + d_{NP,baseline} . \quad (20)$$

Define the effective theta as

$$\theta_{NP}^{**} \equiv \theta_{NP}^* + c_{NP,compensatory} [\theta_{NM} - (\theta_{NDK} - 1)] . \quad (21)$$

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 2), the contribution is zero. 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.

Logically, there are 25 combinations of the parent variables (as each can take on any of five states). However, several of these combinations are impossible, as *Network Modeling* cannot exceed *Network Disciplinary Knowledge*. The remaining 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 3. The effective theta obtained

from eq. (21) is then entered into the GRM to obtain the conditional probability distribution for *Network Proficiency*. 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 zero and renormalizing the probabilities. The corrected probability distributions are given in Table 3. Again, the values of $c_{NP,baseline}$, $d_{NP,baseline}$, and $c_{NP,compensatory}$ reflect expert opinions regarding the conditional probability distribution and will serve as the basis for the prior distributions.

[[Insert Table 3 About Here]]

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 that are operating on the level of fourth semester students; conversely, it is also possible that students might be quite behind, it is even possible that some might be operating at the level of a first semester student or even that of a complete novice. Using an effective theta value of .6 results in an appropriate distribution, which is given in Table 4.

[[Insert Table 4 About Here]]

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, \mathbf{e} , is defined with pseudocounts of examinees. For example, with \mathbf{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 \mathbf{e} . In essence, the values in \mathbf{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 4. Since we desire to have vague prior distributions, we define the pseudocounts accordingly. Operationally, this is accomplished by setting the values in \mathbf{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 4 but on a sample of size 10 (Spiegelhalter et al., 1993).

Summary

In the preceding sections 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 $P(\boldsymbol{\theta} | \boldsymbol{\lambda})$ and hinted at the $P(\boldsymbol{\lambda})$ terms.⁴ $P(\boldsymbol{\theta} | \boldsymbol{\lambda})$ refers to the distribution of the student model variables, while $P(\boldsymbol{\lambda})$ refers to the distribution of the parameters that define the distribution of the student model variables. In terms of the effective theta method, $\boldsymbol{\theta}$ are the student model variables themselves and $\boldsymbol{\lambda}$ consists of

- The various c , and d parameters used to define the distributions of *Network Modeling*, *Network Proficiency*, *Design*, *Implement*, and *Troubleshoot*
- \mathbf{e} parameters used to define the distribution of *Network Disciplinary Knowledge*

In order to enact a fully Bayesian model, distributions of 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 Troubleshoot*. A pictorial representation of a Design evidence model is given in Figure 2. 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 (see Williamson et al., in submission, for the grounding of these and other observables used in NetPASS).

[[Insert Figure 2 About Here]]

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). As of 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*. The consequences of

incorrectly assuming conditional independence can be deleterious in estimating the values of variables and the precision of the estimates (Mislevy & Patz, 1995). We therefore introduce a context variable, *Design ContextE*, meant to account for this possible (construct irrelevant) dependency due to the common task. Note that the distribution for *Design ContextE*, the square to the left of the node in Figure 2, has no directed edges flowing into it meaning that the distribution of *Design ContextE* is not a conditional distribution; *Design ContextE* is an exogenous variable. The two parents, *DK and DesignE* and *Design ContextE*, represent distinct and independent portions of the dependency between *Correctness of OutcomeE* and *Quality of RationaleE*. The observables are conditionally independent only given both parents. Figure 2 represents a complete Design evidence model where the observables are both (1) modeled in relation to student model variables, and (2) 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 3. 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. Specifically, the work products examinees produce in response to this task lead to another observable dependent on *Network Disciplinary Knowledge* and *Network Modeling*, depicted in the lower part of Figure 3. *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*.

[[Insert Figure 3 About Here]]

Note that all the observables have *Implement ContextE* as one parent. Again, this is because all the observables are formed from the same work products 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 4. Its interpretation is analogous to the Implement evidence model.

[[Insert Figure 4 About Here]]

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.⁵ The instrumental variables are modeled as taking on five values, as was the case for the student model variables. Recall that the observables can take

on any of three values. As discussed more below, the context values can take on either of two values.

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 (i.e., the ceiling value), is thus subsequently imposed the probability distribution is adjusted accordingly.

Basic formulas. Let θ_1 and θ_2 be parent variables for a child variable θ_c . Define

$$\theta_c^* \equiv \min(\theta_1, \theta_2). \quad (22)$$

Define a linear transformation of θ_c^* :

$$\theta_c^{**} \equiv c_{c,\theta_c^*} \times \theta_c^* + d_{c,\theta_c^*}. \quad (23)$$

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. (22). This would be a model of a leaky conjunction.⁶ 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 is

$$\theta_c^{**} \equiv [c_{c,\theta_c^*} \times \theta_c^* + d_{c,\theta_c^*}] + [c_{c,\theta_1} \times (\theta_1 - \theta_c^*)] + [c_{c,\theta_2} \times (\theta_2 - \theta_c^*)], \quad (24)$$

where the contents of the first set of brackets is just that defined in eq. (23), the contents of the second set of brackets captures the impact of how high above the minimum θ_1 is, and the contents of the third set of brackets captures the impact of how high above the minimum θ_2 is.⁷

Once the effective theta is obtained, it is entered into the GRM to obtain a probability distribution for the value of the child. The GRM will return probabilities for all possible values, even those outlawed by the leaky conjunction (i.e., those above θ_c^*). To fix this, we force the probabilities for the values above θ_c^* to be zero 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 2. *DK and DesignE* is formed by a leaky conjunction of *Network Disciplinary Knowledge* and *Design*. Thus to calculate the effective theta first instantiate equation (22):

$$\theta_{DKandDesign}^* \equiv \min(\theta_{NDK}, \theta_{Design}). \quad (25)$$

Next instantiate eq. (24) to calculate the effective theta:

$$\begin{aligned} \theta_{DKandDesignE}^{**} \equiv & [c_{DKandDesignE, \theta_{DKandDesign}^*} \times \theta_{DKandDesign}^* + d_{DKandDesignE, \theta_{DKandDesign}^*}] \\ & + [c_{DKandDesignE, NDK} \times (\theta_{NDK} - \theta_{DKandDesign}^*)] \\ & + [c_{DKandDesignE, Design} \times (\theta_{Design} - \theta_{DKandDesign}^*)] \end{aligned} \quad (26)$$

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 5 illustrates the correct structure of the probabilities for several of the combinations of *Network Disciplinary Knowledge* and *Design*.

[[Insert Table 5 About Here]]

The values listed in Table 5 were calculated using eq. (26) 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 modeled by having the value of $c_{DKandDesignE, Design}$ be greater than $c_{DKandDesignE, NDK}$.⁸ 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. The instrumental variables in the Design Medium and Design Hard instances are modeled in exactly the same way. That is, $\theta_{DKandDesignM}^{**}$ and $\theta_{DKandDesignH}^{**}$ are defined analogously to $\theta_{DKandDesignE}^{**}$ in eq. (26), each with their own corresponding c and d parameters.⁹ 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 for the instrumental variables in the Design Easy instance.

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. (22):

$$\theta_{DKandImplement}^* \equiv \min(\theta_{NDK}, \theta_{Implement}) \quad (27)$$

$$\theta_{DKandNM}^* \equiv \min(\theta_{NDK}, \theta_{NM}). \quad (28)$$

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

$$\begin{aligned} \theta_{DKandImplementE}^{**} \equiv & [c_{DKandImplementE, \theta_{DKandImplement}^*} \times \theta_{DKandImplement}^* + d_{DKandImplementE, \theta_{DKandImplement}^*}] \\ & + [c_{DKandImplementE, NDK} \times (\theta_{NDK} - \theta_{DKandImplement}^*)] \\ & + [c_{DKandImplementE, Implement} \times (\theta_{Implement} - \theta_{DKandImplement}^*)] \end{aligned} \quad (29)$$

and

$$\begin{aligned} \theta_{DKandNME}^{**} \equiv & [c_{DKandNME, \theta_{DKandImplement}^*} \times \theta_{DKandNM}^* + d_{DKandNME, \theta_{DKandNM}^*}] \\ & + [c_{DKandNME, NDK} \times (\theta_{NDK} - \theta_{DKandNM}^*)] \\ & + [c_{DKandNME, NM} \times (\theta_{NM} - \theta_{DKandNM}^*)] \end{aligned} \quad (30)$$

The effective thetas for the instrumental variables in the Implement Medium instance, $\theta_{DKandImplementM}^{**}$ and $\theta_{DKandNMM}^{**}$, and the effective thetas for the instrumental variables in the Implement Hard instance, $\theta_{DKandImplementH}^{**}$ and $\theta_{DKandNMH}^{**}$, are defined analogously to their counterparts in the Implement Easy instance (eqs. (29) and (30)), each with their own c and d parameters.

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 for the instrumental variables in the Implement evidence models are expected to be the same as

their counterparts in the Design evidence model instances; the conditional probabilities based on this expectation are therefore the same as were derived for the Design evidence models.

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 identical to those discussed above.

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 define a linear mapping from each parent to the child (separately) and then simply sum up the linear mappings child. More formally, if the marginal contribution of l^{th} parent variable θ_l is the linear mapping function

$$\theta_{c,l}^* \equiv c_{c,l} \times (\theta_l) + d_{c,l} \quad (31)$$

then the combination all L linear mapping functions would be

$$\theta_t^{**} \equiv \sum_{l=1}^L \theta_{c,l}^* . \quad (32)$$

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 are 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, \dots, X_T ;¹⁰ 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, \dots, X_T . Let this context variable take on any of two states, corresponding to values of High, for tasks that produce observables with a strong association due to the task, and Low, for tasks that produce observables with a weaker association. Following the discussion of the previous section, the marginal contribution of θ_1 to the t^{th} observable is

$$\theta_{t,1}^* \equiv c_{t,1} \times (\theta_1) + d_{t,1} \quad (33)$$

and the marginal contribution of θ_2 is given as

$$\theta_{t,2}^* \equiv c_{t,2} \times (\theta_2) + d_{t,2} = c_{t,2} \times (\theta_2). \quad (34)$$

Note that $d_{t,2}$ has been dropped on the right side of eq. (34). This occurs because if the two-level context variable is centered around zero, $c_{t,2}$ captures all the information and $d_{t,2}$ is unnecessary. To specify the expression for the effective theta, instantiate eq. (32):

$$\theta_t^{**} \equiv c_{t,1} \times (\theta_1) + c_{t,2} \times (\theta_2) + d_{t,1}. \quad (35)$$

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 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.¹¹ To obtain the effective theta for the t^{th} observable in the Design Easy instance, instantiate eq. (35)

$$\theta_t^{**} \equiv c_{t,DKandDesignE} \times (\theta_{DKandDesignE}) + c_{t,DesignContextE} \times (\theta_{DesignContextE}) + d_{t,DKandDesignE} \cdot \quad (36)$$

Table 6 is contains calculations of effective thetas and initial conditional probability distributions for the observables in the Design Easy evidence model (in regular typeface). These were calculated by evaluating eq. (36) 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.

[[Insert Table 6 About Here]]

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

$$\theta_t^{**} \equiv c_{t,DKandDesignM} \times (\theta_{DKandDesignM}) + c_{t,DesignContextM} \times (\theta_{DesignContextM}) + d_{t,DKandDesignM} \cdot \quad (37)$$

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

$$\theta_t^{**} \equiv c_{t,DKandDesignH} \times (\theta_{DKandDesignH}) + c_{t,DesignContextH} \times (\theta_{DesignContextH}) + d_{t,DKandDesignH} \cdot \quad (38)$$

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, $d_{t,DKandDesignH} = -7.0$. The expected strength of association between the observables and (both of) the parent variables remains unchanged. Therefore the coefficients in the Design Medium and Design Hard scenarios are expected to be equal to their counterparts in the Design Easy scenario. Table 6 gives the effective thetas and resulting conditional probabilities of response for the Design Medium (*italic typeface*) and Design Hard (**bold typeface**) 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 3. 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. For the first three observables in the Implement Easy instance, we define the effective theta as

$$\begin{aligned} \theta_t^{**} \equiv & c_{t,DKandImplementE} \times (\theta_{DKandImplementE}) \\ & + c_{t,ImplementContextE} \times (\theta_{ImplementContextE}) + d_{t,DKandImplementE} \end{aligned} \quad (39)$$

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

$$\theta_t^{**} \equiv c_{t,DKandNME} \times (\theta_{DKandNME}) + c_{t,ImplementContextE} \times (\theta_{ImplementContextE}) + d_{t,DKandNME} \quad (40)$$

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 6 (regular type).

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. Expressions for the effective thetas for the observables in the Implement Medium evidence model mimic eqs. (39) and (40). Similarly, expressions for the effective thetas for the observables in the Implement Hard evidence model mimic eqs. (39) and (40).

As in the Implement Easy evidence models, each observable in the Implement Medium and Implement Hard evidence models has its own associated *c* and *d* parameters. The coefficients and the intercepts for the observables in the Implement Medium and Implement Hard instances are expected to take on the same values as those listed for the observables in the Design Medium instance and the Design Hard instance, respectively. The distributions corresponding to SME expectation for the Implement Medium and Implement Hard instances are therefore those given in Table 6 (in italic and bold typeface, 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 corresponding

instances in the Design and Implement evidence models. The expected conditional distributions for the Easy, Medium, and Hard instances are therefore those given in Table 6.

Exogenous Variables

In the evidence models, only the Context variables are exogenous. They are modeled as taking on values of -1 and $+1$, each with probability $.5$.

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 $P(\mathbf{X} | \boldsymbol{\theta}, \boldsymbol{\pi})$ and hinted at the $P(\boldsymbol{\eta})$ terms. $P(\mathbf{X} | \boldsymbol{\theta}, \boldsymbol{\pi})$ 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.

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 (the d 's) to be distributed normally with mean defined

by expert expectation and variance of one. Similarly, we define the priors for all coefficients (the c 's) to be distributed normally with mean defined by expert expectation and variance of one, truncated at zero to force all the coefficients to be positive.

Markov Chain Monte Carlo Estimation

The Full Bayesian Model

We have devoted considerable effort to set 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 (i.e., the c 's and d 's). We are

therefore interested in the posterior distributions for these parameters. 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:

$$P(\boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\eta}, \boldsymbol{\lambda} | \mathbf{X}) \propto P(\mathbf{X} | \boldsymbol{\theta}, \boldsymbol{\pi}) \times P(\boldsymbol{\theta} | \boldsymbol{\lambda}) \times P(\boldsymbol{\lambda}) \times P(\boldsymbol{\pi} | \boldsymbol{\eta}) \times P(\boldsymbol{\eta}). \quad (41)$$

Here, $P(\boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\eta}, \boldsymbol{\lambda} | \mathbf{X})$ 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).¹²

An analytic solution for the posteriors for this model is computationally intractable and may very well be impossible. Instead, we pursue an empirical approximation via Markov chain Monte Carlo (MCMC) estimation. MCMC estimation provides an adequate and appropriate framework for computation in Bayesian analyses (Gelman et al., 1995). A complete treatment and description of MCMC estimation is beyond the scope and intent of this work; suffice it to say that for our current purposes, MCMC estimation consists of drawing from a series of distributions that is in the limit equal to drawing from the true posterior distribution (Gilks et al., 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 et al., 1953). For a more complete discussion of this and other MCMC techniques, see Brooks (1998) and Gilks et al. (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 program WinBUGS 1.4 (Spiegelhalter et al., 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 to be discarded. 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 was as high as .50, even for correlations of lag 40. In these cases, the individual chains mix quite slowly, thus chains starting from overdispersed starting values require a great number of iterations to converge.

Prior to data analysis, the first 40,000 iterations of each chain were discarded as "burn-in values" leaving 60,000 iterations per chain. These remaining iterations were pooled in the analysis of the final data for several reasons. First, all these iterations are empirical representations of the true posterior (i.e., values occur with the relative frequencies of the true posterior). Second, though there exists autocorrelations among the values *within* each chain, there is no correlation among the values *between* parallel chains as 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. A metric for summarizing the impact is the percent increase in precision, given as

$$100 \times \frac{(\text{posterior } SD)^{-2} - (\text{prior } SD)^{-2}}{(\text{prior } SD)^{-2}}; \text{ a value of zero indicates no new information is gained}$$

by incorporating the data while a value of 100 indicates that there is twice as much information regarding a parameter after incorporating the data. The average increase in precision for most of the parameters (three parameters were excluded from this analysis, as discussed below) is 118.146305 with a standard deviation of 111.808291. 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 7.

[[Insert Table 7 About Here]]

For the most part, there were mild increases in precision for the variables that define the conditional distributions of the latent variables (i.e., 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 parameters that define the conditional distributions of the observables. This is not a surprising result as the evidence contained in the data (known values for 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 (-12.656127 and -9.124036, respectively). It appears as though two factors are 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 the above analyses 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 814.943554, 1020.053619, and 3820.939739, 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.

To illustrate the impact of parameter estimation, we focus our attention on the parameters for 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 prior distributions for $c_{1,DKandTroubleshootM}$ and $c_{3,DKandTroubleshootM}$ were centered at 2. The posterior means are 2.0290 and 1.9230, respectively, which are both close to the prior mean. Large increases in precision for these parameters (415.8253% and 393.1694%, respectively) indicate that, with respect to these parameters, the data (1) conform to SME expectation and (2) 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 (posterior means of 0.9322 and 0.7590 whereas the mean of the prior was .6). On the other hand, though the priors for the intercept parameters were both the same (prior mean of -6), the posteriors are quite different. The distribution of the intercept for the first observable is lower than the prior (with a posterior mean of -6.7780). Conversely, the distribution of the intercept for the third observable (posterior mean of -4.8650) 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 third.

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 6 (in italic typeface). The posterior conditional distributions (based on the parameters' posterior means) for the first and third observables are given in Table 8 (in regular and italic

typeface, respectively), where it is clearly seen that examinees are more likely to perform well on the third observable, as compared to the first. 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.

[[Insert Table 8 About Here]]

Of particular interest are the parameters associated with the adjustments to the conjunctions (e.g., $c_{DKandDesignE,NDK}$ and $c_{DKandDesignE,Design}$ in eq. (26)). 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 .927650 (the minimum value was .7441) 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 (1) 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 (2) 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 9.

[[Insert Table 9 About Here]]

The average percent increase in precision indicates that the data informs on the student model parameters (Table 9) less than it does on the parameters that define the conditional probability distributions in either the student model or the evidence models (Table 7). 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 2 – 4). 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 for *Network Disciplinary Knowledge* is smaller than that of the other variables.¹³ 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 informs 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 while examinee B completed 6 tasks resulting in 19 data points and examinee C completed only one task, resulting in four 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 (e.g., *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 better warranted.

For example, Table 10 gives the prior and posterior probabilities of *Design* for examinees A, B, and C.¹⁴ 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. As expected, the posterior distribution for examinee A is much more concentrated 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.

[[Insert Table 10 About Here]]

The association between the number of data points and percent increase in precision observed among examinees A, B, and C bear out throughout the data. The correlation between the number of observed values and the percent increase in precision for *Network Disciplinary*

Knowledge was .619. Similarly, for *Network Modeling*, the correlation was .596. For *Design*, *Implement*, and *Troubleshoot*, the correlations were .451, .326, and .489, respectively. In contrast, the correlation for *Network Proficiency* was -.009. All but the last were statistically significant at the .01 level. 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-positing 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, Richardson, and Spiegelhalter (1996b), and Spiegelhalter et al. (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 equally 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.

Notes

¹ 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 case of observable child variables.

² 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.

³ 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 the rows in Table 1 is unchanged.

⁴ 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 $P(\boldsymbol{\theta} | \boldsymbol{\lambda})$ and the $P(\boldsymbol{\lambda})$. See note 11.

⁵ The names of all of the instrumental variables, context variables, and observables in Figures 2, 3, and 4 ended with ‘E’, indicating that these instantiations were the Design Easy, Implement Easy, and Troubleshoot Easy instantiations, respectively.

⁶ 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.

⁷ 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 zero. 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 zero.

⁸ This can be illustrated in much the same way as the expected difference between *Design*, *Implement*, and *Troubleshoot*.

⁹ Note that we need not compute counterparts of eq. (22) 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.

¹⁰ 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.

¹¹ 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.

¹² Note the similarity between eq. (41), the posterior distribution, and eq. (10), the joint distribution. The difference is that in the joint distribution, \mathbf{X} is a random variable, while in the posterior distributions for the parameters, \mathbf{X} is fixed at the values that are actually observed.

¹³ 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 excepting *Network Disciplinary Knowledge*.

¹⁴ The prior was calculated by compiling the distribution with all conditional probability parameters set to the values defined by expert expectation.

References

- Almond, R. G., & Mislevy, R. J. (1999). Graphical models and computerized adaptive testing. *Applied Psychological Measurement, 23*, 223–237.
- Andrich, D. (1982). An extension of the Rasch model for ratings providing both location and dispersion parameters. *Psychometrika, 47*, 105-113.
- 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., and Gelman, A. (1998). Alternative methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics, 7*, 434-455.
- DeMark, S. F., & Behrens, J. T. (in submission). Using statistical natural language processing for understanding complex responses to free-response tasks. *International Journal of Testing*.
- 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 & 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 & Hall.
- Gelman, A. and Rubin, D. B. (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.
- Lindley, D. V. & Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B*, 34, 1-42.
- 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) *On the consequences of ignoring certain conditional dependencies in cognitive diagnosis*. Paper presented at the Annual Meeting of the American Statistical Association, Orlando, FL, August, 1995.
- Mislevy, R. J., Senturk, D., Almond, R. G., Dibello, L. V., Jenkins, F., Steinberg, L. S., & Yan, D. (2002). Modeling conditional probabilities in complex educational assessments. CSE Technical Report 580, Los Angeles: 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. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3-62.
- 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.
- Schum, D. A. (1987). *Evidence and inference for the intelligence analyst*. Lanham, Md.: University Press of America.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph No. 17*, 34, (No. 4, Part 2).
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, 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. <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. (in submission). Design rationale for a complex performance assessment. *International Journal of Testing*.

Table 1

Conditional Probability Table for *Design, Implement, and Troubleshoot*

			$P(\text{Design} = k)$				
<i>Network Proficiency</i>	θ_{NP}	θ_{Design}^{**}	Novice	Semester 1	Semester 2	Semester 3	Semester 4
Novice	1	-3.8	0.6900	0.2527	0.0492	0.0071	0.0011
		-4.2	0.7685	0.1923	0.0337	0.0047	0.0007
		-5.0	0.8808	0.1012	0.0155	0.0021	0.0003
Semester 1	2	-1.8	0.2315	0.4585	0.2527	0.0492	0.0082
		-2.2	0.3100	0.4585	0.1923	0.0337	0.0055
		-3.0	0.5000	0.3808	0.1012	0.0155	0.0025
Semester 2	3	0.2	0.0392	0.1923	0.4585	0.2527	0.0573
		-0.2	0.0573	0.2527	0.4585	0.1923	0.0392
		-1.0	0.1192	0.3808	0.3808	0.1012	0.0180
Semester 3	4	2.2	0.0055	0.0337	0.1923	0.4585	0.3100
		1.8	0.0082	0.0492	0.2527	0.4585	0.2315
		1.0	0.0180	0.1012	0.3808	0.3808	0.1192
Semester 4	5	4.2	0.0007	0.0047	0.0337	0.1923	0.7685
		3.8	0.0011	0.0071	0.0492	0.2527	0.6870
		3.0	0.0025	0.0155	0.1012	0.3808	0.5000

Table 2

Conditional Probability Table for *Network Modeling*

			$P(\text{Network Modeling} = k)$				
<i>Network Disciplinary Knowledge</i>	θ_{NDK}	θ_{NM}^{**}	Novice	Semester 1	Semester 2	Semester 3	Semester 4
Novice	1	-6.0	1	0	0	0	0
Semester 1	2	-4.0	0.7675	0.2325	0	0	0
Semester 2	3	-2.0	0.2823	0.4851	0.2325	0	0
Semester 3	4	0.0	0.0498	0.2325	0.4851	0.2325	0
Semester 4	5	2.0	0.0067	0.0407	0.2215	0.4621	0.2689

Table 3

Conditional Probability for *Network Proficiency*

<i>NDK</i>	θ_{NDK}	<i>NM</i>	θ_{NM}	θ_{NP}^{**}	P(<i>Network Proficiency</i> =k)				
					Novice	Semester 1	Semester 2	Semester 3	Semester 4
Novice	1	Novice	1	-3	1	0	0	0	0
Sem 1	2	Novice	1	-2	0.3679	0.6321	0	0	0
Sem 1	2	Sem 1	2	-1	0.2384	0.7616	0	0	0
Sem 2	3	Novice	1	-1	0.1353	0.4323	0.4323	0	0
Sem 2	3	Sem 1	2	0	0.0649	0.3030	0.6321	0	0
Sem 2	3	Sem 2	3	1	0.0360	0.2024	0.7616	0	0
Sem 3	4	Novice	1	0	0.0498	0.2325	0.4851	0.2325	0
Sem 3	4	Sem 1	2	1	0.0204	0.1149	0.4323	0.4323	0
Sem 3	4	Sem 2	3	2	0.0092	0.0557	0.3030	0.6321	0
Sem 3	4	Sem 3	4	3	0.0050	0.0310	0.2024	0.7616	0
Sem 4	5	Novice	1	1	0.0180	0.1012	0.3808	0.3808	0.1192
Sem 4	5	Sem 1	2	2	0.0067	0.0407	0.2215	0.4621	0.2689
Sem 4	5	Sem 2	3	3	0.0025	0.0155	0.1012	0.3808	0.5000
Sem 4	5	Sem 3	4	4	0.0009	0.0058	0.0407	0.2215	0.7311
Sem 4	5	Sem 4	5	5	0.0003	0.0021	0.0155	0.1012	0.8808

Table 4

Probability Table for *Network Disciplinary Knowledge*

Pr (<i>Network Disciplinary Knowledge</i> = k)				
Novice	Semester 1	Semester 2	Semester 3	Semester 4
0.0148	0.0850	0.3504	0.4080	0.1419

Table 5

Portion of the Conditional Probability Table for the instrumental variables (e.g., *NDK* and *DesignE*)

<i>Network Disciplinary Knowledge</i>		<i>P(NDK and DesignE = k)</i>					
		<i>Design</i>	Novice	Semester 1	Semester 2	Semester 3	Semester 4
Semester 3	Novice		1.0	0	0	0	0
	Semester 1		0.3064	0.6936	0	0	0
	Semester 2		0.0568	0.2787	0.6645	0	0
	Semester 3		0.0092	0.0557	0.3030	0.6321	0
	Semester 4		0.0070	0.0431	0.2564	0.6936	0
Semester 4	Novice		1.0	0	0	0	0
	Semester 1		0.2806	0.7194	0	0	0
	Semester 2		0.0500	0.2564	0.6936	0	0
	Semester 3		0.0080	0.0488	0.2787	0.6645	0
	Semester 4		0.0009	0.0058	0.0407	0.2215	0.7311

Table 6: Conditional distributions for observables in Design Easy, Medium, and Hard Scenarios

<i>DK and Design</i>	$\theta_{DKandDesignM}$	<i>Design Context</i>	$\theta_{DesignContextM}$	θ_t^{**}	P($X=k$)		
					Low	Medium	High
Novice	1	Low	-1	-1.7	0.8022	0.1933	0.0045
				-2.2	0.9168	0.0815	0.0017
		High	1	-2.7	0.9677	0.0317	0.0006
				-1.3	0.6457	0.3444	0.0100
Semester 1	2	Low	-1	-1.8	0.8320	0.1643	0.0037
				-2.3	0.9309	0.0678	0.0014
		High	1	-0.7	0.3543	0.6134	0.0323
				-1.2	0.5987	0.3892	0.0121
Semester 2	3	Low	-1	-1.7	0.8022	0.1933	0.0045
				-0.3	0.1978	0.7330	0.0691
		High	1	-0.8	0.4013	0.5721	0.0266
				-1.3	0.6457	0.3444	0.0100
Semester 3	4	Low	-1	0.3	0.0691	0.7330	0.1978
				-0.2	0.1680	0.7488	0.0832
		High	1	-0.7	0.3543	0.6134	0.0323
				0.7	0.0323	0.6134	0.3543
Semester 4	5	Low	-1	0.2	0.0832	0.7488	0.1680
				-0.3	0.1978	0.7330	0.0691
		High	1	1.3	0.0100	0.3444	0.6457
				0.8	0.0266	0.5721	0.4013
Semester 4	5	Low	-1	0.3	0.0691	0.7330	0.1978
				1.7	0.0045	0.1933	0.8022
		High	1	1.2	0.0121	0.3892	0.5987
				0.7	0.0323	0.6134	0.3543
Semester 4	5	Low	-1	2.3	0.0014	0.0678	0.9309
				1.8	0.0037	0.1643	0.8320
		High	1	1.8	0.0037	0.1643	0.8320
				2.7	0.0006	0.0317	0.9677
High	1	2.7	0.0006	0.0317	0.9677		
		2.7	0.0006	0.0317	0.9677		

Table 7

Average Percent Increase in Precision for Parameters that Define Conditional Distributions by

Model Portion

Model Fragment	Average Increase In Precision
Student Model	54.3026
Latent variables in Design evidence models	82.9780
Observable variables in Design evidence models	216.0490
Latent variables in Implement evidence models	85.0373
Observable variables in Implement evidence models	254.7452
Latent variables in Troubleshoot evidence models	85.9908
Observable variables in Troubleshoot evidence models	147.4303

Table 8

Posterior Conditional Probability Table for the First and Third Observables in the Troubleshoot

Medium evidence model

<i>DK and TrbM</i>	<i>Trb ContextM</i>	<i>P(X=k)</i>		
		Low	Medium	High
Novice	Low	0.9754	0.0241	0.0005
		<i>0.8457</i>	<i>0.1510</i>	<i>0.0033</i>
	High	0.8602	0.1369	0.0030
		<i>0.5456</i>	<i>0.4394</i>	<i>0.0150</i>
Semester 1	Low	0.8392	0.1573	0.0035
		<i>0.4447</i>	<i>0.5329</i>	<i>0.0224</i>
	High	0.4471	0.5307	0.0221
		<i>0.1493</i>	<i>0.7562</i>	<i>0.0945</i>
Semester 2	Low	0.4069	0.5671	0.0260
		<i>0.1048</i>	<i>0.7599</i>	<i>0.1353</i>
	High	0.0961	0.7569	0.1469
		<i>0.0250</i>	<i>0.5584</i>	<i>0.4165</i>
Semester 3	Low	0.0827	0.7485	0.1688
		<i>0.0168</i>	<i>0.4662</i>	<i>0.5170</i>
	High	0.0138	0.4191	0.5671
		<i>0.0037</i>	<i>0.1662</i>	<i>0.8301</i>
Semester 4	Low	0.0117	0.3813	0.6070
		<i>0.0025</i>	<i>0.1177</i>	<i>0.8798</i>
	High	0.0018	0.0894	0.9088
		<i>0.0005</i>	<i>0.0285</i>	<i>0.9709</i>

Table 9

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
<i>Network Disciplinary Knowledge</i>	0.8869	0.8330	25.5337	882.9438	77.1964	-10.4176
<i>Network Modeling</i>	1.0944	0.9506	48.1472	951.4628	107.4095	10.7324
<i>Network Proficiency</i>	1.0770	1.0590	10.2687	54.3379	20.9179	-20.6480
<i>Design</i>	1.2577	1.1262	34.5901	980.0184	146.4760	0.2695
<i>Implement</i>	1.2616	1.0412	52.3769	99.9639	139.7585	63.7278
<i>Troubleshoot</i>	1.2407	1.0479	47.5350	380.8432	53.9321	16.8007

Table 10

Prior and posterior density functions of *Design* for Examinees A, B, and C

	Prior	A	B	C
Novice	0.1137	0.0000	0.0024	0.2106
Semester 1	0.1872	0.0000	0.0614	0.2806
Semester 2	0.2709	0.0052	0.4999	0.2508
Semester 3	0.2446	0.1519	0.3087	0.1537
Semester 4	0.1835	0.8429	0.1274	0.1042

Figure Captions

Figure 1. The NetPASS student model

Figure 2. A Design evidence model

Figure 3. An Implement evidence model

Figure 4. A Troubleshoot evidence model







