Models for Conditional Probability Tables in Educational Assessment

R. G. Almond,[*] L. Dibello, F. Jenkins, D. Senturk,[†] R. J. Mislevy, L. S. Steinberg, D. Yan
Educational Testing Service, Princeton, NJ

Abstract

Experts in educational assessment can often identify the skills needed to provide a solution for a test item and which patterns of those skills produce better expected performance. The method described here combines judgements about the structure of the conditional probability table (e.g., conjunctive, or compensatory) with Item Response Theory methods for partial credit scoring (Samejima, 1969) to produce a conditional probability table or a prior distribution for a learning algorithm. The structural judgements induce a projection of each configuration of parent skill variables onto a single latent response-propensity θ. This is then used to calculate a probability for each cell in the table.

1 Introduction

In an ongoing educational assessment program, a large part of the work goes into bringing new tasks (or "items" or "problems") into the assessment. If test results are used for decisions that have a high perceived impact on the examinee, the problem is exacerbated by the need to retire tasks after a limited exposure. A large part of the work in producing a task for operational use is statistically calibrating it; that is calculating the "weights of evidence" that will be used to update beliefs about the latent abilities we wish to measure. In a Bayesian approach to testing, these weights are determined by the conditional probabilities of obtaining various observable features of the solution—i.e., states of observable variables—given the configuration of latent ability variables.

Obtaining numbers for these tables is in general hard work. We could learn them from pretest data, but the need for a sufficiently large sample size for every configuration of parents makes it expensive to get reliable values using a purely empirical model. Noisy-or and similar models use structural judgements about the conditional probability table to reduce the number of required parameters; Mislevy *et al.* (1999) apply this approach. However, even eliciting priors for a limited number of probabilities is difficult. When assigning conditional probability tables or priors for conditional probability tables, we would like to take advantage of the wealth of experience available from applications using the univariate Item Response Theory (IRT) model. To this end we take an approach based on structured latent class models (Formann, 1985), and extend a Bayesian framework for modeling IRT parameters in terms of expert judgment and collateral information about items (Mislevy, Sheehan, and Wingersky, 1983).

1.1 Graphical model of educational assessment

For student s, let $S_{s,1}, \ldots, S_{s,N}$ be a collection of variables measuring that student's knowledge, skills or abilities in some domain of interest. At any point in time, we represent our knowledge about that student's proficiency by a probability distribution. The prior distribution $\Pr(\mathbf{S}_s)$ is usually based on the distribution of these skills in the population of interest. We are interested in drawing inferences from $\Pr(\mathbf{S}_s|\mathbf{X}_s)$, where $\mathbf{X}_s = \{X_{s,1}, \ldots, X_{s,M}\}$ is the collection of observations made from the student's responses on a collection of M tasks. The student model variables $\mathbf{S}_s$ are purely latent; the observations and prior assumptions about the relationship between the observations and the student model variables are required for inference. Almond and Mislevy (1999) describe this general framework.

The case we address in this paper is the multivariate latent class model. Here all the student model variables $S_{s,n}$ are discrete. In this case, we can represent

[*]Corresponding Author: ralmond@ets.org
[†]U.C. Santa Barbara; Summer intern at ETS.

the distribution $\Pr(\mathbf{S}_s)$ with a Bayesian network. For the models developed below, it is useful to order the states in order of increasing level of proficiency.

If we knew $\Pr(\mathbf{X}_s|\mathbf{S}_s)$ we obviously could apply Bayes theorem to calculate $\Pr(\mathbf{S}_s|\mathbf{X}_s)$. Usually we assume that the observations from different tasks are conditionally independent given the student model variables (see Almond and Mislevy, 1999 for a discussion). What we need is a collection of *evidence models* $\Pr(X_{s,m}|\mathbf{S}_s)$. In most cases, $X_{s,m}$ is conditionally independent of all observation variables from other tasks and all but a subset of student-model variables. Thus, $\Pr(X_{s,m}|\mathbf{S}_s) = \Pr(X_{s,m}|\mathbf{S}_{s,m})$, where $\mathbf{S}_{s,m} \subset \mathbf{S}_s$. We call $\mathbf{S}_{s,m}$ the footprint of evidence model m.

In the Bayesian network case, the evidence model m is a conditional probability table. If we wish to elicit an unstructured prior for $\Pr(X_{s,m}|\mathbf{S}_{s,m})$, we must specify $|\mathbf{S}_{s,m}|$ Dirchlet distributions, where $|\mathbf{S}_{s,m}|$ is the size of the state space of the footprint of Task m. This can be a daunting task. For instance, there are about a hundred observable variables in the Biomass example discussed below, most with three possible values, many with size 18 footprints—over five thousand individual probabilities altogether. In the simple special case of IRT we have a long history of building evidence models. The goal of this paper is draw on that experience to formulate structured models for more ambitious structures for $\Pr(X_{s,m}|\mathbf{S}_{s,m})$.

1.2 Models for partial credit scoring

The most common IRT case posits a single continuous skill variable, θ_s, and binary response variables. The relationship between the observation and the skill variable is a logistic regression. This model is used for equating many well known college entrance exams, including the SAT, the GRE, and TOEFL. Various parameterizations are found in the literature. The two parameter logistic (2PL) model, for example, is

$$\operatorname{logit}\big(\Pr(X_{s,m}|\theta_s)\big) = a_m(\theta_s + b_m).$$

Samejima's graded response model (1969) extends this model to an observable $X_{s,m}$ that can take on one of the ordered values $x_0 \prec \cdots \prec x_K$. We define for $k = 1, \ldots, K$ we define:

$$\Pr(X_{s,m} \geq x_k|\theta_s) = P^*_k(\theta_s) = \operatorname{logit}^{-1}\big(a_m(\theta_s + b_{m,k})\big). \tag{1}$$

The category probabilities $\Pr(X_{s,m} = x_k|\theta_s)$ can be calculated from the differences of the cumulative probabilities given by equations (1). Figure 1 illustrates response category probabilities for a three-category task, with $a_m = 1$, $b_{m,1} = -1$, and $b_{m,1} = +1$. For very low values of θ, the lowest level of response is most likely, then as θ increases, probabilities increase for higher-valued responses in an orderly manner. A single value of θ specifies the full conditional distribution of all possible responses.

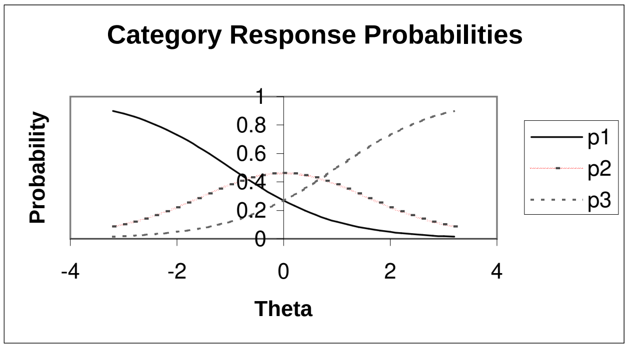

Figure 1: A graded-response IRT model.

2 The θ projection method

We are interested in finding models for $\Pr(X_{s,m}|\mathbf{S}_{s,m})$ in the case where $\Pr(\mathbf{S}_s)$ is a Bayesian network and $X_{s,m}$ is an ordered discrete variable. We employ the following device. First we pick a fixed set of values for a_m and $\mathbf{b}_m = \{b_{m,1}, \ldots, b_{m,K}\}$. Then we define a mapping function $f(\mathbf{S}_{s,m}) = \tilde{\theta}_{s,m}$. We can now apply Samejima's graded response model to fill out the tables.

We gain two advantages with this transformation of the problem. First, in the multivariate case our experts may be comfortable describing the functional form for $f_m(\cdot)$ even if they are uncomfortable with specifying a conditional probability table (e.g., "You have to know how to do A, but then you can solve the problem if you can carry out either procedure B or procedure C").

Second, we have transformed the problem to a scale that is familiar to experts in educational measurement. Thus, they will more comfortable with elicitation process on this scale. The scale of IRT models is often set by standardizing the distribution of θ, and in this metric a value of -1 for b indicates an item that is somewhat easy for the examinees, 0 a typical item, and +1 a somewhat difficult item; further, a parameters typically range from about .3 to 3. When the expert says she expects an item to be easy for the intended population, or that responses will be fairly strongly related to proficiency, we have a good idea of what the a and b parameters will be. If we are planning to refine the evidence models with pretest data we can elicit initial opinions in the form of linguistic parameters (e.g., "hard" or "easy") that are assigned to one of a number of numerical priors predefined by psychometricians.

We describe this setup for the one dimensional and multidimensional cases below, then show how the same approach can also be used to relax the assumption of independent observations.

2.1 Evidence models with univariate footprints

When $\mathbf{S}_{s,m} = \{S_{s,n}\}$, the projection function $g_m(\cdot)$ can be any monotonic function of $S_{s,n}$. Assuming that the levels of $S_{s,n}$ are roughly equally spaced we could use a linear function on the index, $g_m(\texttt{label}_i) = c_m i + d_m$. This model gives us just two parameters to elicit no matter how many states of $S_{s,n}$ or $X_{s,m}$ there are.

The intuition is that $\tilde{\theta}_{s,m}$ is the student's proficiency specific to solving Task m. The function $g_m(\cdot)$ is the projection of $S_{s,n}$ onto that space. The constant parameter d_m is related to the average difficulty of the item, and the slope c_m depends on the sensitivity with which response probabilities discriminate among levels of $S_{s,n}$—that is, the difference in the conditional probabilities for levels of $S_{s,n}$ and hence the weights of evidence for the task.

Table 1 gives conditional response probabilities for a task with three ordered possible outcomes, and a single skill variable with three ordered states. The item parameters used for Figure 1 are again used, and $g_m(\texttt{label}_i) = 1i - 2$. That is, the states $\{\texttt{low}, \texttt{medium}, \texttt{high}\}$ are mapped to θ values of -1, 0, and +1 respectively.

2.2 Evidence models with multivariate footprints

Now suppose that $\mathbf{S}_{s,m} = \{S_{s,n_1}, \ldots, S_{s,n_J}\}$. We construct the projection function as

$$f_m(\mathbf{S}_{s,m}) = g_m\big(h_{m,1}(S_{s,n_1}), \ldots, h_{m,J}(S_{s,n_J})\big).$$

Here each $h_{m,j}(S_{s,n_j}) = \tilde{\theta}_{s,m,j}$ is a projection of the marginal influence of skill S_{s,n_j} on Task m. The structure function $g_m(\tilde{\theta}_{s,m,1}, \ldots, \tilde{\theta}_{s,m,J})$ describes how the skills interact to produce proficiency in solving this particular task. As in the univariate case, if we assume the skill levels are roughly equally spaced we can describe that relationship with two parameters per skill: $c_{m,j}$ and $d_{m,j}$.

The structure function $g_m(\cdot)$ describes how the skills interact. Some common choices are *compensatory* (sum)—skills compensate for each other,—*conjunctive*(min)—all skills are necessary for solving the problem,—and *disjunctive*(max)—any of the skills can be used to solve the problem. We have also experimented with some asymmetric combination functions. Of particular interest is the *inhibitor* model in which one skill must be at a threshold value before the other skills play a role at all. Experts are often able to suggest the functional form of $g_m(\cdot)$; this is particularly so when they have constructed tasks deliberately to evoke certain skills in certain combinations. Section 3.0 provides examples of the mathematical structures for some $g_m(\cdot)$ functions we used in the Biomass project.

2.3 Dependent Evidence

One limitation of the IRT model is that all of the evidence is considered independent given the proficiency of the student. This assumption breaks down in the presence of complex tasks which yield multiple observations, for example, a reading passage followed by several questions. In this example, students who happened to be familiar with the topic of the passage would do better across all of the questions at any proficiency level.

In the Bayesian network model, we can formalize this notion by introducing into the evidence model an independent *context* skill variable C_m to represent familiarity with the topic or context of Task m. When we model the responses for Task m, C_m is treated as an extra parent of the observations from Task m. After absorbing all evidence from Task m this variable can be discarded with the rest of the evidence model. The example in Section 3.1 illustrates the use of a context variable. (See Bradlow, Wainer, and Wang, 1999, discuss a similar model from an alternative perspective.)

3 A numerical example

3.1 The Biomass Project

Biomass was a project carried out at Educational Testing Service in 2000. A computer-based prototype assessment was developed for secondary-school biology, with an emphasis on inquiry skills and model-based reasoning in the context of microevolution and transmission genetics. The student model $\mathbf{S}$ in the Biomass prototype consisted of fifteen variables. It is shown as Figure 2. The ovals represent the SM variables, the squares represent probability distributions, and the edges represent the dependence relationships among variables. Four multistage investigative tasks were developed, which required a total of forty-eight evidence models to manage incoming information about students' proficiencies. An evidence model contained between one and ten observable variables X_m, and has from one to four student-model variables in its footprint.

S_s	$\tilde{\theta}_s$	P_1^*	P_2^*	Pr(low)	Pr(medium)	Pr(high)
low	-1.00	0.50	0.12	0.50	0.38	0.12
medium	0.00	0.73	0.27	0.27	0.46	0.27
high	+1.00	0.88	0.50	0.12	0.38	0.50

Table 1: A univariate model for a task with three ordered response categories.

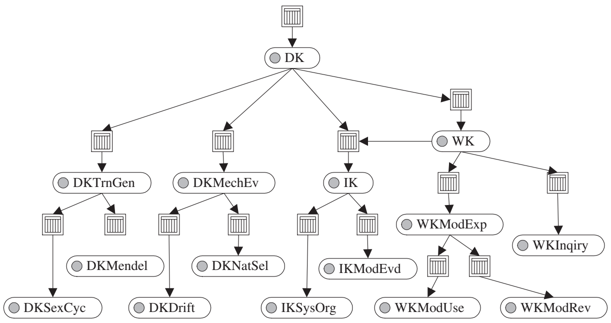

Figure 2: The student model for the Biomass project.

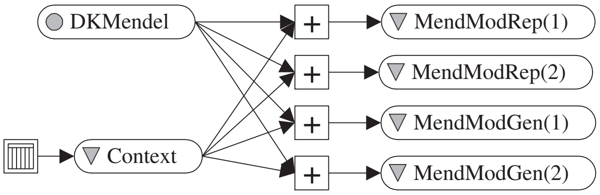

Figure 3: An evidence model with seven (four shown) conditionally dependent observations.

For each evidence model, we elicited the opinions of the experts who developed the tasks as to (a) the structure of the relationship among student model parents for each observable—i.e., compensatory, conjunctive, inhibition; (b) whether the task was harder than usual, easier than usual, or typical in difficulty for the intended population of students; (c) whether tasks within an evidence model should be considered conditionally dependent; and (d) the relative importance of skills, for observable variables with more than one student-model parent. The following sections present, for selected observables, the outcomes of these conversations and the resulting conditional probability matrices.

3.2 Agouti 1: Conditional Dependence

Figure 3 shows the evidence model for the first task in the Agouti mouse scenario, which concerns inquiry skills and knowledge about transmission genetics. The examinee must build a formal representation for the mode of inheritance for hair color in a population of mice, working from the results of some crosses and a colloquial description of a fictitious student's hypothesis. Seven features of the examinee's solution are evaluated, reflecting its accuracy, internal consistency, and correctness in using the representational conventions. (For brevity, only four are shown in the figure. The others are similar.)

The experts intend for all of these observables to bear on a single student-model variable, Disciplinary Knowledge of the Mendelian Model—DKMendel for short. Because all of these observable variables are aspects of the same problem solution, they are to be modeled as conditionally dependent, hence the additional 'Context' parent of the observable variables. They indicated that the first of the seven, say $X_{A1,1}$, was easier than typical. $X_{A1,1}$ has three ordered levels that rate an examinee's solution as to the chromosome type he has indicated. From these verbal specifications, we created the following structure for the conditional probability table for $X_{A1,1}$.

The two parents of $X_{A1,1}$, denoted $S_{s,A1_1}$ and $S_{s,A1_2}$, are DKMendel and Context. Their ranges are $\{\texttt{low}, \texttt{medium}, \texttt{high}\}$ and $\{\texttt{low}, \texttt{high}\}$. Experts told us the first observable is easier than typical, so we set $c_{A1,1} = 1$ and $d_{A1,1} = -1$; that is,

$$h_{A1,1}(S_{s,A1_1}) = \tilde{\theta}_{s,A1,1} = 1i - 1\,.$$

Thus $\{\texttt{low}, \texttt{medium}, \texttt{high}\}$ are mapped to $\tilde{\theta}$ values of 0, 1, and 2, which will lead to higher probabilities for higher level responses. For Context, we set $c_{A1,2} = 1$ and $d_{A1,2} = -1.5$, so

$$h_{A1,2}(S_{s,A1_2}) = \tilde{\theta}_{s,A1,2} = i - 1.5\,.$$

Thus $\{\texttt{low}, \texttt{high}\}$ are mapped to $\tilde{\theta}$ values of -.5 and +.5. In order to effect conditional dependence without affecting the marginal influence of DKMendel, we want to center the prior distribution of Context around zero. The proposed specifications for $c_{A1,2}$ and $d_{A1,2}$ accomplish this. Looking ahead toward refining distributions from data, we will generally want to be able to revise c while keeping the distribution centered. If the index values of Context are $\{1, 2\}$, we can *estimate* c and *set* $d = -1.5c$.

As described in Section 2.3, the structure function that combines the influence of these two factors is compensatory, or summative; that is,

$$\tilde{\theta}_{s,A1} = \sum_{n=1}^{2} c_{A1,n} S_{A1_n} + d_{A1,n} \, .$$

Table 2 shows the resulting conditional probabilities.

3.3 Agouti 4: An Inhibition Relationship

Figure 4 shows the evidence model for a subsequent task in the Agouti mouse scenario. After having proposed a hypothesis for the mode of inheritance of the mice, the question is what to do next. Our experts told us that the examinee's answer depends mainly on her understanding of the inquiry process—it is time to cross some mice, to provide data to test the hypothesis—but the examinee has to know at least enough about the Mendelian model to understand the problem. This means the relationship between inquiry skill (WKInqry) and DKMendel with respect to the observable is now inhibition; the examinee must get 'over the hurdle' of having DKMendel at least `medium` before WKInqry can come into play.

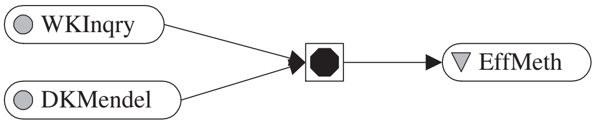

Figure 4: An evidence model with two parents in an inhibition relationship.

The experts indicated that the difficulty of this problem is typical, so we set $c_{A4,1} = 1$ and $d_{A4,1} = 2$ for the marginal projection function for WKInqry. That is,

$$h_{A4,1}(S_{s,A4_1}) = \tilde{\theta}_{s,A4,1} = i - 2 \, ,$$

and {`low`, `medium`, `high`} maps to $\tilde{\theta}$ values of -1, 0, and 1. For DKMendel, we employed a different kind of marginal function:

$$h_{A4,2}(S_{s,A4_2}) = \tilde{\theta}_{s,A4,2} = \begin{cases} 0 & \text{if } i = 0 \\ 1 & \text{if } i = \{1,2\} \end{cases} .$$

In other words, it is 0 if the examinee is not at least `medium` on DKMendel, and 1 if she is. The inhibiting influence of DKMendel is then effected by the structural function

$$g_{A4}\big(\tilde{\theta}_{s,A4,1}, \tilde{\theta}_{s,A4,2}\big) = \\ \big(1 - \tilde{\theta}_{s,A4,2}\big)\big(c_{A4,1}\big(-1\big) + d_{A4,1}\big) + \tilde{\theta}_{s,A4,2}\tilde{\theta}_{s,A4,1} \, .$$

Table 3 gives the resulting conditional probabilities.

3.4 Lizard 1: Disjunction + Conditional Dependence

Figure 5 shows the evidence model Bayes net fragment for Segment 1 in the "lizard scenario." This task addresses whether a student can judge the strength of an argument for disconfirming, supporting, or proving that specific mechanisms of evolution are operating in a problem context. The three nodes on the right represent observable variables—taken together, $\mathbf{X}_{s,L1}$. Domain experts designed these questions to elicit evidence from the student about Knowledge of Microevolution (DKMechEv) and Inquiry (WKInqry). All the student-models and observables in this task have three ordered values, which we again call {`low`, `medium`, `high`}.

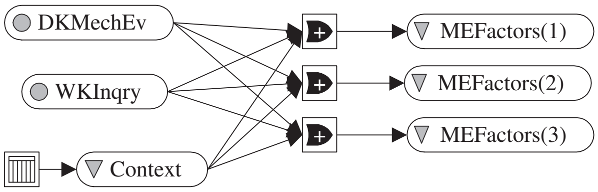

Figure 5: An evidence model with two student model parents and conditional dependence among observable variables.

The domain experts indicated that (a) the relationship between these skills, concerning their effect on the observables, was *disjunctive*; (b) response probabilities were more sensitive to WKInqry than on DKMechEv; and (c) the observables should be modeled as conditionally dependent, because all of the observables are evaluations of the same set of actions in a common problem setting. We thus have three parents for each observable, namely, WKInqry, DKMechEv, and Context. (Note that Context for Lizard L1 is actually a different variable than Context for Agouti 1.)

In accordance with the experts' verbal prior expectations, we set the following values for the marginal projections. For WKInqry, $h_{L1,1}(S_{s,L1_1}) = \tilde{\theta}_{s,L1,1} = 1.5i - 3$. For DKMechEv, $h_{L1,2}(S_{s,L1_2}) = \tilde{\theta}_{s,L1,2} = 1.0i - 2$. And for Context, $h_{L1,3}(S_{s,L1_3}) = \tilde{\theta}_{s,L1,3} = i - 1.5$. The structural function takes the maximum of $\tilde{\theta}_{s,L1,1}$ and $\tilde{\theta}_{s,L1,2}$, followed by a compensatory relationship with $\tilde{\theta}_{s,L1,3}$:

$$g_{L1}\big(\tilde{\theta}_{s,L1,1}, \tilde{\theta}_{s,L1,2}, \tilde{\theta}_{s,L1,3}\big) = \\ \max\big(\tilde{\theta}_{s,L1,1}, \tilde{\theta}_{s,L1,2}\big) + \tilde{\theta}_{s,L1,3} .$$

The resulting conditional probabilities appear in Table 4.

DKMendal	$\tilde{\theta}_1$	Context	$\tilde{\theta}_2$	$\tilde{\theta}$	P_1^*	P_2^*	Pr(low)	Pr(medium)	Pr(high)
low	0.00	low	-0.5	-0.50	0.62	0.18	0.38	0.44	0.18
low	0.00	high	0.5	0.50	0.82	0.38	0.18	0.44	0.38
medium	1.00	low	-0.5	0.50	0.82	0.38	0.18	0.44	0.38
medium	1.00	high	0.5	1.50	0.92	0.62	0.08	0.30	0.62
high	2.00	low	-0.5	1.50	0.92	0.62	0.08	0.30	0.62
high	2.00	high	0.5	2.50	0.97	0.82	0.03	0.15	0.82

Table 2: Conditional probability distributions for a three-valued observable variable, with one student-model parent and one context parent in a compensatory relationship.

WKInqry	$\tilde{\theta}_1$	DKMendel	$\tilde{\theta}_2$	$\tilde{\theta}$	P_1^*	P_2^*	Pr(low)	Pr(medium)	Pr(high)
low	-1.00	low	0.00	-1.00	0.50	0.12	0.50	0.38	0.12
low	-1.00	medium	1.00	-1.00	0.50	0.12	0.50	0.38	0.12
low	-1.00	high	1.00	-1.00	0.50	0.12	0.50	0.38	0.12
medium	0.00	low	0.00	-1.00	0.50	0.12	0.50	0.38	0.12
medium	0.00	medium	1.00	0.00	0.73	0.27	0.27	0.46	0.27
medium	0.00	high	1.00	0.00	0.73	0.27	0.27	0.46	0.27
high	1.00	low	0.00	-1.00	0.50	0.12	0.50	0.38	0.12
high	1.00	medium	1.00	1.00	0.88	0.50	0.12	0.38	0.50
high	1.00	high	1.00	1.00	0.88	0.50	0.12	0.38	0.50

Table 3: Conditional probability distributions for a three-valued observable variable, illustrating an inhibition relationship.

WKInqry	$\tilde{\theta}_1$	DKMechEv	$\tilde{\theta}_2$	Context	$\tilde{\theta}_3$	$\tilde{\theta}$	P_1^*	P_2^*	Pr(low)	Pr(med)	Pr(high)
low	-1.50	low	-1.00	low	-0.50	-1.50	0.38	0.08	0.62	0.30	0.08
low	-1.50	low	-1.00	high	0.50	-0.50	0.62	0.18	0.38	0.44	0.18
low	-1.50	medium	0.00	low	-0.50	-0.50	0.62	0.18	0.38	0.44	0.18
low	-1.50	medium	0.00	high	0.50	0.50	0.82	0.38	0.18	0.44	0.38
low	-1.50	high	1.00	low	-0.50	0.50	0.82	0.38	0.18	0.44	0.38
low	-1.50	high	1.00	high	0.50	1.50	0.92	0.62	0.08	0.30	0.62
medium	0.00	low	-1.00	low	-0.50	-0.50	0.62	0.18	0.38	0.44	0.18
medium	0.00	low	-1.00	high	0.50	0.50	0.82	0.38	0.18	0.44	0.38
medium	0.00	medium	0.00	low	-0.50	-0.50	0.62	0.18	0.38	0.44	0.18
medium	0.00	medium	0.00	high	0.50	0.50	0.82	0.38	0.18	0.44	0.38
medium	0.00	high	1.00	low	-0.50	0.50	0.82	0.38	0.18	0.44	0.38
medium	0.00	high	1.00	high	0.50	1.50	0.92	0.62	0.08	0.30	0.62
high	1.50	low	-1.00	low	-0.50	1.00	0.88	0.50	0.12	0.38	0.50
high	1.50	low	-1.00	high	0.50	2.00	0.95	0.73	0.05	0.22	0.73
high	1.50	medium	0.00	low	-0.50	1.00	0.88	0.50	0.12	0.38	0.50
high	1.50	medium	0.00	high	0.50	2.00	0.95	0.73	0.05	0.22	0.73
high	1.50	high	1.00	low	-0.50	1.00	0.88	0.50	0.12	0.38	0.50
high	1.50	high	1.00	high	0.50	2.00	0.95	0.73	0.05	0.22	0.73

Table 4: Conditional probability distributions for a three-valued observable variable, illustrating an a disjunctive relationship over two student-model variables and conditional dependence with tasks from the same segment.

4 Next steps

Section 3 showed how we used expert judgment to obtain initial values for conditional probabilities in Biomass evidence models. The next step will be to gather empirical data to refine these probabilities. Following Mislevy *et al.* (1999) we are planning to use Markov Chain Monte Carlo to fit the parameters, possibly using BUGS (Spiegelhalter *et al.*, 1995).

We are exploring three approaches for this latter step. In the first, we would use the output of the "projection plus graded response" model to produce Dirchlet priors for the conditional probability tables, then update each conditional probability directly from observations; that is, without further constraint on the likelihood functions. However, as we expect that many of the entries will be quite sparse, this will probably not lead to good results (e.g., unstable, or violating monotonicity relationships). A second possibility is to follow Mislevy *et al.* (1999) and use the structure function to tie similar elements of the conditional probability matrix together. However, this does not allow us to learn information about the relative scaling of the skills in each task. We expect the most promising approach is to assign priors to $c_{m,j}$ and $d_{m,j}$ based on the verbally-valued assessments of our experts, and then learn these parameters directly (recalling that the d parameters for Context variables are not estimated). Again we can draw on thirty years of experience with IRT models to propose 'reasonable' prior distributions. As starting point, we will posit independent normal priors on ds and lognormal priors on cs, with means set at the numerical translations of verbal priors and standard deviations of 1. (Our preliminary work here shows that it is help to "center" the label indexes around 0; i.e., to use $-1, 0, 1$ instead of $1, 2, 3$.)

If we had additional collateral information about the tasks, such as format and content elements that were systematically related to their difficulty, we could use that to reduce the pretest sample size (Mislevy, Sheehan and Wingersky, 1993). In this case we could use the collateral information in our model of $c_{m,j}$, $d_{m,j}$ or b_m. Adams, Wilson, and Wang (1997) have proposed a model along these lines for continuous latent variables, which could be adapted to the structured latent class situation we have discussed here.

5 Conclusion

As interest in increasingly ambitious assessment tasks grows, knowing how to make sense of the resulting complex response data has become a bottleneck. We cannot rely on purely empirical methods to establish relationships between complex observations and multivariate constructs. This research illustrates a way to use expert judgment to produce initial values for the conditional probability tables for tasks in multivariate latent class models. It builds on experience with univariate graded-response IRT models to ease the burden on subject matter experts, and is amenable to the framework of Bayesian estimation for refining these judgments as data accumulate.

References

Adams, R., Wilson, M.R., and Wang, W.-C. (1997) . "The multidimensional random coefficients multinomial logit model." *Applied Psychological Measurement,* **21**, 1–23.

Almond, R.G., and Mislevy, R.J. (1999). "Graphical models and computerized adaptive testing." *Applied Psychological Measurement,* **23**, 223–238.

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

Formann, A.K. (1985). " Constrained latent class models: Theory and applications." *British Journal of Mathematical and Statistical Psychology,* **38**, 87–114.

Mislevy, R.J, Almond, R.G., Yan, D. and Steinberg, L.S. (1999). "Bayes Nets in Educational Assessment: Where the numbers come from." In *Uncertainty in Artificial Intelligence '99*, Laskey, K.B. and Prade, H. (eds.). Morgan-Kaufmann, 437–446.

Mislevy, R.J., Sheehan, K.M. and Wingersky, M.S. (1993). "How to equate tests with little or no data." *Journal of Educational Measurement,* **30**, 55–78.

Spiegelhalter, D.J., Thomas, A., Best, N.G., and Gilks, W.R. (1995). "BUGS: Bayesian inference Using Gibbs Sampling, Version 0.50." MRC Biostatistics Unit, Cambridge.
`http://www.mrc-bsu.cam.ac.uk/bugs/`

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