International Journal of Statistics and Probability; Vol. 6, No. 3; May 2017
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Hybrid Measurement Models for Technology-Enhanced Assessments Through mIRT-bayes

Kathleen Scalise[1]

[1] Department of Methodology, Policy & Leadership, University of Oregon, Eugene, OR, USA

Correspondence: Kathleen Scalise, Department of Methodology, Policy & Leadership, University of Oregon, Eugene, OR, 97403 USA. Tel: 1-542-346-0893. E-mail: kscalise@uoregon.edu

Received: March 21, 2017 Accepted: April 7, 2017 Online Published: April 23, 2017

doi:10.5539/ijsp.v6n3p168 URL: https://doi.org/10.5539/ijsp.v6n3p168

Abstract

Technology-enhanced assessments (TEAs) are rapidly emerging in educational measurement. In contexts such as simulation and gaming, a common challenge is handling complex streams of information, for which new statistical innovations are needed that can provide high quality proficiency estimates for the psychometrics of complex TEAs. Often in educational assessments with formal measurement models, latent variable models such as item response theory (IRT) are used to generate proficiency estimates from evidence elicited. Such robust techniques have become a foundation of educational assessment, when models fit. Another less common approach to compile evidence is through Bayesian networks, which represent a set of random variables and their conditional dependencies via a directed acyclic graph. Network approaches can be much more flexibly designed for complex assessment tasks and are often preferred by task developers, for technology-enhanced settings. However, the Bayesian network-based statistical models often are difficult to validate and to gauge the stability and accuracy, since the models make assumptions regarding conditional dependencies that are difficult to test. Here a new measurement model family, mIRT-bayes, is proposed to gain advantages of both latent variable models and network techniques combined through hybridization. Specifically, the technique described here embeds small Bayesian networks within an overarching multidimensional IRT model (mIRT), preserving the flexibility for task design while retaining the robust statistical properties of latent variable methods. Applied to simulation-based data from Harvard's Virtual Performance Assessments (VPA), the results of the new model show acceptable fit for the overarching mIRT model, along with reduction of the standard error of measurement through the embedded Bayesian networks, compared to use of mIRT alone. Overall for respondents, a finer grain-size of inference is made possible without additional testing time or scoring resources, showing potentially promise for this family of new hybrid models.

Keywords: technology, statistics, technology-enhanced assessment, learning analytics, data science, Bayesian network, Bayes net, IRT, item response models, hybrid models, virtual performance assessment, digital literacy, simulation, probability and proficiency estimates, gaming, measurement models

1. Introduction

1.1 Challenges of New Data Science for Complex Technology-enhanced Tasks

Technology today offers many new opportunities for innovation in educational assessment through rich new tasks (Cayton-Hodges et al, 2012; Haertel et al, 2012; Scalise & Gifford, 2006, 2008), potentially powerful automated scoring (Rosen & Foltz, 2014), innovative reporting (Ainley, Fraillon, Schulz, & Gebhardt, 2014), and real-time feedback mechanisms (Timms, 2016).

However, the dramatic innovations in TEA constructs, observations and scoring come along with interpretation challenges of statistically aggregating scores from the new TEA assessment instruments (Scalise, 2012; Wilson et al, 2012). New statistical models that can provide high quality proficiency estimates for the psychometrics of complex TEA contexts are needed (Scalise, 2014; Wilson et al, 2012). Here, a hybrid mIRT-bayes two-stage modeling approach is introduced and explored through statistical application in a simulation-based Harvard University TEA context (Scalise & Clarke-Midura, 2014).

1.2 Features of TEAs

Data accumulation, or in other words measurement modeling, challenges for complex TEAs are myriad (Scalise & Gifford, 2006; Timms, 2016; Wilson, Scalise, & Gochyyev, 2015; Wilson, Scalise, & Gochyyev, 2016). This paper addresses one concern: Often some important observables, whether process-oriented or content-related, are directly scorable by assessment developers, while other salient data present as "semi-amorphous," or less immediately interpretable in TEAs. In the TEA semi-amorphous situation, any particular observable may not be especially meaningful alone but is key with other data in observing a pattern, or trend over a set of observables. This paper explores approaches to effectively accumulating semi-amorphous TEA data with new models.

Semi-amorphous data include a wide range of possibilities such as click streams of mouse movements in a digital setting, patterns of resource use such as podcasts or vodcasts, or alternating waves of dialog in a chat stream. Semi-amorphous data could also include selected or constructed content responses (SR or CR), such as a choice made by the respondent (SR) or an explanation provided (CR), if the result represents a range of reasoning facets rather than a clearly scorable answer with a correct response, as in a traditional assessment question. Such a range of ways to elicit evidence is advancing what is possible for learning analytics (Papamitsiou & Economides, 2014), especially in hard-to-measure constructs (Baker & Siemens, 2014; Barton & Schultz, 2012; Scalise, 2012).

However, while semi-amorphous data in TEAs do represent a set of observables, they have some characteristics differing from items in traditional ”item bundles” or testlets in the prior measurement literature (Care, Griffin, Scoular, Awwal & Zoanetti, 2015; Scalise & Wilson, 2007). While sharing issues of dependency well explored in the literature (Li, Bolt, & Fu, 2006; Rosenbaum, 1988; Wainer & Kiely, 1987; Wilson & Adams, 1995), traditional bundle structures typically involve strongly organized and predesigned item groups, with well-defined respondent interactions taking place. For instance, a typical traditional example of a testlet or item bundle might involve a reading passage or math problem-solving situation, with stimulus material shared across a set of items, and all respondents answering a sequential set of subsequent questions.

Semi-amorphous data sets, by contrast, are often not strongly structured in terms of the respondent's pattern of interaction. This is not to say that the observables are not well aligned to the construct. On the contrary, because of their very complexity, TEAs for formal measurement are usually robustly designed based on evidence-centered practices, domain modeling, cognitive task analysis and other often intensive means of construct alignment. But, however, semi-amorphous data are specifically designed such that each examinee has substantial self-selected response differentiation choice (Scalise, 2007) for any particular indicator. Thus, saliency to the construct resides in interpretation over a pattern. For instance in the example used here, which will be introduced fully later in this paper, the semi-amorphous data involves whether respondents did or did not take the opportunity to observe a variety of frogs and water samples in a science simulation educational assessment.

For any particular frog or water sample, there could be many reasons why the simulation artifact was observed or not from a given respondent. It could be that the simulation player virtually wandering about in the scenario didn't happen to encounter the location of that particular frog. Or for a specific water sample, the respondent might have spoken to a farmer ”simulated agent” about it instead of making his or her own observation, feeling that he or she has gained sufficient information to move on in the limited timeframe.

So it is not the individual observable that matters. Rather saliency comes in the pattern of observation over the set of frogs and water samples, relative to the simulation goal. While this could be handled in the measurement design with approaches used in the past such as giving partial credit over the set of observables, using decision rules to award a score, or having substantial systematically missing data, there are drawbacks to these approaches. For instance, as the semi-amorphous opportunities become more complex, it becomes very hard for learning scientists designing the TEAs to provide decision rules to score all the possible permutations of patterns, in the complex TEA tasks. There becomes a tendency to reduce the relevant information or discard it altogether for ease of use.

This increases test taking time for respondents as more data then has to be collected to provide a reliable score given that so much useful but semi-amorphous information is discarded as hard to model. Furthermore, certain types of information, e.g. the ”semi-amorphous” data, are systematically discarded and replaced with more direct responses. This may not as well represent the construct, potentially skewing the estimates, and certainly does not take full advantage of the TEA medium. The same is true for partial credit scores, whereby learning scientist find that after accumulating more than a few pieces of salient but semi-amorphous TEA data the myriad permutations of patterns become overwhelming.

Treating the observations as if they were a set of sparsely populated but systematically missing indicators, with any one indicator able to stand on its own, is not a good representation of the conceptual meaning of the TEA data. Any observation is not intended to be a sole indicator. Since it is only the pattern over them that is meaningful, then the pattern is the intended unit of analysis, not the observable, for semi-amorphous data (there may be many other observables in the TEA that are intended to be individual indicators though).

Many artifacts in the item analysis potentially can result from ignoring the intended structure of the data, including misfit for observables treated as items, sparseness for well-estimating item parameters, and lack of stability in parameters.

The end result is that relevant information is lost in many TEAs if potentially meaningful patterns are reduced more than necessary, or when eliminated entirely because the salient semi-amorphous data is deemed too difficult to include in the statistical estimation of the measurement model.

1.3 Relevant Scholarship

Here, a measurement model is considered to be defined as a mathematical model that aggregates or accumulates data from individual indicators or observables to make an inference (Wilson, 2005). The mIRT-bayes two-stage measurement process examined here nests information from small Bayes net structures within an IRT model.

Using Bayes nets within IRT, or conversely IRT within Bayes nets, has been proposed as a potential solution in educational measurement for complex assessments to capture advantages and mitigate disadvantages of each approach for some time (Mislevy, 2001). But little research is available exemplifying the possibilities. The advent of complex TEA is stimulating new interest, so the new mIRT-bayes modeling approach is proposed here.

The research literature has well-described both of these two model families separately previously (e.g., not in hybridized form). For the overarching IRT model to be used here, the Multidimensional Random Coefficients Multinomial Logit Model (MRCML), presented previously in many other publications, is applied (Adams, Wilson, & Wang, 1997; Wang, 1995; Wang, Wilson, & Adams, 1997). Since it generates estimates for respondents on more than one dimension, the MRCML is a multidimensional IRT model, or mIRT model.

The multidimensional form of the MRCML model is a direct extension of the RCML unidimensional model (Adams & Wilson, 1996). Both the RCML and the MRCML have been employed extensively in student assessment for both large scale and classroom applications in numerous applications previously, as summarized by Wilson (2005).

Formally, letting the latent traits define a two-dimensional latent space with the dimensions allowed to be non-orthogonal, the respondent's position in the space is represented by the vector,

$$\boldsymbol{\Theta} = (\Theta_1, \Theta_2) \tag{1}$$

Using notation from the MRCML sources described above and from developers of the software used in the calibration (Wu, Adams, Wilson, & Haldane, 2007), proficiency estimates are generated according to the MRCML as:

$$P(X_{ik} = 1; \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\xi}|\boldsymbol{\Theta}) = \frac{exp(\boldsymbol{b}_{ik}\boldsymbol{\Theta} + a'_{ik}\xi)}{\sum_{k=0}^{K_i} exp(\boldsymbol{b}_{ik}\boldsymbol{\Theta} + a'_{ik}\xi)} \tag{2}$$

with I items indexed $i = 1, \ldots, I$; and each item having $K_i + 1$ possible response alternatives, while $k = 0, 1, \ldots, K_i$.

Items are modeled through a vector,

$$\boldsymbol{\xi} = (\xi_1, \xi_2, ...\xi_P) \tag{3}$$

of P parameters. For a description of linear combinations in which the parameters are estimated for the model, see the description of the MRCML model (Adams, Wilson, & Wang, 1997; Wang, 1995; Wang, Wilson, & Adams, 1997) and the software employed (Wu, Adams, Wilson, & Haldane, 2007).

This multidimensional extension assumes two or more latent traits are to be estimated from the data (Briggs & Wilson, 2003). In this case there will be two traits estimated, scientific inquiry and scientific explanation, from the MRCML statistical modeling of the individuals' responses. Theta estimates form the MRCML are provided as the student proficiency estimates on each dimension, which are generated for each dimension for each student, according to the MRCML model applied here through the ConQuest software package (Wu, Adams, Wilson, & Haldane, 2007).

For this paper, the statistical innovation is to embed Bayesian networks, or Bayes Nets, within the mIRT model during the MRCML estimation. The Bayes Nets results are embedded here as additional item scores in the mIRT model. So the Bayes Nets are essentially used here to provide a "score," or item-level valuing for the respondent, over a *pattern* of semi-amorphous data developed to be salient together as a pattern informing on the construct of interest, and added to the prior more explicit items already in the model. Hence, each semi-amorphous pattern becomes an item.

A range of automated scoring techniques have begun to be explored for TEA, for instance for collaborative problem tasks (Rosen & Foltz, 2014). mIRT-bayes is a new approach. A two-stage estimation is employed here although one-stage could be possible (see Discussion section 4.2 on choices for one-stage and two-stage statistical approaches for estimations).

Bayes Nets are also well understood models previously presented extensively in the research literature as stand-alone models and used for instance in classroom-based assessments (Almond, DiBello, Moulder, & Zapata-Rivera, 2007; Almond, Mulder, Hemat, & Yan, 2009). They have been used in widely ranging examples such as e-learning products that provide statistically modeled assessment evidence (Mislevy, 2001; Scalise, Timms, & Kennedy, 2009).

Bayes Nets are acyclic directed graph structures, represented with nodes and arcs, which are shown in the Results section in an example. To generate the joint probabilities in the network, a strong assumption is made in Bayesian Network approaches that P, the local probability distribution of variable X_i, is made conditional only on the value of the parent nodes connecting to a given node from above. This is a defining assumption of using Bayesian Networks. In this case, each Bayes Net B embedded within the mIRT model generates a single item score,which is added to the mIRT model as a score.

Formally, each Bayes Net consists of Graph $G = (V, E)$, specifying a set V of vertices or nodes, along with a set E of ordered pairs of nodes, called the edges of G. Each node X_i in V, with $V = X_i, \ldots, X_n$, represent a random variable, here in the examples with finitely many states, together with a joint distribution P such that:

$$P(X_i = x_1, ..., X_N = X_n) = \prod_{i=1}^{n} P(X_1 = x_1 | pa(X_1)) \tag{4}$$

Thus, P, the local probability distribution of variable X_i is made conditional only on the value of the parent nodes, $pa(X_i)$, as is always the case in Bayes Nets, since this is a defining assumption of using Bayesian Networks.

Note that this strong assumption employed by Bayes nets requires that the network structure be well validated as compared to alternative structures that might be proposed, which could potentially have quite different sets of direct parent nodes. Even if much of the network includes the same or similar nodes and is broadly but not specifically similar, then substantial difference in estimation of score for the pattern can result.

When assumptions well hold, however, the Bayes Nets give typically a sparser and more parsimonious structure than use of the multiplication rule alone for full nodal structures. Justification for the parent node arrangement of the network structure here in this paper is made in part by employing only very small Bayes Nets. These have been identified for a salient pattern relative to the construct by learning scientists employing techniques such as content analysis of student pattern results on prior data to generate the Bayes Networks, and to provide the relationships between parent and child nodes.

In TEA, Bayes nets alone are perceived by many content developers as desirable to use in technology-enhanced context because, conceptually, the networks are easy to implement and flexible. Content developers like the way Bayes Net software visually represents complex TEA assessment designs. They also like that TEAs can be made almost arbitrarily more complex to meet the perceived creative and content needs of the delivery innovators, simply by dropping in more nodes and arcs to the Bayes Net visualizations.

Disadvantages from a measurement perspective are, in a sense, the same. With Bayes nets, content developers often quickly try to implement quite complex structures with frequently hundreds or thousands of nodes and equally many associated probability tables. Together, these represent statistically the equivalent of very complex path diagrams, with numerous mediating variables hypothesized. Strong but often relatively untested assumptions describe how the variables all relate. Often it is quite easy for learning scientists to propose other alternate hypotheses for these network structures, that would organize the nodes somewhat or very differently. When many of the mediating variables in such large networks can be moved to new locations in the network equally acceptably to the learning scientists developing them, or nearly equally, this is a problem both for interpretation and for stability and reliability of the measurement results. Alternate node and arc positions represent radically different networks. This potentially leads to very different proficiency estimates for respondents simply by moving some of the nodes around.

This is especially the case when network structures grow beyond a small size to validate in comparison to the various alternatives. Large size network structures generate solutions that are difficult to validate due to the size of the outcome space when comparing the many alternative structures and hypotheses of other structures with so many nodes and arcs employed. This is the equivalent in other statistical techniques of validating extremely large and complex path diagrams. Learning scientists also find it hard to fully interpret the meaning of large networks, which can readily be created but not as easily understood. Especially their statistical implications often are difficult to grasp in the context of latent variables.

When the Bayes net structures are kept small and discrete, however, these issues are more ameliorated while the advantages of flexibility and representation are retained. Even though there potentially may be several small networks nested into a mIRT-bayes, each one can be handled on its own merit. Small network by small network, learning scientists are more able to understand and interpret them, and statisticians to validate the small network structures each separately. Furthermore, within a hybrid approach, the overarching mIRT model supplies a great deal of information about data fit and item analysis of the nested structures.

Disadvantages of the more standard operational IRT models, when used stand-alone and not in a hybridized model, are also evident for TEAs. The IRT models, which are latent variable models, often do not offer the flexibility and representation that many TEA designers seek for their innovative content. Additionally and critically for the topic of this paper, IRT often encounters difficulty with the inclusion of semi-amorphous TEA data. Used as item level data, indicators are likely to show inadequate model fit along with well-documented dependency issues (Li, Y., Bolt, D. M., & Fu, J. 2006; Rosenbaum, 1988; Scalise & Wilson, 2007; Wainer et al., 2006; Wainer & Kiely, 1987; Wilson & Adams, 1995). However, aggregating with traditional IRT testlet or item bundle models encounters the issues discussed previously in the introduction section, such as misrepresenting the intended data structure.

In mIRT-bayes, the intention is that Bayes-related inferences from semi-amorphous data patterns might be effectively entered into overarching IRT models. Advantages of IRT as an overarching model include, as Roy Levy (2012) described, that over six decades of research and application, measurement technology has matured to establish well-accepted procedures for important issues, which include calibration and estimation of a student overall score, reliability and precision information, test form creation, linking and equating, adaptive administrations, evaluating assumptions, checking data-model fit, differential functioning, and invariance. All of these remain in place as mature approaches within mIRT-bayes, due to the use of the overarching mIRT model hybridized into the approach.

According to Levy, other new approaches are in their infancy when it comes to their application as measurement models in larger assessment enterprises. Yet current IRT-based approaches used operationally may not always cover the full need of TEAs and new construct challenges. So hybridization routes seem important to investigate.

1.4 Hypotheses and Correspondence to the Research Design

The hypothesis investigated in this paper is whether mIRT models can be extended to better handle complex TEA data by embedding Bayes Nets results into the mIRT data aggregation and model calibration. It is further hypothesized that if the hybrid model fits acceptably, the semi-amorphous data previously not possible to include in the mIRT model alone might be successfully included. This could improve the measurement characteristics of the TEA instrument, specifically to amplify the information available from the complex technology-enhanced assessment, and reduce the standard errors in the proficiency estimates. If these hypotheses are supported, the mechanism of improvement is expected to be through allowing more information that is already captured by the TEA to be made available to include in the proficiency estimation model, which would have been eliminated previously.

The research design therefore is based on a comparative study of model fit statistics and proficiency estimate results, with and without the hybridization of the mIRT model with the Bayes Nets. First, in Exploration 1, model fit and results from applying the mIRT model alone for the data set used are investigated (see Methods section). Exploration 1 employs only the item response data appropriate to the mIRT model alone and not the semi-amorphous process data. Then, for Exploration 2, the same data set and procedures are used, but expanded with the semi-amorphous data patterns and with application of the mIRT-bayes model. Results are then compared to see if the hybridization has offered any advantages, such as described in the hypotheses above.

2. Method

This section describes how the study was conducted, including definitions of the variables, data set, and tasks used in the study, as well as summarizing the analytic methods employed.

2.1 Identify Subsections

Subsections below include descriptions of the data set and information on the procedures used in the study. This includes description of the assessments used and how they were delivered.

2.2 Data Set Characteristics

The Harvard Virtual Performance Assessment (VPA) and New Frog data set was used in this study. The data set consists of 1,986 cases with scores on 23 items and less than 1 percent missing data. Respondents were middle-school students working with Harvard University on the validation of the Virtual Performance Assessments. Students on average spent about 30 to 40 minutes taking the assessment, which contains in addition to the 23 response items also about 55 actions collected by the interface that the students may perform. Some of the actions are considered salient to the construct by the assessment developers, and others are less so or not at all. More information on the assessment is provided below.

2.3 Assessment Characteristics

The Harvard VPA project assesses middle school students' science inquiry and process skills. The VPAs are delivered using technology that ”has the look and feel of a videogame” (Dede & Clarke-Midura, 2011; Dede, Clarke-Midura, & Scalise, 2013). The New Frog VPA from which the data set was drawn for this study is a serious role-playing game in which students take on the identity of an avatar who is confronted with the dilemma of a six-legged frog appearing in a farming community. Acting as virtual scientists, students walk around the environment, explore, ”speak” to other characters such as virtual farmers in the region who respond with text via decision rules, gather data, and attempt to solve the scientific puzzles.

New Frog takes place in a virtual version of a village with four farms. In New Frog, students explore the problem of a frog with six legs. There are five causal factors to investigate in New Frog for the anatomical anomaly: pollution, pesticides, parasites, genetic mutation, or aliens.

The assessments are used without direct prior instruction on the tasks, and are meant to show a model of how to supplement for instance state examinations. Assessment data are captured by the environment as the student proceeds, including both logs of actions and activity completion associated with scores aligned to the measurement context. Data generated are exported and analyzed for research studies.

Regarding goals and objectives of measurement for the Harvard New Frog VPA, the assessments were aligned by Harvard with the then newly released Framework for K-12 Science Education as well as with the College Board Standards for College Success. The May 2012 public draft of the ”Next Generation Science Standards for Today’s Students and Tomorrow’s Workforce” was used for the alignment. The New Frog VPA uses, for instance, the context of 2012 MS.LS-GDRO Growth, Development, and Reproduction of Organisms (draft May 2012).

2.3.2 Measures and Covariates

The items in the New Frog data set and the two dimensions on which they load in in the explored models show in Table 1.

Table 1. Items and Dimensions for New Frog VPA Data Set

Response Items	Dimension
Control 1-3 (3 dichotomous)	1
Research 1-6 (6 dichotomous)	1
Experimentation 1-3 (3 dichotomous)	1
Sample 1-3 (3 dichotomous)	1
Claim (1 polytomous, rescored to dichotomous)	1 and 2
Farm (1 polytomous)	2
Backpack Contents (1 polytomous)	2
Q6-10 (5 polytomous items: Tadpole 6, Frog 7, Water 8, DNA 9, Blood Test 10)	2

Description: The two dimensions modeled in the New Frog Data set, and the items used.

2.3.3 Research Design

The research design is as described above: a comparative study of model fit statistics and proficiency estimate results, with and without the hybridization of the mIRT model with the Bayes Nets. Subjects were observed through data collection in a technology platform used in school settings. Use of the mIRT model alone is Exploration 1. Use of the hybridized mIRT-bayes model is Exploration 2.

3. Results

For Exploration 1, which is the mIRT alone analysis, first a unidimensional and then a multidimensional random coefficients multinomial logit model (MRCML; Adams, Wilson, & Wang, 1997) were fit to the 23-item data set using ACER ConQuest Generalised Item Response Modelling Software. A description of the scored items and two dimensions on which they load in the mIRT model show in Table 1 in the Methods section.

Exploration 2 examined whether accumulating additional semi-amorphous TEA information through a hybrid mIRT-bayes model, for a total of 26 scores, improved results.

3.1 Recruitment

Secondary data analysis of the de-identified data set was the focus of this study. Subjects had previously been recruited by Harvard University according to Harvard’s IRB protocols.

3.2 Statistics and Data Analysis

Exploration 1: mIRT Only. A comparison of the relative fit of the unidimensional and multidimensional models was first made. As the unidimensional model is a submodel of the two-dimensional model, the difference between the deviance of these two models is distributed as a chi-square with two degrees of freedom. The estimated deviance difference between the models of 4433.17 was highly significant ($p<.01$), so the hypothesis is rejected that the unidimensional model fits these data as well as did the two-dimensional model.

Based on statistical significance, the multidimensional model (mIRT) is preferred but there also is evidence of practical significance supporting the multidimensional model as well. As described previously, theoretical support for the two dimensions is based on the standards identified by the STEM content matter experts and learning scientists, and the intention to report on distinctly different skills in these two dimensions, as reflected in the standards table.

The multidimensional model will be used for the rest of Exploration 1. The reliability was high for the New Frog VPA on each dimension, with various indicators ranging from reporting a reliability of .82 to .89, and an overall Cronbach's alpha of .88. High reliability on both dimensions with the mIRT model helps give empirical justification for reporting separate student proficiency scores in each area.

There was also not a strong correlation between the two dimensions; rather the relationship between the two dimensions was only moderate, as estimated by the model. The correlation between the INQUIRY and EXPLANATION latent variables was 0.42. Note that this correlation is effectively already corrected in the ConQuest software for attenuation caused by measurement error. This is a much lower correlation between dimensions than found in many analyses of other science and engineering constructs and instruments using multidimensional models, where dimensions are often found to be very highly correlated, at the .8 level or higher. In STEM education, mIRT may be advocated even at correlations between dimensions as high as .7-.8 (Draney & Peres, 1998).

Item fit to the model was good, with only 4 percent (1 of 23 items) with estimated item difficulty parameters outside 3/4-4/3 mean square weighted fit (Wu, Adams, & Wilson, 1998), for parameters in which the weighted fit T was greater than 2. Even randomly, at the 95 percent confidence level, about 5 percent of item parameters could be expected to be less fitting. Note that this item parameter was also at an extreme high of the difficulty distribution, where fewer were located to provide data to estimate the parameter.

Step parameters for the multidimensional analysis were similar in their fit profile, with no estimated parameters falling outside the outside 3/4-4/3 mean square weighted fit. Itanal item discrimination was also promising at an average of .56 (SD = .10, min = .27, max = .72). No items were negatively discriminating.

Figure 1 is the "Wright Map," or graphical representation of the student ability distribution on each of the dimensions. The left panel shows a representation of the latent INQUIRY ability distribution, the middle panel shows the same for the EXPLANATION ability distribution, and the right panel indicates the difficulty of the items. Note that neither the unit nor origin are necessarily the same between calibrations unless the model is anchored, otherwise sufficiently constrained, or rescaled to match, which was not done here, so the two dimensions are not equated and should not be compared. However, a student's "X" location may be compared within each dimension.

The Wright Map is empirically generated by the analysis, and is based on the actual student data in the New Frog VPA data set. The two sets of X's in the left columns of the figure show vertical histograms (histogram turned on its side) of student performance for the sample group. Lower performing students appear with X's at the bottom of the map and higher performing with X's at the top of the map. There is one histogram per each dimension, with INQUIRY on the left and EXPLANATION on the right.

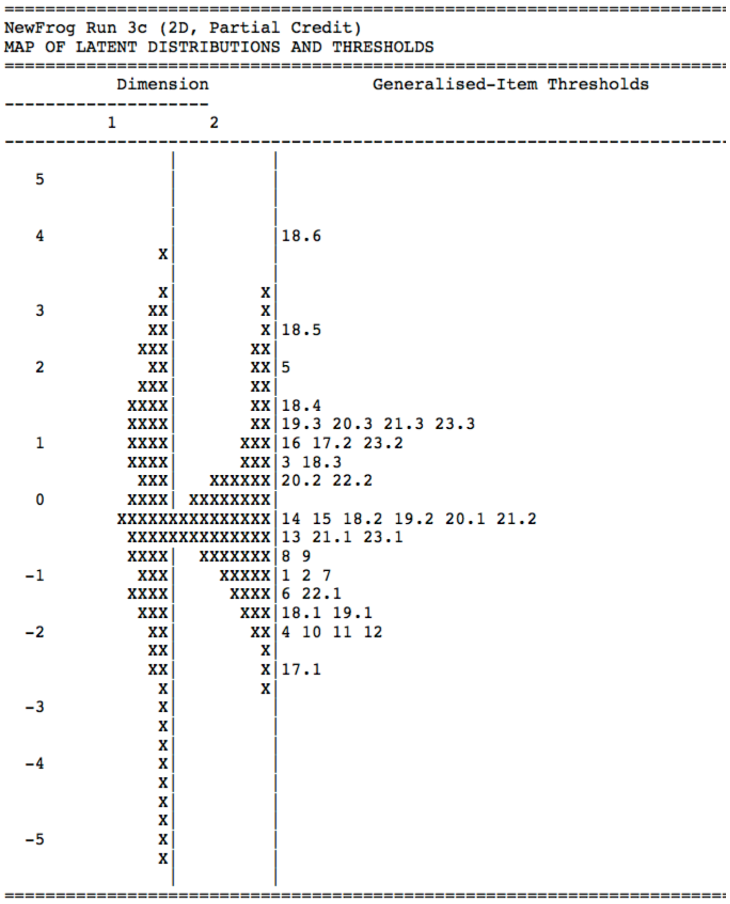
The farthest right column is a scale of the performance, showing the difficulty of achieving the various generalised item Thurstonian thresholds estimated from the data. Briefly stated, Thurstonian thresholds are plotted at the point on the display where a student falling adjacent has a 50 percent chance of achieving at least the indicated level of performance on an item, for that dimension (Wu, Adams, Wilson, & Haldane, 2007).

Using these important techniques of construct mapping (Wilson, 2005) with mIRT models can reveal a portrait of student readiness to learn, or area of the standards where students are actively constructing knowledge.

One major conclusion from the VPA results here is that students in the sample data set are distributed over a wide range of readiness to learn. Also, due to the only moderate correlation between dimensions, students often show different readiness that teachers will need to address in INQUIRY skills and knowledge mapped here as compared to EXPLAINING.

Construct mapping such as here, when models fit, estimates are reliable and a valid case has been made for the uses and purpose of the instruments, gives a "data" picture to teachers, subject matter experts, and learning scientists. It provides empirical evidence on both the construct/standards and where students stand. The next step in the process is for teachers and other experts to apply their knowledge of how to intervene with students to address growth and improvement. Readiness to learn is depicted on the Wright map; steadiness in making learning gains, or growth, is then addressed by teachers, scientists, science educators, schools, districts, states and others who are working with students. The end goal here is for all students to have the opportunity to learn and to achieve success in their efforts.

Exploration 2: mIRT-bayes Hybrid Model. For Exploration 2, the same New Frog data set was used but with information added though two Bayes net structures, Frog Observation and Tadpole Observation, incorporated to amplify the assess-



```
=======================================================================
NewFrog Run 3c (2D, Partial Credit)
MAP OF LATENT DISTRIBUTIONS AND THRESHOLDS
=======================================================================
           Dimension                Generalised-Item Thresholds
--------------------
          1         2
-----------------------------------------------------------------------
                |         |
   5            |         |
                |         |
                |         |
   4            |         |18.6
               X|         |
                |         |
               X|        X|
   3          XX|        X|
              XX|        X|18.5
             XXX|       XX|
   2          XX|       XX|5
             XXX|       XX|
            XXXX|       XX|18.4
            XXXX|       XX|19.3 20.3 21.3 23.3
   1        XXXX|      XXX|16 17.2 23.2
            XXXX|      XXX|3 18.3
             XXX|   XXXXXX|20.2 22.2
   0        XXXX| XXXXXXXX|
           XXXXXXXXXXXXXXX|14 15 18.2 19.2 20.1 21.2
            XXXXXXXXXXXXXX|13 21.1 23.1
            XXXX|  XXXXXXX|8 9
  -1         XXX|    XXXXX|1 2 7
            XXXX|     XXXX|6 22.1
             XXX|      XXX|18.1 19.1
  -2          XX|       XX|4 10 11 12
              XX|        X|
              XX|        X|17.1
               X|        X|
  -3           X|         |
               X|         |
               X|         |
  -4           X|         |
               X|         |
               X|         |
               X|         |
  -5           X|         |
               X|         |
                |         |
=======================================================================
```

Figure 1. New Frog Wright Map from two-dimensional IRT analysis

Description: Graphical representation of the student ability distribution on each of the dimensions. The left panel shows a representation of the latent INQUIRY ability distribution, the middle panel shows the same for the EXPLANATION ability distribution, and the right panel indicates the difficulty of the items.

ment information. This added two items to the original data set, since each Bayes Net explored a pattern and delivered a score for pattern of semi-amorphous data. A third item on six-legged frog observation was also added, see information on this below. To summarize, semi-amorphous information was added about whether respondents observed various frogs, tadpoles and water samples available in the simulation.

In parallel with Exploration 1, first a unidimensional and then a multidimensional random coefficients multinomial logit model (MRCML; Adams, Wilson, & Wang, 1997) were fit to the now 25-item data set, following the Bayes Net application to provide the scoring, using ACER ConQuest Generalised Item Response Modelling Software.

For the original nodes in the Bayes net structures, see Figure 2, frogs and tadpoles located at the ”yellow” farm, the ”blue” farm and the ”red” farm were employed, along with water evidence including observations of laboratory and red farm water samples. Each VPA participant could achieve a 0 (not observed) or 1 (observed) for inspecting each frog, tadpole and water sample, as well as a partial credit score on a water evidence question, which for the purpose of the network was dichotomized into low and high. The original network representation for Frog Observables shows in Figure 2 and the joint distribution tables in Figure 3. Blue nodes show the frogs and water samples possible to observe, green nodes show inferences to be made in the original network. The frog evidence node was queried for posterior probabilities and expectations.

First, for the New Frog network, posterior probabilities for the Frog Evidence node were calculated. For instance for frogs, a student could observe the yellow and red frogs but not the other evidence, for a given posterior probability.

Posterior distributions show how the given evidence gathering pattern is associated with probability of success on Frog Observable, see Figure 3. The result is called the ”belief” given the ”evidence.” Obtaining the belief given the evidence is

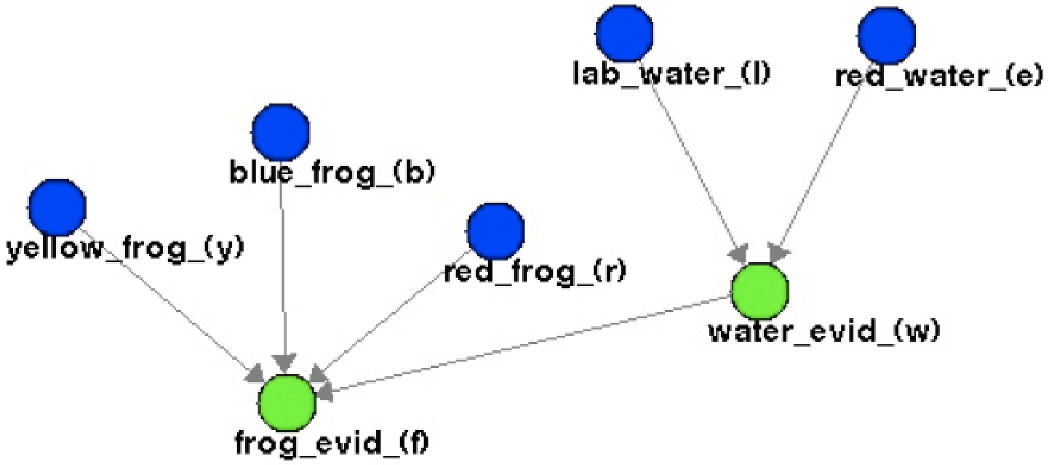


Figure 2. Frog Observable Bayes Net

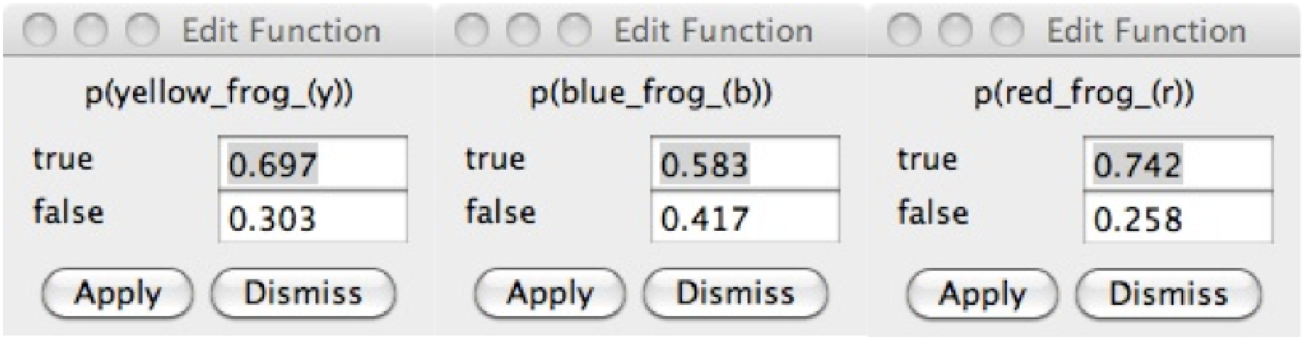
Description: Information is added to the mIRT-bayes Exploration 2 though the two Bayes net structures, Frog Observation and Tadpole Observation. An example is shown through the Frog Observable Bayes Net acyclic diagram in Figure 2. Exploration 2 uses the mIRT-bayes model to incorporate semi-amorphous process data from the TEA not previously possible to use in the mIRT model alone from Exploration 1.

called ”belief updating” and it can occur over a variety of semi-amorphous variables depending on the network structure.

A similar process generated posteriors for the Tadpole Observable network. After all patterns were generated and results were inspected, the observational water evidence (lab and red) was determined to add little useful information while the water evidence node did through the water evidence question directly, so the network was adjusted accordingly.

Posteriors and expectancies for the final network were calculated, and then merged into the original data set as two partial credit items to represent the Frog Observable and Tadpole Observable networks. As well one new dichotomous item was added to the data set tracking whether or the not the six-legged frog itself was observed. (This was not considered semi-amorphous data but possible to treat alone as a response item because the overall goal of the simulation was understanding the six-legged frog appearance in the farming community.) Consequently from the first stage of the mIRT-bayes process, three new items were added to the original 23 in the New Frog data set.

Some results for the mIRT-bayes analysis show in Table 2 and Figure 4. A major feature to notice under mIRT-bayes with the added item information from the network structures is that standard errors for the student estimates are improved over the prior mIRT-only run. As described previously, neither the unit nor origin are necessarily the same between calibrations



p(yellow_frog_(y))

true	0.697
false	0.303

p(blue_frog_(b))

true	0.583
false	0.417

p(red_frog_(r))

true	0.742
false	0.258

Figure 3. Empirical probability tables for the three Frog Observable nodes

Description: Example of Bayes Net approach used for the Frog Observable nodes.

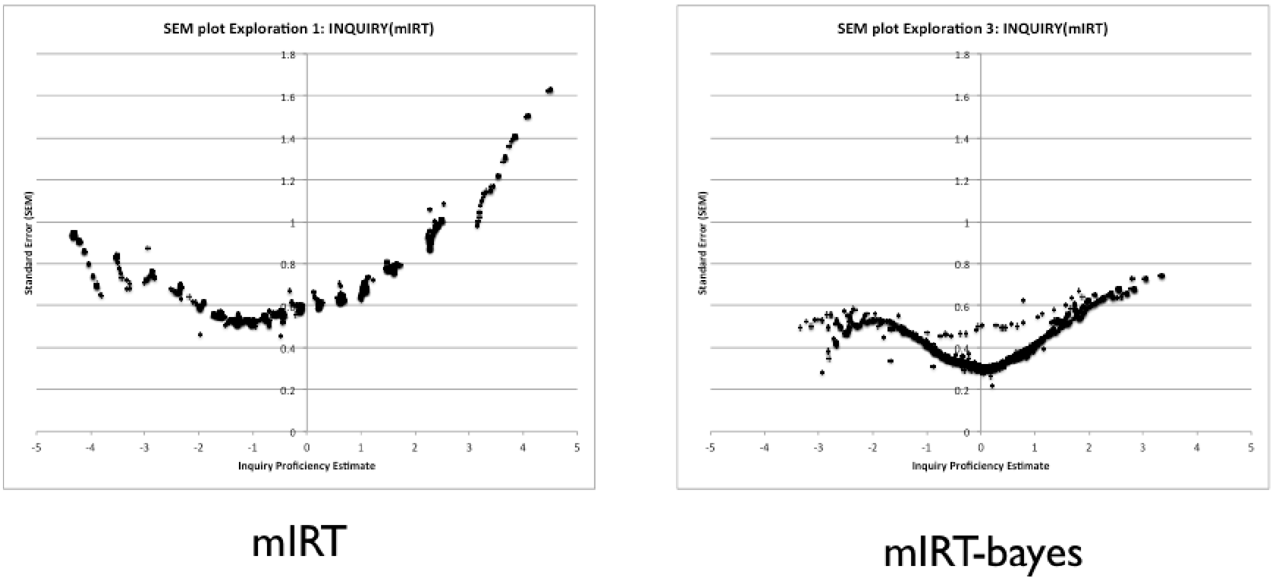


Figure 4. Standard error plots for the mIRT alone and mIRT-bayes investigations

Description: Comparison of some standard error results from Explorations 1 and 2. Previously under mIRT alone for this data set, a range of four average standard errors (4X bins) yielded approximately 3 unique proficiency levels of the 4X bins. Under mIRT-bayes, a range of four average standard errors exceeded 4 unique proficiency levels, indicating that mIRT-bayes may be useful in making finer grain inferences from the data collected in TEA while also reporting reliable scale scores.

unless the model is anchored, otherwise sufficiently constrained, or rescaled to match, which was not done here. However, the standard errors and grain size of inference relative to the overall estimate performance range can be compared. As an example, consider the INQUIRY scale. Previously under mIRT alone for this data set, a range of four average standard errors (4X bins) was 2.85 logits for the INQUIRY scale, yielding approximately 3 unique proficiency levels of the 4X bins. Under mIRT-bayes, a range of four average standard errors is only 1.66 logits for INQUIRY. Student estimates on INQUIRY ranged from a minimum of -3.33 logits to 3.35 logits, giving a range of 6.68 logits, so 4X bins (6.68 logits/1.66 logits) exceeded 4 unique proficiency levels.

Thus the mIRT-bayes approach effectively aggregates more of the available evidence in the technology-enhanced assessment, for this data set, and allows a finer grain size of inference, without requiring additional data to be collected or using more student or teacher time. The impact of improved measurement is most evident at the extremes, as can be seen in Figure 4, since this is where the least measurement information was available previously.

Reliability estimates under mIRT-Bayes are good for both dimensions, ranging from .83 to .89 with a Cronbach's alpha of .87, as shown in Table 2. As before, the missing data percentage remained less than 1 percent for the analysis reported here.

Regarding fit statistics, the original 23 items in A2 and A3 are slightly better fitting under the Exploration 2 mIRT-bayes model than previously under Exploration mIRT only. For instance, the parameter at the extreme high difficulty reported previously for A2 and A3 as falling slightly outside 3/4-4/3 mean square weighted fit no longer falls outside under the new model. This is perhaps a consequence of reducing the ceiling and floor effects of the test by lowering the standard errors at the extremes. In addition the new Frog Observable network and the 6-Legged Frog item showed good fit, within the tolerances described previously, with the Tadpole Observable slightly less well-fitting.

Given, as before, that up to 5 percent lesser fitting is expected in any calibration at the 95 percent confidence level, the model fitted here at 4 percent within expected tolerance remains good (1/26).

As before, under mIRT-bayes, itanal item discrimination remained at nearly the same average values as previously (.53, SD = .14, min = .24, max = .84). However the Frog Observable network at .81 exceeded the maximum item discrimination of .72 seen by any other item in the prior mIRT run. As well, Tadpole Bayes at .71 nearly exceeded all others, with the exception of only one item in the prior mIRT run.

Table 2. Some Results for New Frog VPA mIRT-bayes Exploration 2

Dimensions	INQUIRY/EXPLANATION
Cronbach's alpha overall	.87
MLE person separation reliability	
Dimension 1: INQUIRY	.86
Dimension 2: EXPLANATION	.89
Estimated aprior/person variance reliability (EAP/PV)	
Dimension 1: INQUIRY	.83
Dimension 2: EXPLANATION	.88

Description: Some mIRT-bayes Exploration 2 reliability estimates, overall and by dimension.

4. Discussion

Under mIRT-bayes model with the added item information from the network structures, standard errors for the student estimates are improved over the prior mIRT-only run. Thus the mIRT-bayes approach based on this data set effectively aggregates more of the available evidence in the technology-enhanced assessment, as predicted in the study hypotheses. Lowering the standard error allows a finer grain size of inference without requiring additional data to be collected or using more student or teacher time for the assessment.

Additionally, the Frog Observable network, composed of semi-amorphous data previously discarded, exceeded the maximum item discrimination of .72 seen by any other item in the prior mIRT only run. As well, Tadpole Bayes at .71 nearly exceeded all others, with the exception of only one item in the prior mIRT run. Thus the Bayes network additions to the mIRT model appear to be highly informative, and adding the semi-amorphous data is shown here to have good utility.

At the same time, additional data collection or hand-scoring was not required in any of the Bayes items to achieve the improved performance characteristics. All the observation evidence in both Frog Bayes and Tadpole Bayes had been previously captured but not used, and scoring was based on automated scoring.

4.1 Extensions: Exploring a One-Stage or Two-Stage Estimation Process

For the TEA field, whenever observables are statistically aggregated to make an inference, a measurement model is considered as being applied (Scalise, 2013). Since both stages of the mIRT-bayes process are accumulating data from sets of observables, the full approach is considered here to be a two-stage, or multi-stage, measurement model process.

The hybrid model process employs two major modeling approaches in the measurement field, IRT and Bayes nets. Therefore usual measurement standards of exploring data fit, validity, reliability and other high quality evidence concerns should be addressed for TEA applications, as briefly exemplified here.

In a one-stage process, nested models such as mIRT-bayes could also potentially be estimated. One approach could draw on Bayesian network models with continuous proficiency variables as equivalent to multidimensional IRT models (Almond, Mulder, Hemat, & Yan, 2009). Where such configurations apply, it could be possible to institute design matrices or other approaches through the estimation software that sufficiently constrain portions of the calibration, such that the nested directed acyclic graph portion of the model and the overarching IRT structure are calibrated simultaneously, meaning within one estimation.

There are advantages to both a one-stage and a two-stage process. As in typical in one-stage processes for a number of other statistical models and might be the case here, standard error estimations are often improved in a single stage. However, first, conceptually for learning scientists and assessment developers, it may be easier to think of the nested models separately, better preserving the flexibility desired of the representations. Not fully integrating the calibrations also may help preserve moving nested structures in and out of the overarching model. If structures can be preserved, modified or removed in a more modular fashion, this reflects the role of understanding those cases when semi-amorphous patterns are found to effectively contribute information and when they do not. Finally, many current software packages may have difficulty estimating models for a one-stage process at this time, while the two-stage process can readily be completed using a variety of familiar applications. This is called a ”low threshold” advantage in the technology community. Open access or widely available products for the learning science and developer communities mean there is a lower implementation threshold for a two-stage process at this time.

4.2 Extensions: Considering Size and Number of Nested Bayes Net Structures

For mIRT-bayes, small nested Bayes network structures of no more than four to ten nodes are suggested at this time for entry into mIRT-bayes.

On the lower bound, the aggregation permutations of three nodes or less can usually be readily specified by learning scientists or subject matter experts in TEAs using traditional decision rules or rubrics, so a more extensive model is not needed.

On the high end, much larger aggregations encompassing more than about 10 nodes become subject to the concerns of Bayes net alternate hypotheses discussed previously. Since large aggregations can readily be handled by the overarching mIRT model, there is less need to construct large nested networks within an embedded model. Also, key portions of larger structures that seem meaningful to learning scientists can be ”nodalized” or isolated from the large structure and form a new nesting within mIRT-bayes. At this time, there is no suggestion for how many total nested structures might be included; this is a subject for further investigation.

Note too that for the very small case of three nodes or less, probability estimates and ”beliefs” from prior data sets may be helpful to learning scientists in establishing their decision rules even if a network structure is not needed. So there remains a role for an empirical basis based on ”prior beliefs” at the smaller size as well.

4.3 Limitations of the Study

Limitations of this research study are of course extensive. Data from only one VPA task were examined in this study. Other data sets may indicate differently regarding use of the mIRT-bayes model for accumulating semi-amorphous process data.

The sample was based on an extant data set and employed de-identified secondary data analysis of the assessment data supplied by the Harvard project. Future work might look at samples intentionally designed for representative characteristics. Furthermore, information on incentives, compliance rates, and other specific settings that might have been involved in the study were not available, except for missing data rates directly in the data set.

Furthermore, only one mIRT model was employed in this study, since it was well-fitting to the data here. It was extended with the Bayes Nets. However, other mIRT models might be employed and should be investigated, including more parameterized models. Future work might explore anchoring well-fitting items according to the MRCML model here and treating any less well-fitting networks, if they arise, with a more parameterized IRT model.

The generalizability, or external validity, of the findings is also in question when an assessment is based on a single performance task. Student performance in this one context may not be representative of measuring a fuller science INQUIRY/EXPLORATION construct. Additional measures would need to be incorporated to consider such concerns as content validity and other assessment material sampling concerns over the educational framework. This was not the purpose of this modeling study, but would have important relevance in actual operational assessment settings.

However, incorporating more material is likely to extended the performance event beyond an acceptable assessment window for schools. So other designs such as repeated interim assessments might need to be considered, due to the length of time to measurement for performance assessments.

4.4 Conclusion

These results show the mIRT-bayes hybrid model may be promising to pursue further, based at least on performance with this data set. Operationally it appears to be more capable of distinguishing among student performances, although limitations are noted in the above section. The approach improved measurement results. In terms of evidence quality, this helps support that mIRT-bayes may be useful in making finer grain inferences while also reporting reliable scale scores aggregated at a larger grain size through the mIRT model.

Complex information streams and rich assessment designs have been emerging recently in technology-enhanced assessments. Increasingly these include semi-amorphous data, which can be content or process, such as click streams, patterns of resource use such as podcasts or vodcasts, alternating waves of dialog from live chats, and selected or constructed content responses that represent a range of reasoning facets.

For accumulating such data into well-supported inferences, results of using a hybrid mIRT-bayes measurement modeling approach are reported here. The mIRT-bayes two-stage measurement process introduced here nests information from small Bayes net structures within a Multidimensional Random Coefficients Multinomial Logit Model (MRCML) model (Adams et al., 1997b; Wang, 1995; Wang et al., 1997) although other models could be used. Nested within are small acyclic directed graph structures, shown here in examples of two layers and three layers; again other Bayes net structures could be used.

Conceptually, it is possible to generalize this hybrid to a wide variety of Bayes network structures and to a range of unidimensional, sequential unidimensional, or multidimensional IRT models, depending on the example.

Applied to a simulation-based data set from Harvard’s Virtual Performance Assessments, the first investigation using the mIRT model alone showed the viability of using this approach for aggregating data in the Harvard New Frog virtual performance assessment. The second investigation, using mIRT-bayes, showed additional improvement including reduction of the standard error of measurement compared to use of mIRT alone, for the simulation data set example. Especially when considering extremes of the proficiency scale, salient information difficult to include in the mIRT model directly was recouped through the nested Bayes network structures. At the same time, the two-stage approach preserved interpretability by keeping nested structures small and bounded within the IRT framework.

While a two-stage process is used here, extension to one-stage is also possible (Almond et al., 2009). However, there are advantages in flexibility of assessment development and scoring to keep the analytic process as two-stage.

More broadly, hybrid modeling approaches of nesting Bayes Nets within IRT, or conversely IRT with Bayes Nets, have been suggested as potentially complementary of each other for next-generation measurement models in complex technology-enhanced contexts. This is because strengths and limitations of the two approaches can tend to offset each other in demanding contexts. Together, or blended, they may offer potential, as shown here.

From this study of a hybrid approach, desirable attributes that characterize next-generation approaches to model hybridization include:

1. Borrowing strength from each other, for example, one measurement model within another or one model effectively extending another.

2. Establishing multiple inferential grain sizes.

3. Sometimes drawing on stronger confirmatory data from improved domain modeling.

4. Sometimes using richer range of exploratory data to explain more variance.

5. Functioning with much more data (e.g, big data), but also noisier data, while still being able to generate high quality evidence, often across large-scale and classroom-based practices if needed.

6. Allowing a wider range of constructs, observations, and interpretations for 21st century teaching and learning needs.

Acknowledgements

The research presented here and date set used were supported by the Bill and Melinda Gates Foundation, through the Harvard Virtual Performance Assessment (VPA) project, which designed and developed the VPA task.

References

Adams, R. J., & Wilson, M. (1996). Formulating the Rasch model as a mixed coefficients multinomial logit. In G. Engelhard & M. Wilson (Eds.), *Objective measurement III: Theory into practice*. Norwood, NJ: Ablex.

Adams, R. J., & Wilson, M. (1996). Formulating the Rasch model as a mixed coefficients multinomial logit. In G. Engelhard & M. Wilson (Eds.), *Objective measurement III: Theory into practice*. Norwood, NJ: Ablex.

Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The Multidimensional Random Coefficients Multinomial Logit Model. *Applied Psychological Measurement, 21*(1), 1-23. https://doi.org/10.1177/0146621697211001

Ainley, J., Fraillon, J., Schulz, W., & Gebhardt, E. (2014). Measuring Changes in ICT Literacy over Time. Paper presented at the National Council on Measurement in Education, Philadelphia, PA.

Almond, R. G., DiBello, L. V., Moulder, B., & Zapata-Rivera, J.-D. (2007). Modeling Diagnostic Assessments with Bayesian. *Journal of Educational Measurement, 44*(4), 341-359. https://doi.org/10.1111/j.1745-3984.2007.00043.x

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. https://doi.org/10.3102/1076998609332751

Baker, R. S., & Siemens, G. (2014). Educational data mining and learning analytics. In K. Sawyer (Ed.), *The Cambridge Handbook of the Learning Sciences* (pp. 253-274). https://doi.org/10.1017/cbo9781139519526.016

Barton, K., & Schultz, G. (2012). Using Technology to Assess Hard-to-Measure Constructs in the CCSS and to Expand Accessibility: English Language Arts. Paper presented at the Invitational Research Symposium on Technology Enhanced Assessments, Washington, DC. http://www.k12center.org/events/research_meetings/tea.html

Briggs, D., & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement, 4*(1), 87-100.

Care, E., Griffin, P., Scoular, C., Awwal, N., & Zoanetti, N. (2015). Collaborative Problem Solving Tasks. In P. Griffin & E. Care (Eds.), *Assessment and Teaching of 21st Century Skills* (pp. 85-104). Netherlands: Springer. https://doi.org/10.1007/978-94-017-9395-7_4

Cayton-Hodges, G. A., Marquez, E., van Rijn, P., Keehner, M., Laitusis, C., Zapata-Rivera, D., Bauer, M., & Hakkinen, M. T. (2012). Technology Enhanced Assessments in Mathematics and Beyond: Strengths, Challenges, and Future Directions. Paper presented at the Invitational Research Symposium on Technology Enhanced Assessments, Washington, DC. http://www.k12center.org/events/research_meetings/tea.html

Dede, C., & Clarke-Midura, J. (Producer). (2011, Nov. 3, 2012). Welcome to the VPA Project. Retrieved from http://vpa.gse.harvard.edu/

Dede, C., Clarke-Midura, J., Scalise, K. (2013). Virtual Performance Assessment and Games: Potential as Learning and Assessment Tools. Paper presented at the Invitational Research Symposium on Science Assessment, Washington, DC.

Draney, K., & Peres, D. (1998). Multidimensional modeling of complex science assessment data. Berkeley, CA: University of California, Berkeley. Retrieved from http://bearcenter.berkeley.edu/publications/multidim.pdf

Haertel, G. D., Cheng, B. H., Cameto, R., Fujii, R., Sanford, C., Rutstein, D., & Morrison, K. (2012). Design and Development of Technology-Enhanced Assessment Tasks: Integrating Evidence-Centered Design and Universal Design for Learning Frameworks to Assess Hard to Measure Science Constructs and Increase Student Accessibility. Paper presented at the Invitational Research Symposium on Technology Enhanced Assessments, Washington, DC. http://www.k12center.org/events/research_meetings/tea.html

Levy, R. (2012). Psychometric Advances, Opportunities, and Challenges for Simulation-Based Assessment. Paper presented at the Invitational Research Symposium on Technology Enhanced Assessments, Washington, DC. http://www.k12center.org/events/research_meetings/tea.html

Li, Y., Bolt, D. M. & Fu, J. (2006). A Comparison of Alternative Models for Testlets. *Applied Psychological Measurement, 30*(1), 3-21. https://doi.org/10.1177/0146621605275414

Mislevy, R. J. (2001). Modeling conditional probabilities in a complex assessment: An application of Bayesian modeling in a computer-based performance assessment Presented at the conference Cognition and Assessment: Theory to Practice, August 14-15, at the University of Maryland.

Papamitsiou, Z., & Economides, A. A. (2014). Learning Analytics and Educational Data Mining in Practice: A Systematic Literature Review of Empirical Evidence. *Educational Technology & Society, 17*(4), 49-64.

Rosen, Y., & Foltz, P. (2014). Assessing Collaborative Problem Solving through Automated Technologies. *Research and Practice in Technology Enhanced Learning, 9*(3), 389?410.

Rosenbaum, P. R. (1988). Item Bundles. *Psychometrika, 53*, 349-359. https://doi.org/10.1007/BF02294217

Scalise, K. (2007). Differentiated e-Learning: Five Approaches through Instructional Technology. *International Journal of Learning Technology, 3*(2), 169-182. https://doi.org/10.1504/IJLT.2007.014843

Scalise, K. (2012). Using Technology to Assess Hard-to-Measure Constructs in the CCSS and to Expand Accessibility. Invitational Research Symposium on Technology Enhanced Assessments, Educational Testing Service (ETS), Washington, DC. Retrieved from.
http://www.k12center.org/events/research_meetings/tea.html

Scalise, K. (2013). Innovation and Coherent Assessment: Views for Educational Measurement ETS Seminar. Princeton, NJ.

Scalise, K. (2014). Technology-Enhanced Assessments in NAEP, California Educational Research Association (CER-A) Conference, Building California's Future: Strategies for Achieving Coherence among Standards, Instruction, Assessment, and Evaluation.

Scalise, K., & Clarke-Midura, J. (2014). mIRT-bayes as Hybrid Measurement Model for Technology-Enhanced Assessments. Paper presented at the 2014 National Council on Measurement in Education, Philadelphia, PA.

Scalise, K., & Gifford, B. R. (2006). Computer-Based Assessment in E-Learning: A Framework for Constructing "Intermediate Constraint" Questions and Tasks for Technology Platforms. *Journal of Teaching, Learning and Assessment, 4*(6).

Scalise, K., & Gifford, B. R. (2008). Innovative Item Types: Intermediate Constraint Questions and Tasks for Computer-

Based Testing. Paper presented at the National Council on Measurement in Education (NCME), Session on "Building Adaptive and Other Computer-Based Tests", New York, NY.

Scalise, K., Timms, M. J., & Kennedy, C. (2009). Assessment for e-Learning: Case Studies of an Emerging Field. Paper presented at the Assessment & Teaching of 21st Century Skills First Annual Conference, San Diego, CA, San Diego, CA.

Scalise, K., & Wilson, M. (2007). Bundle Models for Computer Adaptive Testing in E-Learning Assessment. Paper presented at the 2007 GMAC Conference on Computerized Adaptive Testing (Graduate Management Admission Council), Minneapolis, MN.

Timms, M. (2016). Towards a model of how learners process feedback: A deeper look at learning. *Australian Journal of Education*, June. https://doi.org/10.1177/0004944116652912

Wainer, H., Brown, L., Bradlow, E., Wang, X., Skorupski, W. P., Boulet, J., & Mislevy, R. J. (2006). An Application of Testlet Response Theory in the Scoring of a Complex Certification Exam. In D. M. Williamson, I. J. Bejar & R. J. Mislevy (Eds.), *Automated Scoring of Complex Tasks in Computer Based Testing*. Mahway, NJ: Lawrence Erlbaum Associates, Inc.

Wainer, H., & Kiely, G. (1987). Item Clusters and Computerized Adaptive Testing: A Case for Testlets. *Journal of Educational Measurement, 24*, 185-202. https://doi.org/10.1111/j.1745-3984.1987.tb00274.x

Wang, W.-C. (1995). Implementation and application of the multidimensional random coefficients multinomial logit. Unpublished doctoral dissertation. University of California, Berkeley, California.

Wang, W.-C., Wilson, M., & Adams, R. (1997). Rasch models for multidimensionality between items and within items. In W. G. Englehard, Mark (Ed.), *Objective Measurement* (Vol. 4). Greenwich, CN: Ablex Publishing.

Wilson, M. (2005). *Constructing Measures: An Item Response Modeling Approach*. Mahwah, NJ: Lawrence Erlbaum Assoc.

Wilson, M., & Adams, R. J. (1995). Rasch Models for Item Bundles. *Psychometrika, 60*(2), 181-198. https://doi.org/10.1007/BF02301412

Wilson, M., Bejar, I., Scalise, K., Templin, J., Wiliam, D., & Torres Irribarra, D. (2012). Perspectives on Methodological Issues. In P. Griffin, B. McGaw & E. Care (Eds.), *Assessment and Teaching of 21st Century Skills*. Dordrecht; New York: Springer. https://doi.org/10.1007/978-94-007-2324-5_3

Wilson, M., Scalise, K., & Gochyyev, P. (2016). Assessment of Learning in Digital Interactive Social Networks: A Learning Analytics Approach. *Online Learning Journal, 20*(2). https://doi.org/10.24059/olj.v20i2.799

Wilson, M., Scalise, K., & Gochyyev, P. (2015). Rethinking ICT Literacy: From Computer Skills to Social Network Settings. *Thinking Skills and Creativity, 18*, 65-80. https://doi.org/10.1016/j.tsc.2015.05.001

Wu, M., Adams, R. J., & Wilson, M. (1998). The generalised Rasch model ACER ConQuest. Hawthorn, Australia: ACER.

Wu, M., Adams, R. J., Wilson, M., & Haldane, S. (2007). ACER ConQuest, Version 2.0, Generalised Item Response Modelling Software. Camberwell, Victoria: ACER Press.

Copyrights