CRESST REPORT 837

DYNAMIC BAYESIAN NETWORK MODELING OF GAME BASED DIAGNOSTIC ASSESSMENTS

JANUARY, 2014

Roy Levy

National Center for Research
on Evaluation, Standards, & Student Testing
UCLA | Graduate School of Education & Information Studies

Dynamic Bayesian Network Modeling of Game Based Diagnostic Assessments

CRESST Report 837

Roy Levy
Arizona State University

January 2014

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
300 Charles E. Young Drive North
GSE&IS Bldg., Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

This research was supported by the Center for Advanced Technology in Schools (CATS), PR/Award Number R305C080015, as administered by the Institute of Education Sciences, U.S. Department of Education.

The findings and opinions expressed here do not necessarily reflect the positions or policies of the Center for Advanced Technology in Schools (CATS), the National Center for Education Research (NCER), the Institute of Education Sciences (IES), or the U.S. Department of Education.

The author wishes to thank Deirdre Kerr for invaluable insights regarding interpretations of student activities that form the basis for this work, and Mark Hansen and Ron Dietel for comments on earlier versions that have yielded improvements to this report.

To cite from this report, please use the following as your APA reference: Levy, R. (2014). *Dynamic Bayesian Network Modeling of Game Based Diagnostic Assessments* (CRESST Report 837). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

TABLE OF CONTENTS

DYNAMIC BAYESIAN NETWORK MODELING OF GAME BASED DIAGNOSTIC ASSESSMENTS

Roy Levy
Arizona State University

Abstract

Digital games offer an appealing environment for assessing student proficiencies, including skills and misconceptions in a diagnostic setting. This paper proposes a dynamic Bayesian network modeling approach for observations of student performance from an educational video game. A Bayesian approach to model construction, calibration, and use in facilitating inferences about students on the fly is described.

Introduction

Games offer an appealing environment for conducting assessment, in part because games and assessments share many structural features (Behrens, Frezzo, Mislevy, Kroopnick, & Wise, 2008). Games that employ digitally based simulations and data recording afford opportunities for more complex assessment arguments than typical paper and pencil tests (Levy, 2013), including the monitoring of learning and change over time (Rowe & Lester, 2010; Shute, 2011). Finally because of their motivational nature, games are often attractive to students and may support the integration of assessment and learning activities in a dynamic, longitudinal approach (Shute, 2011).

To date, psychometric modeling strategies for game-based assessments have been somewhat limited. To expand these strategies, this paper illustrates the construction, calibration, and use of a dynamic Bayesian Network (DBN) psychometric model for student performance in *Save Patch*, an educational game targeting rational number mathematics. Bayesian Networks (BNs) provide a fruitful framework for modeling student performance during game-based assessments (e.g., Iseli, Koenig, Lee, & Wainess, 2010; Rowe & Lester, 2010; Shute, 2011). Similarly, BNs have been gainfully employed in a variety of assessments that share features with games, such as simulation-based assessments that tend to share digital modalities of presentation and data collection (e.g., Almond, Mulder, Hemat, & Yan, 2009) and intelligent tutoring systems in which students receive feedback and there is the possibility (really, hope) that students learn during their interactions with the system (e.g., Mislevy & Gitomer, 1996; Reye, 2004; Sao Pedro, Baker, Gobert, Montalvo, & Nakama, 2013; VanLehn, 2008).

The conditional probability structures underlying the BN in game-based assessment are commonly specified in advance by subject matter experts (Iseli et al., 2010; Rowe & Lester,

2010; Shute, 2011). As with other assessment applications, it may be desirable to obtain estimates of the conditional probabilities for game-based assessments based on student performance data. However, the longitudinal dependence structures of game-based assessments (discussed below) pose challenges to estimating these conditional probabilities (Iseli et al., 2010; Rowe & Lester, 2010). Procedures for estimating the parameters of DBNs have, to date, been applied to models with dichotomous latent and observable variables, as are appropriate for tutoring systems (e.g., Baker, Pardos, Gowda, Nooraei, & Heffernan, 2011; Chang, Beck, Mostow, & Corbett, 2006; Sao Pedro et al., 2013), which may be leveraged for game-based assessment with similar assumptions about latent and observable variables. However, in game-based assessment, the student workspace is typically quite open, in which case students can engage in a variety of behaviors. When distinctions among multiple types of performance are warranted, polytomous evaluation of student performance may be necessary, making the aforementioned strategies prohibitively difficult. The author knows of no application in which a DBN psychometric model has been formally specified and calibrated via fitting the model to a dataset from a complex game-based assessment, or an intelligent tutoring system characterized by (a) polytomous evaluations of student performances which inform on (b) multiple aspects of proficiency.

This paper proposes methods to aid in filling this void, describing the construction and calibration of a DBN psychometric model for student performance during a complex game. The proposed methods combine the data with subject matter expertise in the form of hard and soft constraints. Challenges to model specification germane to game-based assessments, as well as challenges to the estimation of such models, are addressed by formulating the model in a fully Bayesian framework using Markov chain Monte Carlo procedures to obtain the posterior distribution for the unknown parameters. In addition, we demonstrate the use of a DBN for supporting inferences about students.

The next section describes key features of the game that motivate the modeling choices. The model for student performance in *Save Patch* follows, as well as a description of the calibration. This paper describes uses of the model to support inferences about students, then concludes with a brief summary and discussion.

Context: Save Patch

Save Patch (Chung et al., 2010) is an educational video game targeting rational number addition, developed by the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) at the University of California, Los Angeles, and the Game Innovation Lab at the University of Southern California. This brief overview is oriented toward the description of

the psychometric model developed in this work; more complete descriptions of the game are given by Center for Advanced Technology in Schools (2012), Chung et al. (2010), and Kerr and Chung (2012a, 2012b).

In *Save Patch*, students engage in a number of game *levels*, in which the student is presented with the setup of the board, and a set of resources in the form of ropes. The aim is for the student to strategically place the ropes such that the game character, Patch, successfully makes it from the starting position to target destination. As is typical in games, gameplay starts at Level 1 with subsequent levels presented in order, if the student reaches them. On each attempt, the student lays out the ropes, and then sets Patch in motion. If Patch successfully reaches the destination, the student proceeds to the next level. If an attempt is unsuccessful, the student remains at the current level and tries again. Simple descriptive summaries of student performance in *Save Patch* have been developed in service of evaluations of the game as a learning experience as well as an assessment (Delacruz, Chung, & Baker, 2010; Kerr & Chung, 2012b). The current work seeks to complement those summaries with potentially richer summaries through more formal measurement modeling for use in such studies.

The levels of *Save Patch* are explicitly designed to target various aspects of proficiency in the domain of rational number addition. For brevity, we refer to targeted aspects of proficiency as skills. Earlier levels target more basic skills; later levels target more advanced skills. In the current work we consider the first 23 levels of the game. The skills and the levels of the game that target them are listed in Table 1.

Table 1

Targeted Aspects of Proficiency (Skills) and Associated Levels

Skills	Levels
Whole numbers	1-3
Unit fractions	4-8
Whole numbers and unit fractions	9-12
Crossing the unit bar	13-15
Adding unit fractions	16-19
Adding improper fractions	20-23

Save Patch was designed in a principled manner so that student behaviors would be reflective of the various skills, with key game mechanics linked to mathematical operations (Chung et al., 2010). Student behaviors are recorded in log files, which were likewise designed

to record and distinguish between salient features of performance indicative of student thinking and decision making. Cluster analyses of log files obtained from students playing early versions of the game led to the identification of a number of strategies adopted by students (Kerr & Chung, 2012b; Kerr, Chung, & Iseli, 2011). These included various solution strategies, as well as strategies that corresponded to misconceptions about mathematics as they occur in the game as well as more broadly. Moreover, student behaviors were associated with these strategies. These analyses support the characterization of each student attempt in terms of these strategies, be it a solution strategy or a misconception. For example, a student who behaved in ways that indicated they did not identify the fractional representation correctly was making a partitioning error, which is indicative of the misconception that, in the game, the denominator is determined by counting the dividing marks along the level, which is in turn indicative that the student does not understand that the denominator in a fraction represents the number of identical parts in one whole unit.

Other strategies were identified that reflected approaches to gameplay. For example, students sometimes used everything given to them in the order that it was presented. This was interpreted as having the misconception that the order in which resources are given corresponds to the solution to the level. This does not provide much evidence about the student's mathematical proficiencies beyond their attempt to "game the system" rather than trying to solve the problem in the intended manner.

These analyses and the results are described in detail by Kerr et al. (2011) and Kerr and Chung (2012b), who framed the assessment activities in *Save Patch* in terms of evidence-centered design (ECD; Mislevy, Steinberg, & Almond, 2003). In the ECD framework, the characterization of salient features of student performances constitutes the definition of evidence identification rules. For our purposes, these amount to rules to process the log files and produce a variable for each student's attempt on each level that summarizes performance. These variables are referred to as *observable* variables, as they are summaries of student actions, and play the role of observable variables in latent variable measurement models, as discussed in following sections.

Table 2 lists the 18 possible observable values for any attempt. The first five represent different solutions. For each level there is a Standard Solution, which is taken as evidence of mastery of the skill. Certain levels can be successfully completed—that is, Patch can be directed to the target destination safely—using other solutions, these are denoted by Fractional Solution, Shortcut Solution, and Alternate Solution. Incomplete Solution refers to the situation where the student partially lays out a correct solution; in this case Patch does not make it to the destination and the student retakes the level.

Table 2

Observable Values and Corresponding Misconceptions

Observable value	Corresponding misconceptions
Standard Solution	
Fractional Solution	
Alternate Solution	
Incomplete Solution	
Shortcut Solution	
Reset Solution	
Wrong Direction	
Skipped Key	
Wrong Numerator	Iterating error
Saw As Mixed Number	Converting to wholes error
Counted Hash Marks	Partitioning error
Counted Hash Marks and Posts	Partitioning error
Saw As One Unit	Unitizing error
Saw As Wholes	Unitizing error
Saw As One Unit and Counted Hash Marks	Partitioning error, Unitizing error
Saw As One Unit and Counted Hash Marks and Posts	Partitioning error, Unitizing error
Everything In Order	Avoiding math
Unknown Error	

Reset Solution refers to the situation where the student lays out a correct solution, but instead of setting Patch along the path, s/he elects to reset the level and try again. Wrong Direction refers to an attempt where the student lays out ropes in a way that the math appears correct, but the orientation of the ropes is not correct. This is viewed as indicative of misunderstanding some of the game mechanics, rather than the mathematics. Similarly, Skipped Key refers to the situation where the attempt is unsuccessful because the path laid out fails to obtain a key needed to open a lock and successfully complete the level.

The remaining values refer to different types of incorrect attempts. The second column of Table 2 lists the misconceptions for these values. The last value, Unknown Error, refers to unsuccessful attempts that could not be otherwise characterized.

Not all observable values are possible on every level. Most levels only have a few possible values, in addition to the Standard Solution and Unknown Error, which are possible on every level.

A simpler approach would be to characterize each attempt dichotomously, such as being successful if Patch reaches the destination and unsuccessful if Patch does not reach the destination. Such an approach would simplify both the characterization of student performances and the accompanying psychometric model, which is the focus of this work, and may be sufficient for high level analyses (Kerr & Chung, 2012b). However, such a characterization might not be sufficient for supporting other sorts of inferences, including real-time inferences regarding student proficiencies and misconceptions. The more complex evidence identification processes adopted here may be warranted if they provide a richer summary of performance. Far from arbitrary, these choices represent beliefs about how distinct behaviors have differential evidentiary bearing on the desired inferences about students (Levy, 2013). In *Save Patch* we wish to draw distinctions between different types of errors, which may differentially constitute evidence of lack of proficiency or misconceptions, as well as distinctions between different types of solutions, which may differentially constitute evidence of proficiency, efficiency, strategies, etc.

Note that the category Unknown Error is used to denote attempts for which the approach could not be summarized with one of the other categories. This is something of a catch-all category, capturing unsuccessful attempts that cannot otherwise be characterized. Thus, the specification of evidence identification rules, used to evaluate student attempts and yield a categorization into one of the categories listed in Table 2, represents purposeful choices made to capitalize on the digital records of log files to provide a more nuanced view, without fully requiring that all different behaviors be interpreted and categorized separately (Levy, 2013).

Dynamic Bayesian Network Psychometric Model

In this section we develop a DBN psychometric model for attempts in *Save Patch*. Latent variables are employed to capture beliefs about proficiencies, and observable variables are used to capture student performance. More specifically each observable can take on any of the values listed in the first column of Table 2 except for Reset Solution, Wrong Direction, and Skipped Key. Attempts with these results were ignored from the current analysis because they were associated with game mechanics rather than underlying mathematics proficiency such that there are not firm beliefs about the evidentiary relevance of these behaviors. Such attempts were ignored in the model-fitting and employment, described in later sections.

The model departs from traditional psychometric models that include a single, static latent variable for each aspect of proficiency and a single observable variable for each task. These psychometric features may not be appropriate for games and related systems wherein students may take multiple attempts at a particular level, accompanied by feedback provided to the

student during the game, leading to the possibility (if not outright desirability) of student learning during the game, both within and between levels. Feedback is often made explicit in intelligent tutoring systems (e.g., VanLehn & Niu, 2001), but is present in games like *Save Patch* because a student knows whether or not they were successful on an attempt at a level—if they are successful, they move on to the next level, if not, they remain at the same level and try again. Importantly, this stands in contrast to traditional assessment experiences where the student is not informed whether they successfully completed the task until (sometimes long) after the assessment, and the student does not get repeat attempts at tasks on which they are unsuccessful. In this section, we develop the model more suitable for these features of the game.

A BN (Pearl, 1988) is a statistical model that represents the joint distribution of a set of discrete variables via recursive conditional distributions of the variables. A BN can be represented as an acyclic directed graph (commonly referred to as a directed acyclic graph, DAG), which depicts the dependence and conditional independence relationships in the model. BNs afford considerable flexibility in modeling dependence structures that arise in assessment (Almond, DiBello, Moulder, & Zapata-Rivera, 2007; Almond, et al., 2009; Almond, Williamson, Mislevy, & Yan, in press; Levy & Mislevy, 2004; Mislevy et al., 2002).

For the current application, we develop a DBN (Reye, 2004; Rowe & Lester, 2010; VanLehn, 2008) which is a type of BN oriented toward modeling time series and related longitudinal data structures. Figure 1 shows a DAG for a simple DBN for a model with a possibly vector valued latent variable ($\boldsymbol{\theta}$), and possibly vector-valued observable variable ($\mathbf{X}$), for each student i at each time point t. The plate over i indicates the structure is repeated over students i, which reflects an exchangeability assumption, facilitating the construction of the model at the individual student level. Likewise, the plate over t indicates a repeated structure over time. The model has two main components. The first is the component within time points, represented in the DAG by the directed edge from the latent to observable variable at any point. This reflects a structuring where, within time points, the observable variable is modeled as stochastically dependent on the latent variable. The second is a transition component between time points, represented in the DAG by the directed edges from the latent variable and observable variable at one time point to the latent variable at the next time, reflecting the stochastic dependence of proficiency at the later time on proficiency and performance at the earlier time.

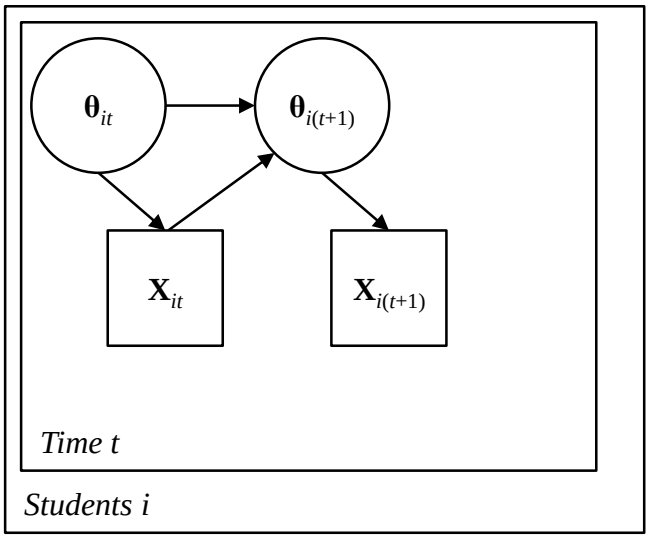

Figure 1. Graph for a dynamic Bayesian network.

Within-Time Component

The within-time component specifies, at each time point, the joint distribution of the observable and the latent variables. Let $\boldsymbol{\theta}_{it} = (\theta_{i1t},\ldots,\theta_{iMt})$ denote the values of the M latent variables for student i at time t.

As is common in diagnostic classification models (Rupp, Templin, & Henson, 2010), for each of the six skills in *Save Patch* we specify a dichotomous latent variable with categories corresponding to mastery, coded as 1, and nonmastery, coded as 0. Each level targets exactly one of these skills. In addition, whether or not the student possesses different misconceptions or (listed in the second column of Table 2) is posited to influence performance. Accordingly, we specify a dichotomous latent variable for each of the misconceptions with categories corresponding to the student possessing/not possessing that misconception, coded as 1 and 0, respectively (Bradshaw & Templin, in press).

At any time point, the latent variables for each student is modeled as the following independent Bernoulli distribution

$$P(\theta_{imt}) \sim B(p_{imt}) \quad (1)$$

where p_{imt} is a prior probability of mastery/possession for skill or misconception m for person i at time t. The notion of it being the prior for time t is meant to indicate that it represents the beliefs just before the observation made at time t. It will serve as the prior distribution for an instantiation of Bayes Theorem that synthesizes the data observed from time t. Note that in general, each student has their own probability distribution for each latent variable at each time. At the initial time, t=1, an assumption of exchangeability implies a common prior across students:

$$P(\theta_{im1}) \sim B(p_{m1}) \quad (2)$$

The parameters governing the distribution of the initial values of the latent variables are assigned independent of diffuse prior distributions

$$P(p_{m1}) \sim Beta(1,1). \quad (3)$$

In *Save Patch*, not every latent variable is measured in every level, including the first level. Accordingly, we specify the prior distribution in (2)-(3) for each latent variable on the first attempt for the first level in which the latent variable is measured. As examples, the prior distribution for each student's latent variable for Whole Number is specified on the first attempt at Level 1, but the prior distribution for each student's latent variable for Unit Fractions is specified on the first attempt at Level 4. The specification of independent prior distributions for the latent variable is a simplification assumed for ease of modeling. A more complex model may specify and parameterize dependence structures among the latent variables if they are of inferential interest (e.g., de la Torre & Douglas, 2004; Levy & Mislevy, 2004). Importantly, the specification of independent priors does not force independence in the posterior distribution, as may emerge based on incorporation of the data.

Distributions of the subsequent values for the latent variables may change over time, as governed by the transition component, described in a later section. The measurement model is assumed to be time-invariant, and as such time is not salient for the discussion of the measurement model and we therefore drop the time index *t* for the current exposition.

Let K denote the number of possible categories of the observables, and recall that M is the number of latent variables. We note that not every latent variable is measured and not every category of the observables is possible in every level of *Save Patch*. The model can be formulated for the individual levels, but for ease of notation we present the model in terms of the general case. Where necessary, we will denote the number of latent variables measured in Level j and the number of categories for the observable based on Level j as M_j and K_j, respectively.

As each latent variable is dichotomous, there are 2^M possible values of $\boldsymbol{\theta}$, each of which corresponds to a different profile of the latent variables. Let $\pi_{jk|\boldsymbol{\theta}=c} = P(X_{ija} = k \mid \boldsymbol{\theta}_i = c)$ denote the probability that student i with a latent variable profile c yields an observable value of k on attempt a on level j. Let $\boldsymbol{\pi}_{j|\boldsymbol{\theta}=c} = (\pi_{j1|\boldsymbol{\theta}=c}, \ldots, \pi_{jK|\boldsymbol{\theta}=c})$ denote the collection of the K category specific conditional probabilities for level j and latent variable profile c. Finally, let $\boldsymbol{\pi}_j = (\boldsymbol{\pi}_{j|\boldsymbol{\theta}=1}, \ldots, \boldsymbol{\pi}_{j|\boldsymbol{\theta}=M})$ denote the full collection of conditional probabilities for level j. Because not every level measures every latent variable or has every possible category for the observable, in total $\boldsymbol{\pi}_j$ contains $(2^{M_j})(K_j)$ conditional probabilities, of which $(2^{M_j})(K_j - 1)$ are free owing to the restriction that

$$\sum_{k=1}^{K_j} \pi_{jk|\boldsymbol{\theta}=c} = 1 \tag{4}$$

(i.e., for each latent variable profile the conditional probabilities for the K_j possible values must sum to 1). For example, Level 1 contains $K_1 = 6$ possible observable values (Standard Solution, Alternate Solution, Saw as One Unit, Saw as One Unit and Counted Hash Marks, Everything in Order, Unknown Error) and measures $M_1 = 4$ latent variables (the targeted proficiency Whole Numbers, and misconceptions/strategies Partitioning Error, Unitizing Error, and Avoiding Math), yielding 96 total and 80 free conditional probabilities.

To simply the specification and estimation of these probabilities, we employ item response theoretic models for structuring the conditional probabilities (Almond et al., 2001; Levy & Mislevy, 2004), ultimately reducing the parameterization. Specifically, we leverage innovations inherent in the Scaling Individuals and Classifying Misconceptions (SICM) model (Bradshaw & Templin, in press). The SICM model employs a continuous latent variable representing proficiency and dichotomous latent variables representing misconceptions. The model developed here departs from the SICM by using discrete rather than continuous latent variables to represent proficiency, along with the discrete latent variables for misconceptions. More generally, we may model any skill, misconception, or other attribute as a discrete latent variable, and as such may then be viewed as a Skill, Misconception, or Attribute Classification (SMAC) model. In addition, the model here extends the SICM model of Bradshaw and Templin (in press) to model multiple targeted aspects of proficiency. The proceeding exposition is similar to that of Bradshaw and Templin (in press), though departs in several places owing to the aforementioned differences.

The model parameterizes $\boldsymbol{\pi}_{j|\boldsymbol{\theta}=c}$ via a polytomous logistic regression framework (e.g., Agresti, 2002). Owing to the restriction in (4), for each observable we specify one category as a baseline and $K - 1$ non-redundant logits that model the probability of a value in another category relative to the baseline category. The model for each of the K categories can be expressed as

$$\pi_{jk|\boldsymbol{\theta}=c} = P(X_{ija} = k \mid \boldsymbol{\theta}_i = c) = \frac{\exp(\lambda_{jk0} + \boldsymbol{\lambda}'_{jk}\boldsymbol{\theta}_i)}{\sum_{k=1}^{K_j} \exp(\lambda_{jk0} + \boldsymbol{\lambda}'_{jk}\boldsymbol{\theta}_i)}. \tag{5}$$

As formulated in (5), the model takes on a form similar to a multidimensional item response model (Reckase, 2009). Alternatively, it may be formulated in terms of incidence matrices indicating whether certain levels measure certain proficiencies and misconceptions (e.g., Bradshaw & Templin, in press), in which case it is an unordered-category extension of the model introduced by von Davier (2005).

We choose the category of Unknown Error to serve as the baseline category because it is a possible value in every level and facilitates a natural interpretation of the parameters, as discussed below. Without loss of generality, this baseline category is coded as the highest category K. For the remaining categories $k \neq K$, the intercept λ_{jk0} is the logit of the observable taking on value k over the baseline category of Unknown Error for a student who has not mastered the targeted skill and possesses none of the misconceptions relevant for category k. Larger values of λ_{jk0} indicate that category k on level j is more likely, holding all else constant.

$\boldsymbol{\lambda}_{jk} = (\lambda_{jk1},\ldots, \lambda_{jkM})'$ is the $(M_j \times 1)$ vector of discrimination parameters capturing the effect of the corresponding M latent variables on the probability of the student's attempt being in category k. Larger magnitudes of any λ_{jkm} indicate larger differences in the probability of the attempt being in category k on level j for students with and without the associated skill or misconception. Positive values of λ_{jkm} indicate that students who possess the skill or misconception are more likely to have a value of category k; negative values of λ_{jkm} indicate that students who possess the skill or misconception are less likely to respond with category k. A value of 0 for λ_{jkm} indicates that the probability of having a value of k on level j does not vary with whether or not the student possesses skill or misconception m.

As discussed above, each level measures only one of the targeted skills and possibly only some of the misconceptions. As a result, not all of the categories for the observable are possible in each level. If category k on level j is not possible, its probability is set to 0 (which may be viewed as fixing $\lambda_{jkm} = 0$ for all m and $\lambda_{jk0} = -\infty$).

Similarly, we set $\lambda_{jkm} = 0$ for all m if the level does not measure the associated skill or misconception, and $\lambda_{jkm} = 0$ if category k does not measure skill or misconception m. Accordingly, for each level, the discriminations for all of the skills *other than* the target skill for the level are set to 0, as are the discriminations for all misconceptions not measured in that level.

The implications of the previous are as follows. For any of the solution categories (Standard Solution, and if possible on the level, Fractional Solution, Alternate Solution, and Incomplete Solution), the discrimination for the latent variable for the targeted skill is included, but the remaining discriminations are set to 0. For any category representing a misconception, the discrimination parameter for the latent variable representing that particular misconception is included, and the discriminations for the remaining latent variables (i.e., those that represent skills or other misconceptions) are set to 0. Finally, to identify the model, the parameters for the baseline category of Unknown Error are fixed to be $\lambda_{jKm} = 0$, $m=1,\ldots,M$. This is summarized in Table 3, which generically represents the prior distributions and the structure of the discriminations for a hypothetical level in which all categories are possible.

Table 3

Prior Distributions for Model Parameters (Within-Time Component)

Observable value	Location parameter	Targeted aspect of proficiency (Skill)	Iterating error	Converting to wholes error	Partitioning error	Unitizing error	Avoiding math
Standard Solution	$N(0, 10)$	$\lambda_{SS} \sim N(2, 10)\ C(0,)$					
Fractional Solution	$N(0, 10)$	$N(2, 10)\ C(0, \lambda_{SS})$					
Alternate Solution	$N(0, 10)$	$N(2, 10)\ C(0, \lambda_{SS})$					
Incomplete Solution	$N(0, 10)$	$N(2, 10)\ C(0, \lambda_{SS})$					
Shortcut Solution	$N(0, 10)$	$N(2, 10)\ C(0, \lambda_{SS})$					
Wrong Numerator	$N(0, 10)$		$N(2, 10)\ C(0,)$				
Saw As Mixed Number	$N(0, 10)$			$N(2, 10)\ C(0,)$			
Counted Hash Marks	$N(0, 10)$				$N(2, 10)\ C(0,)$		
Counted Hash Marks and Posts	$N(0, 10)$				$N(2, 10)\ C(0,)$		
Saw As One Unit	$N(0, 10)$					$N(2, 10)\ C(0,)$	
Saw As Wholes	$N(0, 10)$					$N(2, 10)\ C(0,)$	
Saw As One Unit and Counted Hash Marks	$N(0, 10)$				$N(2, 10)\ C(0,)$	$N(2, 10)\ C(0,)$	
Saw As One Unit and Counted Hash Marks and Posts	$N(0, 10)$				$N(2, 10)\ C(0,)$	$N(2, 10)\ C(0,)$	
Everything In Order	$N(0, 10)$						$N(2, 10)\ C(0,)$
Unknown Error							

Note. For all the columns except that for the location parameter, cells with nonzero entries indicate the discrimination is included in the model.

These entries specify the prior distribution for the (included) discriminations. For Standard Solution, the prior distribution for the discrimination parameter (along the latent variable for the targeted skill) is $N(2, 10)$ censored below by 0. This prior is relatively diffuse, and serves to resolve the possible label switching of the latent variable (Chung, Loken, & Schafer, 2004); likewise for the prior distributions for the discrimination parameters on the other latent variables.

The prior distributions for the discrimination parameters for the other solution categories (along the targeted skill) are also $N(2, 10)$ censored below by 0, but are additionally censored above by the just-mentioned discrimination for Standard Solution. This additional constraint reflects the theory that the Standard Solution is most reflective of mastery of the skill.

Finally, for all intercept parameters included in the model we employ the following diffuse normal prior distributions

$$\lambda_{jk0} \sim N(0, 10). \quad (6)$$

Transition Component

The transition component specifies the probability distribution for subsequent values of the latent variables. Each latent variable is modeled as being dependent on the corresponding latent variable at the preceding time and the observable variable at the preceding time. That is, we define a transition structure for each latent variable based on the previous value of the latent variable and the just observed value of the observable variable. This is only done for latent variables that are measured by the level in question, and this is only done for observable categories that occur; if an observable category does not occur for a level, no transition probability is specified for that level.

Let θ_{imjt} denote the value of latent variable m for student i on level j at time t, and let X_{ijt} denote the value of the observable for person i on level j at time t. We begin by considering the model for each skill, and we set the probability of mastery at time t+1 given mastery at time t to be 1, regardless of the level and value of the observable at time t:

$$P(\theta_{imj(t+1)} = 1 \mid \theta_{imjt} = 1, X_{ijt}) = P(\theta_{imj(t+1)} = 1 \mid \theta_{imjt} = 1) = 1 \quad \forall j \forall t\,. \quad (7)$$

This reflects the hypothesis that mastery, once attained, cannot be lost. This is a firm constraint, and calls for some discussion about its implications. Mastery here takes on the meaning of success with levels that measure the skill in the game. This is not to say that if the student is a master of a skill in the game, then they are master outside of the game. The latent variables are interpreted in the context of the game and its modality. The probability distribution for a student's latent variable for Whole Numbers captures our understanding of their performance on the levels of the game that reflect that aspect of proficiency; likewise the probability distribution

for the latent variable for Partitioning Error captures our understanding of the student's performance and our thinking about their propensity to exhibit behaviors consistent with partitioning errors in the game. Of course, our desire is for the interpretation of performance in the game to generalize to other contexts. The use of a rigorous design process lends support to this, but further empirical work relating the characterization of student performance in the game to performance outside of the game would be beneficial (Delacruz et al., 2010; Kerr & Chung, 2012b).

In modeling the transition from nonmastery to mastery, in general we wish to specify the probability that a student is a master at time t+1 given they were a nonmaster at time t and had a particular value for the observable at time t. For example, what is the probability that a student who was not a master and made a partitioning error on the previous attempt is now a master?

In the current application, we encounter several problems related to the sparseness of data in attempting to estimate these probabilities. These challenges are likely to be present in other educational game environments. To begin, certain observations are somewhat rare. For example, in the dataset used to fit the model, in Level 3 there were 36 occasions where a student had a value of Saw As One Unit and Counted Hash Marks and Posts. Even if we assume that the student was a nonmaster of Whole Numbers on each of these occasions, it still does not afford a lot of data with which to estimate the transition probability. This may be thought of as a problem of too little data. A related problem occurs if the data suggests that the transition probability is 0. In this case it is unclear if the data from a larger sample would imply the transition probability is 0. If we had a larger sample, might we find someone who does transition to mastery? This problem is akin to that of sampling vs. structural 0s in contingency table analyses, and may be thought of as a problem of empirical sparseness. Importantly, the use of a fully Bayesian framework with prior distributions for parameters helps to mitigate these problems.

There is a third problem, which we term a logical problem of sparseness, related to solutions on the last level of the game that measures a latent variable. Consider again Level 3, which is the last level that depends on the latent variable for Whole Numbers. Once a student provides a viable solution to Level 3, they proceed on to the remaining levels, none of which measure the latent variable for Whole Numbers. Thus, there is no subsequent data on which to base the estimate of the transition from nonmastery to mastery of Whole Numbers when students provide viable solutions. Here again, prior specifications can resolve this. For each solution strategy, we can simply set the probability of transition from nonmastery to mastery at a certain value. This may be viewed as specifying a prior distribution with all its mass at a particular point. Alternatively we can set a prior distribution with a density over a region, expressing uncertainty. In these cases the resulting posterior will be identical to the prior.

We opt for the related but slightly different approach of specifying the transition probabilities via a hierarchical prior construction. For each level *j*, the probability of mastery for the latent variable *m*, here the latent variable for the targeted aspect of proficiency, at time *t*+1 given nonmastery at time *t* and the observable at time *t* had a value of *k* is

$$P(\theta_{imj(t+1)} = 1 \mid \theta_{imjt} = 0, X_{ijt} = k) \sim Beta(\alpha_{\theta_m=0,k}, \beta_{\theta_m=0,k}). \quad (8)$$

where $\alpha_{\theta_m=0,k}$ and $\beta_{\theta_m=0,k}$ are hyperparameters. Note that the subscript of θ_m=0 indicates that this is for the case where the value of latent variable *m* is 1, that is, the student does not possesses the skill. Additionally note that they are not indexed by *j*, indicating that these are not level-specific, but rather apply to all levels that measure the latent variable for the skill in question. Prior distributions for these hyperparameters are specified as

$$\alpha_{\theta_m=0,k} - 1 \sim Poisson(1); \quad (9)$$

$$\beta_{\theta_m=0,k} - 1 \sim Poisson(9); \quad (10)$$

The use of these Poisson priors and the "subtract 1" construction defines a Beta distribution in (8) that is not U-shaped and additionally reflects the hypothesis that transitions from nonmastery to mastery are not likely. The specification in (9)-(10) expresses the belief that the transition probability is most likely low, around .10, but that there is considerable uncertainty. With this hierarchical construction, the information in the data regarding the transition probabilities (in (8)) from the levels *before* the last level that measures the targeted skill flows up to the parameters α_{mk} and β_{mk}, which gets synthesized with the fairly diffuse prior (in (9) and (10)) to yield a posterior distribution, which in turn flows down to the transition probabilities (again, (8)) for the last level that measures the target aspect of proficiency. For latent variables representing the targeted skill, this construction is used for each observable category that occurs on a level; if an observable category does not occur for a level, no transition probability is specified.

We specify the transition probabilities for the latent variables that measure misconceptions in a similar fashion, with some slight differences. As just mentioned, if an observable category does not occur for a level, no transition probability is specified.

We do not assume that if a student possesses a misconception then they always will. Indeed, the hope is that they will learn by simply playing the game. Accordingly, for latent variable *m* representing a misconception, we specify the probability that a student will retain the misconception given they just provided the Standard Solution as

$$P(\theta_{imj(t+1)} = 1 \mid \theta_{imjt} = 1, X_{ijt} = k) \sim Beta(\alpha_{\theta_m=1,g_j,k}, \beta_{\theta_m=1,g_j,k}), \; k = \text{Standard Solution} \quad (11)$$

where $\alpha_{\theta_m=1,g_j,k}$ and $\beta_{\theta_m=1,g_j,k}$ are hyperparameters. Note that the subscript of θ_m=1indicates that this is for the case where the value of latent variable *m* is 1, that is, the student possesses the misconception. In addition to being indexed by θ_m and *k*, they are indexed by g_j, which stands for the group that level *j* belongs to, where groups of levels are defined by the targeted skill. This implies that we will have a hierarchical prior structure for the transition probabilities for misconceptions for each group of levels separately. For example, we have a hierarchical prior structure for the transition probabilities for the latent variable for Partitioning Error in Levels 1-3, another hierarchical prior structure the transition probabilities for the latent variable for Partitioning Error in Levels 4-8, another hierarchical prior structure the transition probabilities for the latent variable for Partitioning Error in Levels 9-12, and so on (Note that this was also done for the latent variables for skills, but as the groups were defined by which levels measured the same skills, the additional notation was not needed.). Prior distributions for these hyperparameters are specified as

$$\alpha_{\theta_m=1,g_j,k} - 1 \sim Poisson(1),\ k = \text{Standard Solution}; \tag{12}$$

$$\beta_{\theta_m=1,g_j,k} - 1 \sim Poisson(1),\ k = \text{Standard Solution}. \tag{13}$$

These choices reflect considerable uncertainty regarding the transition probabilities.

For the remaining solution strategies (Fractional Solution, Alternate Solution, Incomplete Solution, Shortcut Solution), we specify a common transition probability with a hierarchical prior structure

$$P(\theta_{imj(t+1)} = 1 \mid \theta_{imjt} = 1, X_{ijt} = k) \sim Beta(\alpha_{\theta_m=1,g_j,k}, \beta_{\theta_m=1,g_j,k}),\ k = \text{Other Solution}; \tag{14}$$

$$\alpha_{\theta_m=1,g_j,k} - 1 \sim Poisson(1),\ k = \text{Other Solution}; \tag{15}$$

$$\beta_{\theta_m=1,g_j,k} - 1 \sim Poisson(1),\ k = \text{Other Solution}. \tag{16}$$

Similarly, for the error category or categories associated with the latent variable for the misconception we specify a distinct hierarchical prior distribution

$$P(\theta_{imj(t+1)} = 1 \mid \theta_{imjt} = 1, X_{ijt} = k) \sim Beta(\alpha_{\theta_m=1,g_j,k}, \beta_{\theta_m=1,g_j,k}), \tag{17}$$

k = Category(ies) associated with misconception captured by latent variable *m*;

$$\alpha_{\theta_m=1,g_j,k} - 1 \sim Poisson(1); \tag{18}$$

k = Category(ies) associated with misconception captured by latent variable *m*;

$$\beta_{\theta_m=1,g_j,k} - 1 \sim Poisson(1), \tag{19}$$

k = Category(ies) associated with misconception captured by latent variable *m*.

For the error categories not associated with the latent variable for the misconception we specify a distinct hierarchical prior distribution

$$P(\theta_{imj(t+1)} = 1 \mid \theta_{imjt} = 1, X_{ijt} = k) \sim Beta(\alpha_{\theta_m=1,g_j,k}, \beta_{\theta_m=1,g_j,k}), \quad (20)$$

k = Category(ies) associated with other errors;

$$\alpha_{\theta_m=1,g_j,k} - 1 \sim Poisson(1); \quad (21)$$

k = Category(ies) associated with other errors;

$$\beta_{\theta_m=1,g_j,k} - 1 \sim Poisson(1), \quad (22)$$

k = Category(ies) associated with other errors.

The preceding equations specify the architecture of the transition probability structure given the student possessed the misconception at the preceding time point. This is reinstantiated for the situation in which the student did not possess the misconception at the preceding time point; which amounts to the probability of acquiring the misconception. Formally,

$$P(\theta_{imj(t+1)} = 1 \mid \theta_{imjt} = 0, X_{ijt} = k) \sim Beta(\alpha_{\theta_m=0g_j,k}, \beta_{\theta_m=0,g_j,k}), \; k = \text{Standard Solution}; \quad (23)$$

$$\alpha_{\theta_m=0,g_j,k} - 1 \sim Poisson(1), \; k = \text{Standard Solution}; \quad (24)$$

$$\beta_{\theta_m=0,g_j,k} - 1 \sim Poisson(1), \; k = \text{Standard Solution}; \quad (25)$$

$$P(\theta_{imj(t+1)} = 1 \mid \theta_{imjt} = 0, X_{ijt} = k) \sim Beta(\alpha_{\theta_m=0,g_j,k}, \beta_{\theta_m=0,g_j,k}), \; k = \text{Other Solution}; \quad (26)$$

$$\alpha_{\theta_m=0,g_j,k} - 1 \sim Poisson(1), \; k = \text{Other Solution}; \quad (27)$$

$$\beta_{\theta_m=0,g_j,k} - 1 \sim Poisson(1), \; k = \text{Other Solution}; \quad (28)$$

$$P(\theta_{imj(t+1)} = 1 \mid \theta_{imjt} = 0, X_{ijt} = k) \sim Beta(\alpha_{\theta_m=0,g_j,k}, \beta_{\theta_m=0,g_j,k}), \quad (29)$$

k = Category(ies) associated with misconception captured by latent variable m;

$$\alpha_{\theta_m=0,g_j,k} - 1 \sim Poisson(1), \quad (30)$$

k = Category(ies) associated with misconception captured by latent variable m;

$$\beta_{\theta_m=0,g_j,k} - 1 \sim Poisson(1), \quad (31)$$

k = Category(ies) associated with misconception captured by latent variable m;

$$P(\theta_{imj(t+1)} = 1 \mid \theta_{imjt} = 0, X_{ijt} = k) \sim Beta(\alpha_{\theta_m=0,g_j,k}, \beta_{\theta_m=0,g_j,k}), \quad (32)$$

k = Category(ies) associated with other errors;

$$\alpha_{\theta_m=0,g_j,k} - 1 \sim Poisson(1); \quad (33)$$

k = Category(ies) associated with other errors;

$$\beta_{\theta_m=0,g_j,k} - 1 \sim Poisson(1), \quad (34)$$

k = Category(ies) associated with other errors.

Results of Model Fitting

The model was calibrated using Markov chain Monte Carlo estimation, via the OpenBUGS software (Lunn, Spiegelhalter, Thomas, & Best, 2009), using data from 851 6^{th}-8^{th} graders playing *Save Patch*. Student attempts were recorded in log files and evaluated in ways described by Kerr and Chung (2012a). Attempts to fit the entire model in one analysis were unsuccessful, as the software crashed. The model was fit by running separate analyses for groups of levels. With one exception the groups were defined by levels corresponding to the same target skill (Table 1). The exception was that unsuccessful attempts at fitting the model for Levels 4-8 and led to them being split into Levels 4-6 in one analysis and Levels 7-8 in another analysis.

To preserve continuity between the separate analyses, for parameters that carried over one group of levels to another, the posterior distribution from the former were used as prior distributions in the latter. For example, for each student, the posterior distribution for the latent variable for Partitioning Error after Level 3 may be viewed as a Bernoulli distribution with a certain parameter, denoted here as $p_{Partitioning_i|Level\,3}$. Empirically, $p_{Partitioning_i|Level\,3}$ is estimated as the proportion of draws from the MCMC process that the latent variable for Partitioning Error after Level 3 for student *i* takes on corresponding to possessing the misconception. The prior distribution for the student's latent variable for Partitioning Error in Level 4 is then specified as a Bernoulli distribution with parameter $p_{Partitioning_i|Level\,3}$. More generally, to ensure continuity across separate analyses, the posterior probability for the parameter from the preceding analysis serves as the prior probability for the parameter in the following analysis.

For each analysis, two chains were run from dispersed start values. Convergence of the chain was evaluated via inspection of the trace plots. Once it appeared that the chains converged, an additional 10,000 iterations were obtained for use in summarizing the posterior distribution. Density plots and summary statistics of the parameters were inspected for interpretability. In all cases, there was no evidence of label switching within or between chains, suggesting the prior specifications were sufficient for resolving the indeterminacies inherent in the use of discrete latent variables.

To illustrate the results, Tables 4-6 contain estimates of the conditional probability tables for Level 19 based on the posterior means from the analysis in BUGS. Table 4 is the conditional probability table for the observable, given the combinations of the targeted skill (Adding Unit Fractions) and the misconception associated with Iterating Error. The highest probability of Standard Solution occurs when the student is a master of Adding Unit Fractions. If the student possesses the Iterating Error misconception, the probability of a solution is lower, and the probability of a Wrong Numerator attempt increases. Table 5 is the conditional probability of

mastery for Adding Unit Fractions at time *t*+1 given Adding Unit Fractions at time *t* and the observable at time *t*. Nonmasters are most likely to become masters following giving the Standard Solution (probability = .38), and least likely following an Unknown Error (probability = .09). Table 6 is the conditional probability of mastery for Iterating Error at time *t*+1 given Iterating Error at time *t* and the observable at time *t*. Students who possess the misconception associated with Iterating Error are fairly likely to keep the misconception even if they correctly solve the Level (probability = .51). They are somewhat less likely to keep the misconception if they have it and just exhibited Wrong Numerator on the previous attempt. The probabilities of acquiring the error if they do not possess it are generally low.

Table 4

Conditional Probabilities for the Observable for Level 19

Latent variables		Observable for Level 19				
Adding unit fractions	Iterating error	Standard Solution	Alternate Solution	Incomplete Solution	Wrong Numerator	Unknown Error
Master	Not Possess	0.95	0.00	0.01	0.03	0.01
Nonmaster	Not Possess	0.58	0.02	0.01	0.25	0.13
Master	Possess	0.77	0.00	0.01	0.21	0.01
Nonmaster	Possess	0.33	0.01	0.01	0.58	0.07

Table 5

Transition Probabilities for Adding Unit Fractions for Level 19

Adding unit fractions at time *t*	Observable for Level 19 at time *t*				
	Standard Solution	Alternate Solution	Incomplete Solution	Wrong Numerator	Unknown Error
Master	1	1	1	1	1
Nonmaster	.38	.17	.19	.20	.09

Table 6

Transition Probabilities for Iterating Error for Level 19

Iterating error at time *t*	Observable for Level 19 at time *t*				
	Standard Solution	Alternate Solution	Incomplete Solution	Wrong Numerator	Unknown Error
Possess	.51	.51	.51	.29	.46
Not possess	0	0	0	.15	.14

Model-Based Reasoning of Student Proficiencies and Misconceptions

This section details the use of the model for facilitating inferences about students. In a Bayesian network (BN), once values of variables are known, they can be entered and their information propagated throughout the network to yield posterior distributions for unknown variables (Pearl, 1988). This updating is fast, particularly when the size of the network is of moderate size and complexity.

Thus, BNs are attractive for psychometric models with latent variables, particularly when the network updating and propagation can be localized. This is indeed possible in the psychometric model introduced here. In psychometric applications, making inferences for students requires entering known values for observables, then propagating that information throughout the network to yield a posterior distribution for the latent variables and any as-of-yet unknown observables.

To facilitate the exposition of this process for the current model, recall the overall structure of the model depicted in Figure 1. As the structure and procedures apply for each student, we drop the subscript *i*. Suppose a value for the observable at the current time, $\mathbf{X}_t$, is observed. The posterior distribution for the remaining variables is

$$\begin{aligned} P(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}, X_{t+1}, | X_t) &\propto P(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}, X_t, X_{t+1}) \\ &\propto P(X_t \mid \boldsymbol{\theta}_t)\, P(\boldsymbol{\theta}_t)\, P(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t, X_t)\, P(X_{t+1} \mid \boldsymbol{\theta}_{t+1}) \\ &\propto P(\boldsymbol{\theta}_t \mid X_t)\, P(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t, X_t)\, P(X_{t+1} \mid \boldsymbol{\theta}_{t+1}) \end{aligned} \tag{35}$$

The second line in (35) follows from the conditional independence assumptions implied by the structure of the model in Figure 1. The third line results by recognizing that the first two terms on the right side of (35) constitute the posterior distribution for $\boldsymbol{\theta}_t$ given X_t.

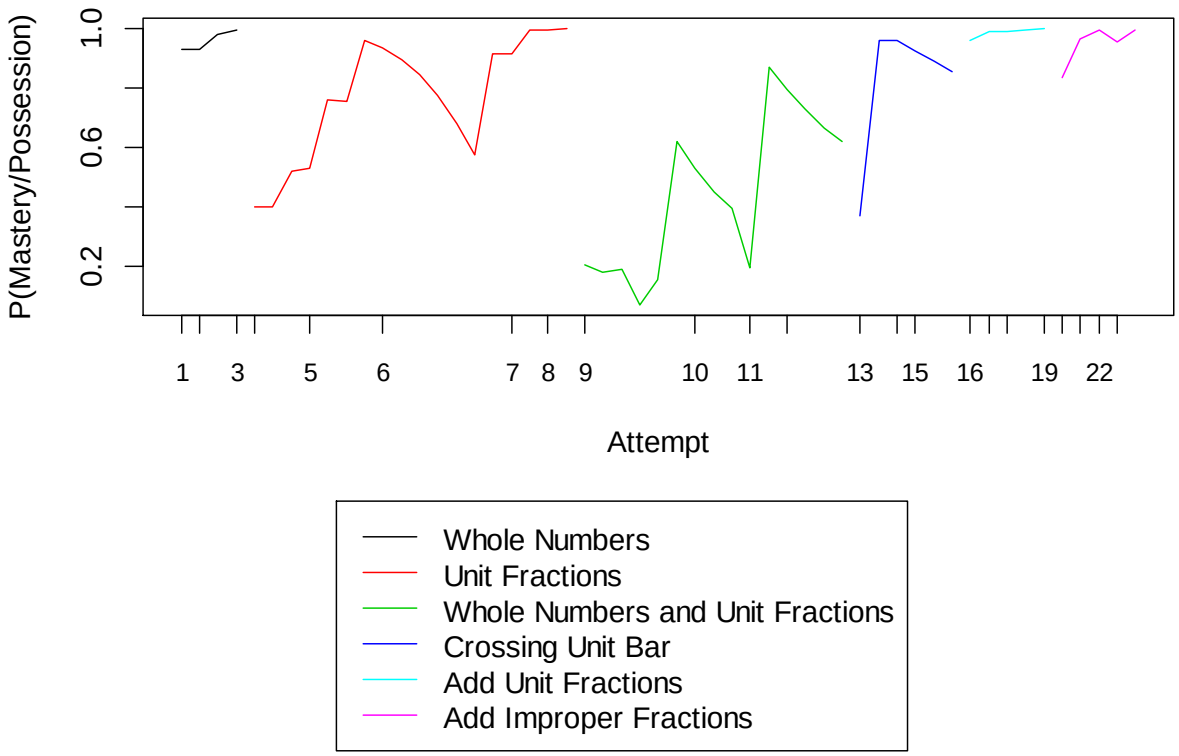

Figure 2. Running history of mastery of skills for a student.

This last factorization supports a multi-phase updating scheme (Reye, 2004). In the first phase, we obtain the posterior distribution for the latent variables immediately prior to the observation via Bayes Theorem

$$P(\boldsymbol{\theta}_t \mid X_t) \propto P(X_t \mid \boldsymbol{\theta}_t)\, P(\boldsymbol{\theta}_t), \tag{36}$$

where $P(\boldsymbol{\theta}_t)$ is the distribution for the latent variables prior to observing X_t, and $P(X_t \mid \boldsymbol{\theta}_t)$ is the measurement model, here the SMAC model. This represents the updated beliefs about the student's proficiency prior to the attempt, where the measurement model $P(X_t \mid \boldsymbol{\theta}_t)$ governs the revision to our beliefs about $\boldsymbol{\theta}_t$.

In the second phase, we obtain the model-based expectations for the latent variables at the next time point. This is given by the posterior predictive distribution for the latent variables $\boldsymbol{\theta}_{t+1}$ given the observable, obtained by marginalizing over the posterior distribution for $\boldsymbol{\theta}_t$

$$P(\boldsymbol{\theta}_{t+1} \mid X_t) = \sum_{\boldsymbol{\theta}_t} P(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t, X_t) P(\boldsymbol{\theta}_t \mid X_t). \tag{37}$$

This distribution represents the updated beliefs about the student's proficiency subsequent to the attempt, where the transition model $P(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t, X_t)$ captures our beliefs about how student proficiency changes in light of their previous proficiency state and performance.

These two phases represent the core parts of inference: given an observation of student performance, update beliefs about their proficiency prior to the attempt (phase one), and then update beliefs about their proficiency subsequent to the attempt (phase two).

A third phase is possible, in which we obtain the model-based expectations for performance at the next time point. This is given by the posterior predictive distribution for X_{t+1}

$$P(X_{t+1} \mid X_t) = \sum_{\boldsymbol{\theta}_{t+1}} P(X_{t+1} \mid \boldsymbol{\theta}_{t+1}) P(\boldsymbol{\theta}_{t+1} \mid X_t). \tag{38}$$

This distribution represents the updated beliefs about the student's performance on the next attempt, given the just updated beliefs about student proficiency for the next attempt. Note the role of the measurement model $P(X_{t+1} \mid \boldsymbol{\theta}_{t+1})$. Here, reasoning proceeds from the latent variables to the observables, which constitutes deductive reasoning through the measurement model. In contrast, reasoning from the observables to the latent variables (in (36)) constitutes inductive reasoning through the measurement model (Mislevy, 1994).

The procedure proposed here allows for fast, local computations that capitalize on the conditional independence assumptions inherent in the graphical model to work with local computations based on one observable at a time, without requiring for the BN to be expanded. The decomposition of the propagation of evidence supports a strategy where only a few distributions need to be stored. At any time point, we have the probability distribution for the latent variables. Once a value for an observable is known, this distribution serves as the prior distribution, $P(\boldsymbol{\theta}_t)$ in Bayes Theorem, which in conjunction with the measurement model $P(X_t \mid \boldsymbol{\theta}_t)$ yields the posterior distribution for the latent variables at this time, $P(\boldsymbol{\theta}_t \mid X_t)$. In turn, we employ the transition probability structure $P(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t, X_t)$ to obtain the posterior predictive distribution for the latent variables at the following time, $P(\boldsymbol{\theta}_{t+1} \mid X_t)$. This then serves as the prior distribution in the next analysis, once additional data are observed. Thus at any point, only a few structures need to be stored/employed: the current distribution of the latent variables, the conditional distribution of the observable (i.e., the measurement model), and the conditional distribution of latent variables in the future (i.e., the transition model). Previous distributions of the latent variable need not be maintained, but can be written out, creating a running history of beliefs about student proficiency.

Importantly, the network for each student can be built on fly. We begin with a probability distribution for each of the latent variables. The BN fragments for the measurement model and the transition model for the level are docked to the latent variables (Almond & Mislevy, 1999). The probability distribution for the latent variables is updated via the two-phase updating scheme

in (36) and (37). Once values for observables are known and their evidentiary implications propagated, they can be discarded (Almond & Mislevy, 1999). The BN fragments for the level are dropped from the model and the BN fragments for the measurement model and transition model for the next observable are docked.

The fully Bayesian framework and OpenBUGS software is advantageous for specifying and calibrating the model. However it is not optimal for conducting inference for students in using the procedures outlined above. To conduct inference, BN fragments corresponding to the prior probabilities for the latent variables, conditional probabilities for the observables, and the transition probabilities for the latent variables were specified using the posterior means for these parameters from OpenBUGS. The gRain package in R (Højsgaard, 2012) was used to conduct inference. As described above, each computation involved the current probability distribution for observables, the conditional probability distribution for the level of the observable at hand, and the transition probability for the latent variables for the level.

A running history of the latent variables for each of the 851 students in the calibration dataset was obtained. To illustrate, Figures 2 and 3 depict the running histories of probability of mastery for the latent variables representing the targeted aspects of proficiency and possessing a misconception for one student. The points are spaced equally for each attempt along the horizontal axis, but the axis is labeled by level, indicating when the student first attempted that level. We can interpret the gaps between the tick marks on the axis as representing the number of attempts spent on each level. Figures 2-3 reveal the student took several attempts on Levels 6, 9, and 12, and relatively few attempts on Levels 19-23. The rises and falls in the trajectories depict the change in the probability distributions for the latent variables that occur as the data arrive and the distributions are updated. This reflects the changing beliefs about the student, based on the model, updated as new information arrives in the form of new observations from student attempts on the levels. Figure 2 depicts the change in beliefs over time regarding the student's mastery of the targeted skills, likewise Figure 3 for beliefs regarding the student's possessing certain misconceptions. We see that the student struggled with Level 6, though that is not attributed to possessing a misconception, as none of the latent variables for the misconceptions have high probabilities. In this case, on the student's first six attempts at Level 6, they committed an Unknown Error. On their seventh attempt, they provided the Standard Solution. In contrast, when the student attempted Levels 9 and 10, they exhibited behaviors consistent with having a misconception associated with Partitioning Errors, as well a strategy of Avoiding Math by putting everything in order.

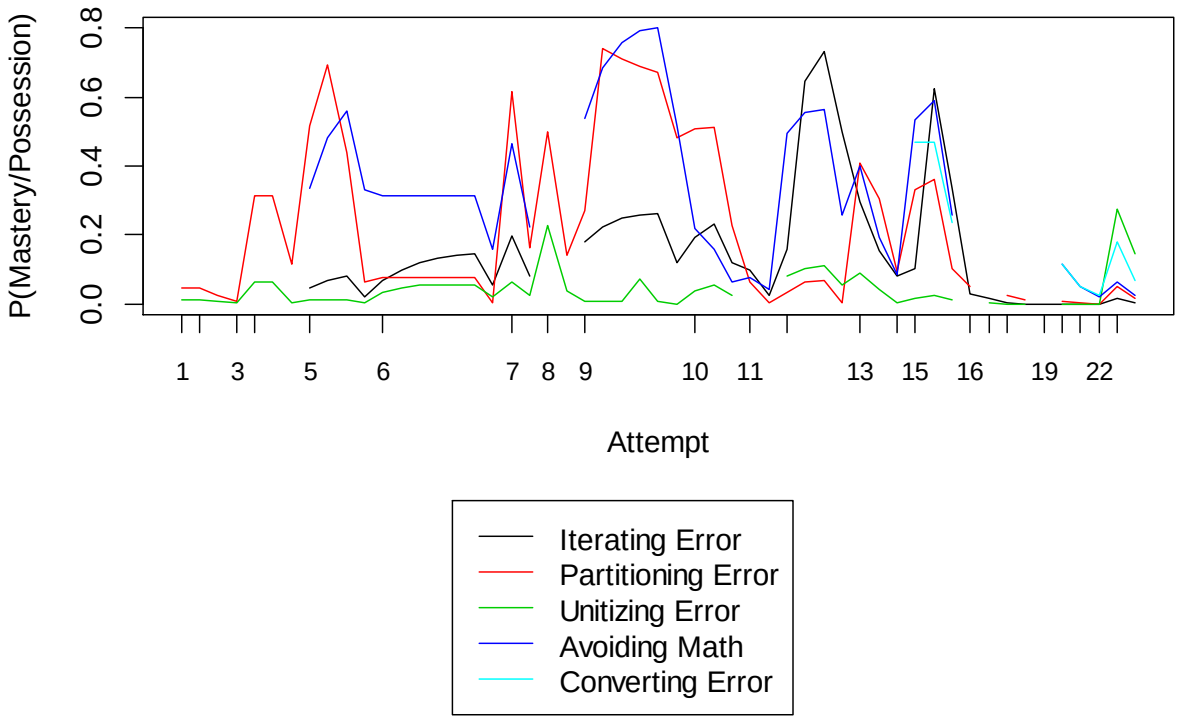

Figure 3. Running history of possession of misconceptions for a student.

Summary and Discussion

This work demonstrated the construction, calibration, and use of a DBN psychometric model for game-based assessment. The structure of the model, in terms of the specification of observable variables, the use of latent variables, and the dependencies of the former on the latter, was grounded on prior work that characterizes the evidentiary bearing of student performance on inferences about their capabilities, strategies, and misconceptions. This prior work came in the form of a principled design process for the game (Chung et al., 2010) as well as empirical analyses primarily in the form of cluster analyses and interpretations of log files (Kerr & Chung, 2012a; Kerr, Chung, & Iseli, 2011).

The judgments and interpretations on which the current developments are based may be in error; as with all statistical models, the current one is a simplification of the more complex real world situation. The best we can hope for is that the model is useful, which is less likely to be the case when the grounding and simplifying assumptions poorly reflect the real world situation at hand. In such a case, inferences based on the model, including characterizations of tasks and students, may be suspect. Critiquing the BN model (e.g., Sinharay, 2006), and perhaps some of the underlying assumptions regarding student proficiency and performance, is left for future work.

The current work focused on the construction of the model based off of the current theoretical and empirical grounding, followed by an empirical approach to calibrating the model. Specifically, a Bayesian approach to modeling facilitated the estimation of parameters based on student performance data as well as subject matter expertise. The context for this work, *Save Patch*, was a game that targets multiple skills and misconceptions, with student performances that can be characterized polytomously. To facilitate a reduced parameterization of the conditional probability tables, a diagnostic SMAC measurement model that used a parameterization similar to the SICM model was developed. The procedures developed here may support a more nuanced view of students based on data from *Save Patch* than more descriptive summaries, which may then be leveraged in studies that relate *Save Patch* to other assessments or measures of learning (Delacruz et al., 2010; Kerr & Chung, 2012b). Extensions to this work include the use of other condensation rules for the latent variables in the measurement component (e.g., Almond et al., 2001; Levy & Mislevy, 2004; Mislevy et al., 2002), modeling the relationships among the latent variables (e.g., de la Torre & Douglas, 2004; Levy & Mislevy, 2004). In addition, the modeling framework adopted here could be expanded by allowing for time-varying structures in the measurement component, transition component, or both.

References

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). Hoboken, NJ: Wiley.

Almond, R. G., DiBello, L., Jenkins, F., Mislevy, R. J., Senturk, D., Steinberg, L. S., & Yan, D. (2001). Models for conditional probability tables in educational assessment. In T. J. Jaakkola & Richardson (Eds.), *Artificial intelligence and statistics 2001* (pp. 137–143). San Francisco, CA: Morgan Kaufmann.

Almond, R. G., DiBello, L. V., Moulder, B., & Zapata-Rivera, J. D. (2007). Modeling diagnostic assessments with Bayesian networks. *Journal of Educational Measurement*, *44*(4), 341–359.

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

Almond, R. G., Mulder, J., Hemat, L. A., & Yan, D. (2009). Bayesian network models for local dependence among observable outcome variables. *Journal of Educational and Behavioral Statistics*, *34*(4), 491–521.

Almond, R. G., Williamson, D. M., Mislevy, R. J., & Yan, D. (in press). *Bayes nets in educational assessment*. New York, NY: Springer.

Baker, R. S. J. D., Pardos, Z., Gowda, S., Nooraei, B., & Heffernan, N. (2011). Ensembling predictions of student knowledge within intelligent tutoring systems. *Proceedings of the 19th International Conference on User Modeling, Adaptation, and Personalization,* 13-24.

Behrens, J. T., Frezzo, D. C., Mislevy, R. J., Kroopnick, M., & Wise, D. (2008). Structural, functional, and semiotic symmetries in simulation-based games and assessments. In E. Baker, J. Dickieson, W. W Wulfeck, and H.F. O'Neill (Eds.), *Assessment of problem solving using simulations* (pp. 59-80). New York: Routledge.

Bradshaw, L., & Templin, J. (in press). Combining item response theory and diagnostic classification models: A psychometric model for scaling ability and diagnosing misconceptions. *Psychometrika*.

Center for Advanced Technology in Schools (2012). *CATS-developed games*. (CRESST Resource Paper No. 15). Los Angeles: National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA. Retrieved from http://www.cse.ucla.edu/products/resource/cresst_resource15.pdf

Chang, K.-m., Beck, J., Mostow, J., & Corbett, A. (2006). A Bayes net toolkit for student modeling in intelligent tutoring systems. In K. Ashley & M. Ikeda (Eds.), *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*. Berlin: Springer-Verlag.

Chung, G. K. W. K., Baker, E. L., Vendlinski, T. P., Buschang, R. E., Delacruz, G. C., Michiuye, J. K., & Bittick, S. J. (2010, April). Testing instructional design variations in a prototype math game. In R. Atkinson (Chair), *Current perspectives from three national R&D centers focused on game-based learning: Issues in learning, instruction, assessment, and game design*. Structured poster session at the annual meeting of the American Educational Research Association, Denver, CO.

Chung, H., Loken, E., & Schafer, J. L. (2004). Difficulties in drawing inferences with finite-mixture models. *The American Statistician, 58,* 152-158.

de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69,* 333-353.

Delacruz, G. C., Chung, G. K. W. K., & Baker, E. L. (2010). *Validity evidence for games as assessment environments* (CRESST Research Report No. 773). Los Angeles: National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA. Retrieved from http://www.cse.ucla.edu/products/reports/R773.pdf

Højsgaard, S. (2012). Graphical independence networks with the gRain package for R. *Journal of Statistical Software, 46,* 1-26. URL http://www.jstatsoft.org/v46/i10/

Iseli, M. R., Koenig, A. D., Lee, J. J., & Wainess, R. (2010). *Automatic assessment of complex task performance in games and simulations* (CRESST Research Report No. 775). Los Angeles: National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA. Retrieved from http://www.cse.ucla.edu/products/reports/R775.pdf

Kerr, D., & Chung, G. K. W. K. (2012a). Identifying key features of student performance in educational video games and simulations through cluster analysis. *Journal of Eductional Data Mining, 4,* 144-182.

Kerr, D., & Chung, G. K. W. K. (2012b). *The mediation effect of in-game performance between prior knowledge and posttest score* (CRESST Research Report No. 819). Los Angeles: National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA. Retrieved from http://www.cse.ucla.edu/products/reports/R819.pdf

Kerr, D., Chung, G. K. W. K., & Iseli, M. R. (2011). *The feasibility of using cluster analysis to examine log data from educational video games* (CRESST Research Report No. 790). Los Angeles: National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA. Retrieved from http://www.cse.ucla.edu/products/reports/R790.pdf

Levy, R. (2013). Psychometric and evidentiary advances, opportunities, and challenges for simulation-based assessment. *Educational Assessment, 18,* 182-207.

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

Lunn, D., Spiegelhalter, D., Thomas, A., Best, N. (2009). The BUGS project: Evolution, critique, and future directions. *Statistics in Medicine, 28,* 3049-3067.

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

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

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 No. 580). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

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. (2009). *Multidimensional item response theory*. New York, NY: Springer.

Reye, J. (2004). Student modeling based on belief networks. *International Journal of Artificial Intelligence in Education, 14*, 1–33.

Rowe, J. P., & Lester, J. C. (2010). Modeling user knowledge with dynamic Bayesian networks in interactive narrative environments. In G.M. Youngblood & V. Bulitko (Eds.), *Proceedings of the sixth AAAI conference on artificial intelligence and interactive digital entertainment, AIIDE 2010*. Retrieved from: http://aaai.org/ocs/index.php/AIIDE/AIIDE10/paper/view/2149

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guilford Press.

Sao Pedro, M. A., Baker, R. S. J. D., Gobert, J. D., Montalvo, O., & Nakama, A. (2013). Leveraging machine-learned detectors of systematic inquiry behavior to estimate and predict transfer of inquiry skill. *User Modeling and User-Adapted Interaction, 23,* 1-39

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. In S. Tobias & J.D. Fletcher (Eds.), *Computer games and instruction* (pp. 503-524). Charlotte, NC: Information Age Publishers.

Sinharay, S. (2006). Model diagnostics for Bayesian networks. *Journal of Educational and Behavioral Statistics, 31,* 1-33.

VanLehn, K. (2008). Intelligent tutoring systems for continuous, embedded assessment. In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 113–138). New York, NY: Erlbaum.

VanLehn, K., & Niu, Z. (2001). Bayesian student modeling, user interfaces and feedback: A sensitivity analysis. *International Journal of Artificial Intelligence in Education*, *12*(2), 154–184.

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (RR-05-16). Princeton, NJ: Educational Testing Service.