

---

# “I can name that Bayesian Network in Two Matrixes!”

---

Russell G. Almond\*

ETS

Princeton, NJ 08541

## Abstract

The traditional approach to building Bayesian networks is to build the graphical structure using a graphical editor and then add probabilities using a separate spreadsheet for each node. This can make it difficult for a design team to get an impression of the total evidence provided by an assessment, especially if the Bayesian network is split into many fragments to make it more manageable. Using the design patterns commonly used to build Bayesian networks for educational assessments, the collection of networks necessary can be specified using two matrixes. An inverse covariance matrix among the proficiency variables (the variables which are the target of interest) specifies the graphical structure and relation strength of the proficiency model. A  $Q$ -matrix — an incidence matrix whose rows represent observable outcomes from assessment tasks and whose columns represent proficiency variables — provides the graphical structure of the evidence models (graph fragments linking proficiency variables to observable outcomes). The  $Q$ -matrix can be augmented to provide details of relationship strengths and provide a high level overview of the kind of evidence available in the assessment. The representation of the model using matrixes means that the bulk of the specification work can be done using a desktop spreadsheet program and does not require specialized software, facilitating collaboration with external experts. The design idea is illustrated with some examples from prior assessment design projects.

Key words: Bayesian Networks, Elicitation,  $Q$ -Matrix, Assessment Design, Covariance Selection Models

## 1 Problem

Bayesian networks are an attractive modeling paradigm because they can capture a wide variety of complex interactions among variables. However, they require a fairly intensive amount of knowledge engineering to build the models. This is especially true if the models contain a large number of latent variables, as is true in many psychological measurement applications, because the expert knowledge may be required to identify latent variables and their states, even if the model parameters are later refined through data.

Evidence-centered assessment design (ECD; Mislevy, Steinberg, & Almond, 2003) is a knowledge engineering method for building Bayesian network models for educational assessment. It starts by factoring the complete network for the assessment into a central core *student proficiency model* and a collection of *evidence models* corresponding to the tasks (Almond & Mislevy, 1999). However, when there are multiple forms of an assessment, there can be a large number of tasks and evidence models to manage. The library of fragments metaphor does not provide a convenient overview of the properties of the entire assessment or any particular form. Section 2 describes our first ECD design repository and its limitations.

Note that a graph may also be expressed through an incidence matrix, a matrix whose rows and columns correspond to nodes in the graph and where a positive value indicates an edge between the corresponding nodes. The original ECD method split the complete Bayesian network for an assessment into a central proficiency model and a collection of evidence model fragments for each task. The revised ECD method shown here will similarly use two matrixes to express the design. The first is called the  $Q$ -Matrix (Section 3) and describes the relationship between proficiency vari-

---

\*Paper submitted to 5th Application Workshop at Uncertainty in Artificial Intelligence Conference 2007, Vancouver, BC, Canada.

ables and observable outcome variables. The second is a correlation matrix among observable variables (Section 4). These two matrixes not only provide a good overview of the model, but they also can be specified using common spreadsheet programs available on the experts desktop, and hence do not require specialized software. This suggests a lighter weight, more nimble procedure for knowledge engineering (Section 5).

## 2 The Evidence-Centered Design Data Repository

In ECD, a complete design for an educational assessment consists of a number of design objects called models (Mislevy et al., 2003). The four central models for the assessment lay out the basic evidentiary basis for the assessment as follows:

1. *Student Proficiency Model*. Identify the aspects of student knowledge, skill and ability about which the assessment will make claims.
2. *Evidence Model*. Identify observable evidence for the student having (not having) the targeted proficiencies.
3. *Task Model*. Design situations which provide the student with opportunities to provide that evidence.
4. *Assembly Model*. Describe rules for how many of what kinds of tasks will constitute a valid form of the assessment.

The last explicitly recognizes that the space of all possible tasks the student could encounter is usually so large that administering all possible tasks to the student is logistically impossible. Often for reasons of repeated testing or security, all students do not receive the same form (collection of tasks). It is common in high stakes assessments for several forms of a test to be printed. In the extreme case of computer adaptive testing, the computer selects a potentially unique sequence of tasks for each student taking the assessment. In all cases, the assembly model controls what constitutes a valid form of the assessment.

The ECD design tool Portal (Steinberg, Mislevy, & Almond, Pending) represented each design object (model) as an electronic entity in a database. The design team, usually through a designated design librarian, entered the design through a series of on-screen forms corresponding to the models described above. The forms were complex, requiring multiple tabs for each model to represent various ways the information could be used in different contexts. The models could

be created in any order; the true design process is iterative. At the end of the process the design team selects a set of models which work together to make a coherent assessment design, called a *conceptual assessment framework* (CAF).

The ECD process recognizes the fact that designs go through several phases. In the first phase, called *Domain Analysis*, the design team organizes requirements for the assessment, and existing bodies of knowledge about the domain to be tested (cognitive theories about the domain, information gleaned from similar assessments). In the second phase, called *Domain Modeling*, the design team builds a preliminary sketch of the assessment argument. Similar in concept to knowledge maps (Howard, 1989), this part of the modeling process is designed to help the design team with trade-off decisions, and selecting the variables and grain size appropriate to the purpose of the assessment. The third phase is the CAF where the final specifications would be determined for a particular assessment. The design tool included support for information pedigree, linking representations of concepts in the later design phases to the prototypes in earlier phases (Bradshaw, Holm, Kipersztok, & Nguyen, 1992). The final model produced in the CAF could be exported as XML data to be sent to StatShop (Almond, Yan, Matukhin, & Chang, 2006), our tool for Bayes net scoring and calibration (parameter estimation).

The CAF editing tool offered three different modes: a database editing tool for defining and documenting variables and models, a graphical drawing tool for drawing graphical structures for student and evidence models, and a spreadsheet tool for entering conditional probability tables. To speed implementation, the latter two views used linked external programs (Microsoft Visio and Excel, respectively) to handle graphical data and conditional probability tables. This ultimately proved to be Portal's downfall, as when upgrades to the linked software broke the Portal interface, the decision was made not to continue supporting Portal.

While it was generally agreed that the ECD process was capturing valuable information for assessment design, Portal as a tool had many problems. First, as it was designed to cover all cases, it had a large number of fields that were not used in any given project. Second, the view of the data was not always the most natural or convenient. In particular, drawing a separate graph fragment for each evidence model could be a daunting task. For the NetPASS project (Behrens, Mislevy, Bauer, Williamson, & Levy, 2004), there were a total of nine tasks from three task models, so building a custom evidence model for each was manageable. For the ACED project (Shute, Hansen, & Almond, 2006),

there were around six hundred potential tasks in the design from about 100 task models, each of which had a single observable outcome variable. Switching back and forth between the three views took a heavy load on the computer (constantly launching helper applications, which was tediously slow) and the operator (constantly switching interfaces provided ample opportunity for confusion and mistakes). Another problem with the Portal view of assessment design as a collection of models is that it did not provide a clear overview of what the assessment was about.

The ACED project is an interesting study in model design. The design team handled the design management problem by creating a big spreadsheet of tasks and models. The rows were labeled with tasks, and the first several columns indicated which proficiency variables were relevant for each task. Other columns indicated the difficulty target or tracked the person responsible for the task and its current status. The team librarian then laboriously transferred the design information from the spreadsheet into Portal.

Upon reflection, it appears that 80% of cases will look more like ACED than NetPASS. In particular, even if an assessment contains large, complex simulation tasks, it usually contains a supporting collection of small discrete tasks as well. Furthermore, the spreadsheet view used by the ACED design team provides a good overview of the entire assessment. Thus, in designing a replacement for Portal we looked to this spreadsheet view of the graph.

### 3 Q-Matrix

If we restrict our attention to tasks which yield a single binary observable outcome variable, then we can use the  $Q$ -matrix (Fischer, 1973; Tatsuoka, 1984) to represent the relationship between observable and proficiency variables. The  $Q$ -matrix is a simple incidence matrix in which the columns represent proficiency variables and the rows represent tasks (items). There is a one in a cell if the skill is relevant to solving the task represented in the row, and a zero if it is irrelevant. Following the (Almond & Mislevy, 1999) notation, each row of the  $Q$ -matrix corresponds to an evidence model Bayes net fragment, when the one entries indicate that the indicated proficiency variable is a parent of the observable variable for that task.

Although the  $Q$ -matrix gives the general shape of the graphical model, it does not tell us how to parameterize that model. (Almond et al., 2001) introduced the idea of providing a small vocabulary of possible parameterizations that the domain expert could pick from. The following four parameterizations for conditional probability tables were the most useful:

**Conjunctive.** All skills are necessary to solve the problem. This model tends to look like a noisy-and or noisy-min model.

**Disjunctive.** The skills represent alternative solution paths, and only one is necessary to solve the problem. This model tends to look like a noisy-or or noisy-max.

**Compensatory.** Having more of one skill will compensate for having less of another. This is an additive model, where the probability of success depends on a (weighted) sum of the skill levels.

**Inhibitor.** Success in the problem is primarily dependent on one skill, but unless the student has a minimal level of another skill, then they are unlikely to be able to solve the problem at all. An example of this is a mathematics word problem where part of the challenge is extracting the relevant data from a natural language description of the problem. Students with insufficient familiarity with the language of the test will be unable to solve the problem, but once the minimum language threshold has been reached, additional language skill will not help solve the problem.

A second key idea introduced by Lou DiBello (Almond et al., 2001) is that if we can map the categorical levels of the Bayesian network to values on a unit normal distribution, then we can press well understood models from item response theory (IRT) into service. As this trick was frequently used with Samejima's graded response model, this class of models became known as DiBello-Samejima models.

#### 3.1 Translating between discrete and continuous variables

Assume that we have a discrete variable  $S_m$  which can take on values  $\{s_{m,1}, \dots, s_{m,K}\}$  and a continuous mirror variable  $Y_m$ . We consider the states of  $S_m$  to be ordered, so that  $s_{m,k} \succ s_{m,k'}$  if and only if  $k > k'$ . Let  $p_{m,k} = P(S_m = s_{m,k})$  and  $P_{m,k} = P(S_m \succeq s_{m,k})$  and as a special case define  $P_{m,0} = 0$ . Furthermore, let  $\mu_{Y_m}$  and  $\sigma_{Y_m}$  be the mean and standard deviation of  $Y_m$  (which will be zero and one if  $Y_m$  is scaled to a unit normal distribution).

We can think of the variable  $Y$  as partitioning the distribution of  $Y_m$  into a number of bins (Figure 1). The widths of the bins are determined by the probabilities  $p_{m,k}$ . Thus, we need to set the cutpoints between the bins,  $c_{m,k}$  at the point so that the area under the curve for that bin equals  $p_{m,k}$ . This can be done using the formula:

$$c_{m,k} = \mu_{Y_m} + \sigma_{Y_m} * \Phi^{-1}(P_{m,k}) \quad \text{for } k < K_m, \quad (1)$$

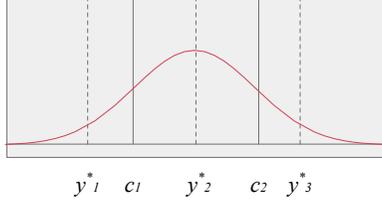


Figure 1: Cutpoints for a normal distribution

where  $\Phi^{-1}(\cdot)$  is a function that produces quantiles of the normal distribution (the inverse normal c.d.f.).

To convert from the discrete variable  $S_m$  to the continuous variable  $Y_m$  we can represent each interval with its midpoint. Note that the first and last intervals actually stretch to infinity, that is  $c_{m,0} = -\infty$  and  $c_{m,K} = +\infty$ . We can work around this problem by taking midpoints with respect to the normal density. Thus, we define:

$$y_{m,k}^* = \mu_{Y_m} + \sigma_{Y_m} * \Phi^{-1}(P_{m,k} - p_{m,k}/2) . \quad (2)$$

Reversing the procedure is also straightforward. Suppose that we learn (through building a regression model, Section 4.1) that  $E[Y_m|X = x] = \mu_{Y_m|x}$  and  $\text{var}[Y_m|\mathbf{X} = \mathbf{x}] = \sigma_{Y_m|x}^2$ . We can then calculate the conditional probability of  $S_m$  given  $\mathbf{X} = \mathbf{x}$  as follows:

$$\begin{aligned} P(S_m = s_{m,k}|\mathbf{X} = \mathbf{x}) &= P(c_{m,k-1} \leq Y_m \leq c_{m,k}|\mathbf{X} = \mathbf{x}) \\ &= \Phi\left(\frac{c_{m,k} - \mu_{Y_m|x}}{\sigma_{Y_m|x}}\right) \\ &\quad - \Phi\left(\frac{c_{m,k-1} - \mu_{Y_m|x}}{\sigma_{Y_m|x}}\right) \end{aligned} \quad (3)$$

where  $\Phi(\cdot)$  is the cumulative normal distribution function.

These procedures assume that the modeler has a fixed marginal distribution for  $S_m$  in mind. In some situations, it may be more natural to think of a fixed set of cut points,  $c_{m,k}$  (for example, if the cut scores were set by a standard setting committee). In this case, inverting Equation 1 produces values for  $P_{m,k}$  and the rest of the calculations follow.

### 3.2 Augmenting the $Q$ -matrix to support DiBello-Samejima models.

The DiBello-Samejima framework (Almond et al., 2001) goes on to use item response theory (IRT) models to calculate the actual probability. Each level of each proficiency variable is assigned an “effective theta” value, a number on a Normal (0,1) scale representing the average skill level of people in that group.

The skill levels are combined using a combination function based on the type of model (e.g., sum, min, max). This is fed into a latent variable logistic regression model. For the compensatory distribution, the model looks like:

$$P(Y_{ij} = \underline{Correct}|\theta_i) = \text{logit}^{-1} \left[ \sum_{k \in pa(j)} a_{jk} \theta_{ik} / \sqrt{|pa(j)|} - b_j \right] , \quad (4)$$

where  $\theta_{ik}$  represents Person  $i$ ’s effective theta value on Skill  $k$ , and  $pa(j)$  is the set of skills which are parents of the observable outcome variable  $Y_{ij}$ . Here  $|pa(j)|$  represents the number of parents of the observable and  $1/\sqrt{|pa(j)|}$  is a variance stabilization constant.

The parameters  $b_j$  and  $a_{jk}$  are known as the difficulty and discrimination parameters in item response theory. As assessment tasks can behave unexpectedly, domain experts will not be able to supply exact values for these parameters before an assessment is pretested. However, it is not unreasonable for the expert to place these values into broad categories, say “Easy”, “Medium” and “Hard” tasks. The analyst could then map these linguistic categories onto prior distributions. For example, the term “Easy” might map onto a normal prior with mean -1 and variance 1, while the “Hard” prior would have a mean of 1 and the same variance. In addition, to changing the mean, one could also change the variance, so that a classification of “Unknown” would translate into a prior with mean zero and variance of 2. If pretest data becomes available, then the model can later be refined with data.

We can specify all of this information in matrix form, by augmenting the  $Q$ -matrix representation. First, we add an additional column to indicate the parameterization for the conditional probability table. Next, we add a column to represent the difficulty. As with the original  $Q$ -matrix, zeros are used to indicate parent variables (skills) which are irrelevant to the task at hand. However, the entry in the relevant cells is now a numeric or linguistic value giving the strength of the relationship. Table 1 shows an example.

We note in passing that it is possible to add extra columns to this representation whose use is purely for the benefit of the analysts. In Table 1 the “Item” column gives the sequence number of the item on the form; information which is useful to the developers reviewing the form but is not used in constructing a Bayes net model for the assessment. Also, we can use additional columns to represent additional kinds of information. In the excerpt, we can see two values for the “ObsName” column, ISCORRECT (for binary observables) and PC4 (for four level partial credit mod-

Table 1: An augmented Q-Matrix from an experimental Reading test.

EvidenceModel	TaskName	ObsName	Form	Item	CPTType	Diff	S1	S2	S3	S4
EM8Word	VB533037	isCorrect	ReadA	1	Comp.	0	1	0	0	0
EM2ConnectInfo	VB533038	isCorrect	ReadA	2	Comp.	0	0	0	1	0
EM8Word	VB533039	isCorrect	ReadA	3	Comp.	0	1	0	0	0
EM4SpecInfo	VB533041	isCorrect	ReadA	4	Comp.	0	0	1	0	0
EM3ConnectSynthPC4	VB533431	pc4	ReadA	12	Comp.	0	0	0	1	1

els<sup>1</sup>). No inhibitor relationships are used in this example, but the additional information needed for the inhibitor model could again be represented as additional columns.

The information necessary to fill out each row of the augmented  $Q$ -matrix could be collected through a structured interview technique, however, just as the earlier Portal method of specifying the graph and the CPT separately, a separate interview for each row would not provide the designer with an overview of the assessment. In the matrix view, if two tasks are very similar, the designer can copy and paste information about an earlier row to construct the new row.

Furthermore, the  $Q$ -matrix provides a visual summary of the design of the assessment. Certain kinds of problems can be identified from the assessment. For example, if two skills always (or almost always) appear together as parents of observables, then the assessment will have difficulty distinguishing between them. In many cases, the principles of assessment form design are like experimental design.

Finally, note that the augmented  $Q$ -matrix can be stored in a spreadsheet. This means that members of the design team (including off-site consultants) can edit the data using standard office tools and do not need specialized software on their computers to access the data. Our strategy for building evidence models is now to elicit the necessary information from the experts using a spreadsheet like Table 1 customized for the project. We then use a package of functions written in R (R Development Core Team, 2005) to translate this spreadsheet into the XML model descriptions needed to drive the StatShop calibration and scoring system. Standard R programming style breaks the translation process into many small pieces, most of which are re-usable in new contexts. Thus, a minimal amount of custom coding is needed to support each project.

Even when using the individual models, the  $Q$ -matrix view has proved beneficial when checking our work. In development of ETS's ICT Literacy assessment, we

<sup>1</sup>(Almond et al., 2001) describes the extension of this type of model to observables with more than two levels

took the XML models exported by our Portal tool and ran them through a series of inverse functions, building the  $Q$ -matrix from the XML. We then used this to make sure that all parameters had been correctly specified.

## 4 Correlation Matrix

The augmented  $Q$ -matrix solves a substantial fraction of the problem. However, in order to specify a complete Bayes net scoring model for an assessment, the design team must also specify a proficiency model. This is a complete Bayesian network and not just a fragment (the evidence models borrow nodes from the proficiency model, and hence are incomplete without the proficiency models). So in principle, any Bayesian network tool could be used for the job, although in practice there is still the difficulty of translating from the format of the Bayes net tool to StatShop's XML format.

In my experience with design teams, they have little difficulty identifying the relevant proficiency variables. The issues of how many variables to include in the model, and how many levels each variable should have always produces a lively debate, but the design team usually understands the issues when explained to them. In ECD practice, levels of the proficiency variables are defined through *claims* (statements about what students at a given proficiency level can and cannot do) that give the variables clarity and help to resolve some of the grain size issues. Showing a draft  $Q$ -matrix can help the design team resolve trade-offs of assessment scope versus length (and cost).

When it comes to the issue of creating graphical structure, however, the design team needs firm guidance. Without input from statistically sophisticated team members, the structure of the proficiency model tends to be a hierarchical breakdown of the domain rather than a statement of dependence and independence conditions among the variables.

The situation is even worse when it comes to the numbers. The proficiency variables are abstract and latent, and hence they provide little real world experience for which the expert can provide a judgment. In the pro-

cess of designing ACED, the expert charged with developing the proficiency model had a great deal of difficulty with the numbers. Although she understood the Bayesian networks and what was required, she simply did not have confidence in her numerical judgments.

#### 4.1 Regression Models

For the ACED project, a simple spreadsheet based on linear regression models provided our expert with a coherent framework for elicitation of these conditional probability tables. The “effective theta” mapping described in Section 3.1 mapped the levels of the parent variable (the proficiency model was tree shaped and all nodes had at most one parent) to continuous variables. The expert then specified a correlation and intercept for a regression model. This was used to create a new mean and variance for the output variable on the continuous scale. This was mapped back into continuous probabilities using the inverse mapping technique described in Section 3.1.

This was highly successful from the standpoint of interaction with the expert. She now only needed to specify two parameters: the correlation with the parent variable and the intercept, which could be interpreted as a difference in level between the parent and child variable. Although there was still one source of difficulty for the expert, both the parent and child variables were latent constructs. The expert seldom sees the latent variables, but rather sees the manifestation of those variables as performance on tasks. Consequently, the correlations were lower than perhaps appropriate in order to account for measurement error.

Using the spreadsheet, the expert was able to fill in the conditional probability tables in the ACED model. This method of generating conditional probability tables was later incorporated into StatShop as the DiBello-Normal model.

#### 4.2 Inverse Covariance Matrix

The final step in this story comes when I was working with another expert to build a model for reading. The expert and I had identified five different proficiencies involved, however, when I started asking questions about the relationship among the variables, and the expert responded by handing me correlations among observed scores on tests meant to reflect the various proficiency scales. I began to realize that in general, expert knowledge about the relationships among psychological variables comes from factor analysis and structural equation modeling studies involving both manifest and latent variables. Often these analyses produce correlation matrixes.

This is all the more interesting as there is a close connection between the inverse of the correlation matrix and the graphical model (Whittaker, 1990). In particular, zeros in the inverse covariance matrix represent conditional independence between variables (Dempster, 1972). Thus, the pattern of zeros in the inverse covariance matrix provides an undirected graphical structure for the proficiency model.

Suppose that we are given the following information about the proficiency model:

1. A collection of categorical variables  $\mathcal{S}$  which belong in our proficiency model. Additionally we assume that for each categorical variable  $S_m$  there is a corresponding continuous factor  $Y_m$  coming from a factor analysis or structural equation model for the domain.
2. A collection of marginal distributions,  $P(S_m)$ , over the variables in  $\mathcal{S}$ .
3. The matrix  $\Sigma = \text{cov}(\mathbf{Y})$ , or at least an estimate of that matrix.
4. The expected value  $\mu_{\mathbf{Y}}$  of  $\mathbf{Y}$  or at least an estimate of that quantity.

The following steps should then produce a proficiency model.

1. Construct the inverse correlation matrix,  $W$ , by inverting and scaling the covariance matrix.
2. Select a threshold,  $t_{\min}$  and construct an undirected graph by adding an edge between Node  $i$  and Node  $j$  if  $|w_{ij}| > t_{\min}$ .
3. Use maximum cardinality search (Tarjan & Yannakakis, 1984) to produce a perfect ordering<sup>2</sup> of the nodes. Direct each edge from the lower to the higher numbered node.
4. Produce a regression model regressing each variable on its parents in the graph. The intercept and residual standard deviation in each regression is set to match the specified marginal distributions for the parent and child variables.
5. Produce conditional probability tables by discretizing the regression models.

This procedure above assumes that the covariance matrix expresses the relationships between the latent variables. Such matrixes are commonly available from

<sup>2</sup>A perfect ordering exists only if the graph is triangulated. As a normal graphical model exists over the variables  $\mathbf{Y}$  over the graph produced in Step 2, that graph must be triangulated.

Table 2: Covariance Matrix for Math Grade example (Whittaker, 1990).

	MECHANICS	VECTORS	ALGEBRA	ANALYSIS	STATISTICS
MECHANICS	302.29	125.78	100.43	105.07	116.07
VECTORS	125.78	170.88	84.19	93.60	97.89
ALGEBRA	100.43	84.19	111.60	110.84	120.49
ANALYSIS	105.07	93.60	110.84	217.88	153.77
STATISTICS	116.07	97.89	120.49	153.77	294.37

Table 3: Partial Correlation Matrix for Math Grade example (Whittaker, 1990).

	MECHANICS	VECTORS	ALGEBRA	ANALYSIS	STATISTICS
MECHANICS	1	-0.33	-0.23	0.00	-0.02
VECTORS	-0.33	1	-0.28	-0.08	-0.02
ALGEBRA	-0.23	-0.28	1	-0.43	-0.36
ANALYSIS	0.00	-0.08	-0.43	1	-0.25
STATISTICS	-0.02	-0.02	-0.36	-0.25	1

factor analysis or structural equation model results. If only observed score correlations are available, then that correlation matrix can be used instead, however, these will generally be lower than the latent variable correlations due to the measurement error in the instruments which measure them.

As with the  $Q$ -matrix, the covariance matrix and the supporting information about marginal distributions (and levels) for proficiency variables can be captured via any convenient means. A collection of R functions then translates the matrix into the XML model descriptions needed by StatShop.

### 4.3 An Example of the Inverse Covariance Matrix

We illustrate this procedure with using a data set analyzed in Whittaker (1990) originally taken from Mardia, Kent, and Bibby (1979). Table 2 gives the variance/covariance matrix for scores on five mathematics tests for a number of college students. Inverting and scaling the covariance matrix produces the partial correlation matrix shown in Table 3, where off-diagonal entries greater than 0.1 in absolute value have been colored gray. These correspond edges we wish to include in the model. The corresponding graph is given in Figure 2(a).

Next, we need to go from an undirected to a directed graph. A straightforward method for doing this is to choose an ordering of the variables. If an edge connects two variables, the orientation of the edge is set to go from the variable earlier in the list to the one later in the list. Although the choice of order is arbitrary, the subject matter experts should be consulted as some orderings may be more natural than others.

In going from the undirected to the directed representations not any ordering is appropriate. The choice of directions of the arrows must not induce any moralization edges which are not in the original graph. Consequently, the selected ordering must be a perfect ordering. As the graph in Figure 2(a) is triangulated, a perfect ordering exists. Figures 2(a) and 2(b) illustrate this idea. Because it seems natural that ALGEBRA is a pre-requisite for the other skills, it is put first in the list. The chosen ordering is ALGEBRA, MECHANICS, VECTORS, ANALYSIS, and STATISTICS is perfect. This induces the graph shown in Figure 2(b).

As it turns out, any ordering with ALGEBRA first is perfect. We get into trouble only if we put algebra after nodes from both the left and right wings of the butterfly. Thus the order MECHANICS, ANALYSIS, ALGEBRA would cause difficulties because then  $\text{pa}(\text{ALGEBRA}) = \{\text{MECHANICS}, \text{ANALYSIS}\}$  inducing a moralization edge between MECHANICS and ANALYSIS not present in the undirected graph Figure 2(a).

This procedure yields for every variable  $S_m$  in the model a set of parents  $\text{pa}(S_m)$ . It also yields a natural ordering of the variables, so that if we simply built a regression of each variable  $Y_m$  on its parents  $\text{pa}(Y_m)$  then we would get a normal graphical model for the continuous variables. All we need to do now is discretize the variables.

We define the categorical mirrors of these variables by defining three categories High, Medium and Low, where High corresponds to the upper quartile, Low corresponds to the lower quartile and the remaining half of the data are designated Medium. This means that the marginal distribution for all five variables in the model should be (0.25, 0.5, 0.25).



Figure 2: Graphical models for Five Math Test Scores (Whitakker, 1990)

Table 4: Unconditional probability table for ALGEBRA.

High	Med	Low
0.25	0.50	0.25

The first variable ALGEBRA has no parents and so it is an easy case. The CPT for algebra is just the marginal distribution (Table 4). This same rule applies for any other variables which have no parents in the directed graph.

The second variable, MECHANICS, has a single parent, ALGEBRA. This requires a regression model for MECHANICS given ALGEBRA. To begin, we calculate midpoint,  $x^*$ , values for the three states of ALGEBRA these are given in the second column of Table 5. Next, we solve the regression equations giving a slope of 0.90 (for ALGEBRA), an intercept of -5.59 and a residual standard deviation of 14.6. Cranking through the calculations using Equation 3 yields the conditional probability distribution shown in Table 5.

Table 5: Conditional probability table for MECHANICS.

ALGEBRA	$x^*_{ALGEBRA}$	High	Med	Low
High	38.45	0.48	0.46	0.06
Med	50.60	0.21	0.58	0.21
Low	62.75	0.06	0.46	0.48

The rest of the calculations proceed in a similar fashion.

One potential issue with this construction is that the original data presented in Mardia et al. (1979) are based on observed scores, rather than latent proficiency variables. We expect such observed scores to be lower due to measurement error, and a better procedure would take this into account. We could “bump up” the correlations to compensate, or use the generate CPTs as priors and learn better parameters for the proficiency model from data.

## 5 A New Philosophy of Knowledge Engineering

The preceding discussion shows how the bulk of the work of specifying a Bayesian network for an assessment can be expressed as specifying two matrixes: the augmented  $Q$ -matrix which provides the basis for the evidence models, and the (inverse) covariance matrix which provides the basis for the proficiency model. Additional details are still necessary (such as exact definitions for all the variables), however, these two matrixes provide the bulk of the elicitation process.

By switching to the matrix view, the design team is able to see more of the model at once. In particular, issues like insufficient tasks addressing a particular proficiency variable are difficult to see when mired in the details of drawing graphs and specifying conditional probability tables. The  $Q$ -matrix provides a high level view.

Another important feature of the new system is that the universal design database (Portal) has been replaced with a series of forms expressed as text documents and spreadsheets. This has several critically important consequences. First, the design team is free to focus on those parts of the ECD model relevant to their process. The Portal database still serves a useful role in listing issues that the design team needs to consider; however, the design team can choose from among those issues and organize them in the way that they please. This includes important representational issues. For example, in the NetPASS project the Portal database was designed to accommodate rules of evidence (instructions for how to set values for the observable variables) expressed as production rules. However, the programmer charged with implementing the rules said that he would rather the requirements be expressed as natural language constructions.

A second consequence of the custom forms translated to XML paradigm is that members of the design team no longer need custom software to edit the specifications. Design documents which can be edited with software installed on a typical desktop system supports collaboration with outside experts via email, as well as reducing the need for the librarian (although

a librarian still plays a useful role in managing design changes). Rational RequisitePro® (IBM, 2007), a product which supports the requirements analysis phase of software design, provides a similar paradigm. In particular, the design team edits word processor documents (using templates provided by RequisitePro) and then runs a software tool to extract details into a requirements database.

The new paradigm involves additional effort at the front end, customizing forms and data collection procedures, and at the back end, customizing form to XML translators. However, that effort pays off in more streamlined operations of the teams. In particular, much of the irrelevant information for a given project is stripped away, allowing the design team to focus on the issues important to that project. The open source and functional programming nature of the R tools provides strong support for reusing existing translation code. It also supports global changes (such as changing the prior variance for the difficulty parameter from 1 to 2) as a single translation function can be written rather than tediously enter the same change in a multitude of different distribution editing forms.

Another benefit of the new paradigm for design is expandability. At the design phase, adding new capability is as simple as adding a new column in the spreadsheet, or a new possible value to a list of values. Additional work may be needed in other parts of the production apparatus (the task authoring environment; the scoring and statistical analysis environment, StatShop; the test delivery infrastructure; the reporting infrastructure). But that work can take place concurrently with the design effort. In contrast working with a design tool like Portal would require changes in the design tool to support each new application.

One place where there is room for customization is in tasks with multiple observable outcome variables. Here the  $Q$ -matrix view can be modified by either assigning one row per observable or one row per task (essentially collapsing all of the observables into a single row). The former view is probably better for specifying models, but the latter for getting an impression of the total collection of evidence provided by the assessment. Another difficulty is that there is often local dependence forcing additional connections between observables from the same task. Almond, Mulder, Hemat, and Yan (2006) list a number of possible design patterns for modeling local dependence. A key observation is that not all of these need be supported within the augmented  $Q$ -matrix, only the ones which are needed for the project at hand.

Although the focus of this paper has been on the educational application, the two matrixes approach can

be used to represent any Bayesian network which can be partitioned into a system model and a collection of evidence models (Almond & Mislevy, 1999). In particular, many diagnostic applications fall into this category. What is called the “proficiency model” above becomes a system model describing the state of a patient (medical diagnosis) or machine (mechanical diagnosis). The evidence models (the rows of the  $Q$ -matrix) represent potential tests which the diagnostician can perform. This technique could be especially valuable if the test fall into a small number of design patterns; however, even if each test requires a custom Bayes net fragment, the  $Q$ -matrix can still reveal places where new kinds of test could provide potentially valuable evidence.

The biggest advantage using the matrixes rather than the traditional graph and spreadsheet approach to constructing the graphical models is that it brings the process closer to what experts see in their day-to-day experience. Kadane (1980) states that the closer the elicitation procedure gets to “observed data” an expert might actually see, the better the expert will be at supplying the numbers. The form customization procedure allows these two matrix views to be modified so that they use the organization and language of the experts.

## References

- Almond, R. G., Dibello, L., Jenkins, F., Mislevy, R. J., Senturk, D., Steinberg, L., et al. (2001). Models for conditional probability tables in educational assessment. In T. Jaakkola & T. Richardson (Eds.), *Artificial intelligence and statistics 2001* (p. 137-143). Morgan Kaufmann.
- Almond, R. G., & Mislevy, R. J. (1999). Graphical models and computerized adaptive testing. *Applied Psychological Measurement*, *23*, 223-238.
- Almond, R. G., Mulder, J., Hemat, L. A., & Yan, D. (2006). *Models for local dependence among observable outcome variables* (ETS Research Report No. RR-06-36). Educational Testing Service.
- Almond, R. G., Yan, D., Matukhin, A., & Chang, D. (2006). *Statshop testing* (ETS RM No. 06-04). Educational Testing Service.
- Behrens, J., Mislevy, R. J., Bauer, M., Williamson, D., & Levy, R. (2004). Introduction to evidence centered design and lessons learned from its application in a global e-learning program. *International Journal of Measurement*, *4*, 295-301.
- Bradshaw, J. M., Holm, P., Kipersztok, O., & Nguyen, T. (1992). eQuality: An application of DDUCKS to process management. In T. Wetter, K.-D. Althoff, J. H. Boose, B. R. Gaines, M. Linster, &

- F. Schmalhofer (Eds.), *Current developments in knowledge acquisition: EKAW-92*.
- Dempster, A. (1972). Covariance selection. *Biometrics*, 28, 157-175.
- Fischer, G. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, 37, 359-374.
- Howard, R. (1989). Knowledge maps. *Management Science*, 35, 903-922.
- IBM. (2007). *Rational RequisitePro (r)*. Software description downloaded on April 4, 2007 from.
- Kadane, J. (1980). Predictive and structural methods for eliciting prior distributions. In A. Zellner (Ed.), *Bayesian analysis and statistics*. North-Holland.
- Mardia, K., Kent, J., & Bibby, J. (1979). *Multivariate analysis*. Academic Press.
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessment (with discussion). *Measurement: Interdisciplinary Research and Perspective*, 1(1), 3-62.
- R Development Core Team. (2005). *R: A language and environment for statistical computing*. Vienna, Austria. (ISBN 3-900051-07-0)
- Shute, V. J., Hansen, E. G., & Almond, R. G. (2006). *An assessment for learning system called aced: The impact of feedback and adaptivity on learning*. (Research Report). ETS. (Draft in preparation, December, 2006.)
- Steinberg, L. S., Mislevy, R. J., & Almond, R. G. (Pending). *Portal assessment design system for educational testing*. U.S. Patent application. (Attorney Docket No. 246400.0159, Wilmer, Culter and Pickering.)
- Tarjan, R., & Yannakakis, M. (1984). Simple linear-time algorithms to test chordality of graphs, test acyclicity of hypergraphs, and selectively reduce acyclic hypergraphs. *Siam J. Comput.*, 13, 566-579.
- Tatsuoka, K. (1984). *Analysis of errors in fraction addition and subtraction problems* (Vol. 20; NIE Final report No. NIE-G-81-002). University of Illinois, Computer-based Education Research.
- Whittaker, J. (1990). *Graphical models in applied multivariate statistics*. Wiley.
- R functions. Interactions with John Sabatini, Richard Roberts and Aurora Graf, each experts in their respective fields, has taught me much about communicating ECD designs and methodology with others. Finally, the experiences of the ACED design team, Val Shute, Aurora Graf, Jody Underwood, David Williamson, and Peggy Redman have been very influential in this reconception of the ECD tools. ACED development and data collection was sponsored by National Science Foundation Grant No. 0313202.

## Acknowledgments

Lou DiBello provided a lot of the impetus for “simplifying Bayesian networks” which spurred this research. Debbie Pisacreta and Holly Knott did some early prototypes of a  $Q$ -matrix tool. Peggy Redman and Lisa Hemat have both served in the capacity of ECD librarian and their feedback has been a valuable source of input. Lisa additionally was an early user of many of the