Bayesian Networks: A Teacher's View

Russell G. Almond[*]
ETS
Princeton, NJ 08541

Valerie J. Shute
ETS
Princeton, NJ 08541

Jody S. Underwood
ETS
Princeton, NJ 08541

Juan-Diego Zapata-Rivera
ETS
Princeton, NJ 08541

Abstract

Teachers viewing Bayesian network-based proficiency estimates from a classroom full of students face a different problem from a tutor looking at one student at a time. Fortunately, individual proficiency estimates can be aggregated into classroom and other group estimates through sums and averages. This paper explores a few graphical representations for group level inferences from a Bayesian network.

Key words: Bayesian Networks, Computer Graphics, Probabilities, Aggregation

1 Teachers' Questions

Bayesian networks are becoming an increasingly popular way of representing the state of a student's knowledge, skills, or abilities, especially in intelligent learning environments (for example, ACED; Shute et al., 2005). The display capability of most Bayesian network software is designed to work with one individual at a time. A teacher, however, is typically concerned with making inferences about a classroom full of students. This paper looks at the problem of making inferences about groups individuals using the same Bayesian network.

Suppose that a teacher has 20–30 students who have taken an assessment which is scored using a Bayesian network. For each student, the teacher has a Bayesian network over a collection of *proficiency variables* which represents our best estimate of the student's state of proficiency. There are a number of questions the teacher might want to ask:

- How is the class doing overall? How many students are meeting or exceeding the curriculum objectives (standards)?
- Which students are not meeting the objectives? Which students are on the cusp of meeting the objectives?
- How are the students doing on each of the individual standards and skills (sub-proficiencies)?
- How does this class compare to other similar classes?
- How do previously identified groups within the classroom differ?
- What are the typical patterns of skill acquisition?
- How can the students be grouped into clusters which require similar kinds of instruction?
- Are there individuals with atypical patterns that require special attention?
- How much credence should I put in the estimates from the Bayes nets relative to other sources of information?
- What should I teach next?

The goal of this presentation is to explore some graphical representations which might start to answer those questions. We will do this using data collected from a prototype system called ACED (Shute et al., 2005).

2 ACED

ACED (Adaptive Content with Evidence-based Diagnosis; Shute et al., 2005) is a computer assessment of sequences appropriate for a course in middle school mathematics. ACED is an experimental prototype designed to explore: (a) the use of the Madigan and Almond (1995) algorithm to select the next task in a assessment, (b) the use of targeted diagnostic feedback,

[*]Email: ralmond@ets.org, almond@acm.org

and (c) the use of technological solutions to make the assessment accessible to students with visual disabilities.

Graf (2003) describes the construction of the Proficiency Model—a collection of latent variables describing the student's proficiency with sequences. ACED spanned three sequence types—arithmetic, geometric and other recursive sequences—commonly taught in 8th grade, but only the geometric sequence model is described here. The model is expressed as a tree shaped Bayesian network with the following proficiency variables:

Table 1: Proficiency Variables and their Mutual Information with the Overall Geometric Proficiency
The parent variable for each proficiency is the variable with fewer '+' markings above it in the list. Note that only the relative size of the mutual information is considered below.

Proficiency	**Mutual Information**
Solve Geometric Problems	1.39
+Visual Representations	0.34
+Examples	0.32
+Table Representation	0.26
+Model Geometric	0.25
+Common Ratio	0.20
+Extend Sequence	0.20
+Induce Rules	0.29
++ Verbal Rule	0.19
++ Algebra Rule	0.07
+++ Explicit Rule	0.02
+++ Recursive Rule	0.01

The model was constructed through expert (Graf) judgment about the correlation between the variables and their parents in the hierarchy. The mutual information (Nicholson and Jitnah, 1998) between the overall "Solve Geometric Problems" proficiency with each of the proficiency variables naturally decreases as you move down the hierarchy. Each variable can take on one of three proficiency levels: `Low`, `Medium` and `High`. The variables were chosen to reflect how the geometric sequences were represented in the tasks. There were 63 tasks in the geometric sequence portion of the assessment. In the evidence model for each task, the task outcome (evaluated as `right` or `wrong`) was directly related to (had as a parent) a single proficiency variable.

ACED is based on the National Council of Teachers of Mathematics standards which in turn form the basis of the standards of all 50 U.S. states. Although firmly based on those standards, true alignment is difficult to achieve because (a) ACED has a finer level of detail in it's (diagnostic) proficiency model than is found in most standards, and (b) all 50 states have set the cut point for "proficient" with the general category of sequences at different places. Thus, although performance on ACED should be strongly correlated with each state's standards, the `Medium` proficiency level in ACED may be higher or lower than the proficient point set in any given state.

The data used in the graphs below comes from an evaluation of ACED (Shute, 2006). It consists of data from 157 students who received the adaptive version of ACED. Roughly half the students received diagnostic feedback designed to help them understand their mistakes and the remaining half had accuracy-only feedback. For this paper, we will ignore the evaluation component of the study (including pre- and post-test measurements) and focus on the data that can be used to produce a collection of representative scores that a teacher might see. Note that for these students, geometric sequences were not an explicit part of the curriculum, although some geometric sequence problems may have been taught as part of other topics in algebra.

3 Scores coming out of a Bayesian Network

ACED scores student responses using the Bayesian network. The individual task outcome variables are entered as findings in task specific nodes and the results are propagated through the proficiency model. After evidence from all tasks are entered, the posterior proficiency model gives our beliefs about the proficiency state for this particular student. Technically, any *statistic*—that is any functional of that posterior distribution—can be used as a score. In practice the marginal distributions for each of the proficiency variables in Table 1 were recorded for each participating student.

As each variable can take on the values `Low`, `Medium`, and `High`, the "score" for any student is three numbers which sum to 1. To reduce the three numbers to a single number, we employed the "trick" of assigning numeric values -1, 0 and +1 to the three proficiency states, and taking the expected value. This *Expected A Posteriori* (EAP) score can also be written, $\mathrm{P}(\theta_{ij} = \texttt{High}) - \mathrm{P}(\theta_{ij} = \texttt{Low})$ (where θ_{ij} is the value for Student i on Proficiency j). The EAP score has a monotonic relationship to the score created through another method, Item Response Theory, which is commonly used to score high-stakes assessments (Hemat and Almond, Under Review).

An alternative to the EAP score is to report the mode of the marginal distribution. Thus, a student for whom

$P(\theta_{ij} = \texttt{Medium}) \geq P(\theta_{ij} = \texttt{High}), P(\theta_{ij} = \texttt{Low})$ would be reported as `Medium` on Proficiency j. While this score ignores information about our certainty about this classification, it has the advantage of simplicity. It could be further refined through two improvements. First, students for whom the modal probability differed from the next highest value by less than a threshold value (say 5%) should be identified as being on the cusp of gaining the next level. Teachers will want to pay special attention to these individuals. Second, when the marginal distribution is very flat (all states having roughly the same probability) the system should identify those individuals as ones about which it has a lot of uncertainty.

4 Group Level Plots

Teachers often want to look at the average performance for a group of students. Fortunately, the average of the probabilities has a very natural interpretation. The sum of the probability of each student being in the `High` state on a proficiency variable is the expected number of students at the `High` state. The average probability is then the expected proportion of students in the `High` state. This is expressed as a percentage to make it more accessible to teachers. Figure 1 shows how this might be depicted.

We use a split bar plot to represent the probabilities because humans are typically better at judging length than angles (Cleveland and McGill, 1987). One of these proficiency levels (in this case probability of `Medium`) is chosen as an anchor for the anchor line. The length of the bar below the line is the probability of being below the anchor state, and the length of the bar above the line is the probability of being at or above the anchor state. The colors are chosen as an intensity scale[1] because (a) this is least likely to present difficulties for viewers with limited color perception and (b) the figures are quite likely to be reproduced on a black and white printer or copier.

This plot poses some difficulties in interpretation for the intended audience. Seeing that 17% of the students are at the lowest category for the *Extend* skill, the natural question for the teacher is "Who are those students?" The answer coming from the Bayes net is a probability for each student, although the number of students classified at the `Low` state is likely to be similar to the numbers from the expected values. That is, the number of students in this sample in the `Low`, `Medium`, and `High` states according to the modal classification rule is 19, 3, and 3, compared to expected values of 18.75, 3.5 and 2.75. If many students are on the cusp, then these numbers could be off by as many as two or three students.

An alternative would be to first classify the students using the modal category for each ability and then count the number of students falling into each category. This sacrifices some precision in the estimates but is easier to explain. A plot similar to Figure 1 could be produced, but it has the advantage that it could be annotated with actual counts rather than percentages.

5 Individual Level Plots

An almost immediate question of a teacher when confronted with a plot like Figure 1 is "Which students are below the line (representing minimal proficiency)?" This is a complex question, as the bar below the line is made up of small fractions of all of the students. But the deeper question concerning identifying the students in need of additional help can be answered by simply plotting all of the scores for all of the students. Figure 2 is one realization of that idea.

Figure 2 is essentially a table of bar plots, one row for each student and one column for each proficiency variable. The bars are drawn horizontally to facilitate comparisons between students, particularly adjacent students. Some simple sorting helps to make patterns in the variables more apparent. Sorting by the probability of the overall proficiency variable puts all of the low performing students up near the top of the display. (The ability to dynamically re-sort the table would be useful as teachers might find other sortings useful for other purposes: for example, alphabetical sorting helps in finding students by name). The columns are sorted according to their mutual information with the overall proficiency variable, but there may be better ways to do this sorting (particularly if there is knowledge about the way the skills are ordered in the curriculum).

Figure 2 is a complex report and may take a bit of training before teachers are comfortable interpreting it. Some training could be done through linking the display to explanatory information (i.e., linking the proficiency names to definitions and explanations). But a better approach would be to produce a natural language summary of important features in the display. For example, Student S276 seems to do better than expected with tabular representations and Student S258 seems to do better than expected with pictorial representations. This might suggest instructional strategies for those students. Automatically generating such natural language descriptions is an interesting

[1] This can be difficult to get right because of the difference between rendering devices. In particular, on-screen rendering using X11 and PDF-based graphics requires different color values to give good visual discrimination.

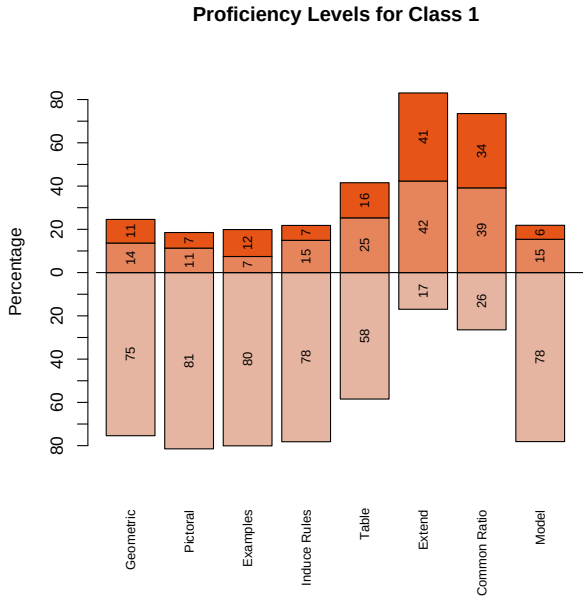

Figure 1: Bar plot for a "classroom" of 25 students.
A group of 25 students were selected randomly from the ACED data to form a "class" and the average marginal distributions are reported for this class. The shading in the bars gives the expected proportion of students in the high, medium, and low ability groups for each skill. The numbers in the bars give the expected proportion as a percentage. The bars are offset so that the percentage of students above the low ability is the height of the bar above the reference line.

task beyond the scope of this paper.

6 Comparing Groups

There are a number of different questions a teacher might want to ask which basically involve comparing groups of students. One kind of question involves comparing the current classroom to a larger group (e.g., school, district, or state). A second involves comparing groups within a classroom (e.g., predefined ability groups). If there are only two groups then the stack bars can be placed side by side for easy comparison. Figure 3 shows an example. The same graphical display could be used to compare an individual student to the class.

Note that the "School" in Figure 3 is the entire sample collected by Shute (2006), and the classroom is a randomly selected group of 25 students. Therefore, although it appears that the "class" is doing slightly worse than the school overall, the difference is not one which should cause concern. Teachers and administrators would require guidance about which differences are meaningful and which can be explained by "sampling."

As the number of groups increases, plots like Figure 3 grow increasingly crowded. One way of getting around this problem is to generate separate plots for each of the proficiency variables. Figure 4 shows one possible realization of this idea, comparing students by their mathematical performance level, as assigned by their school.

One way to improve Figure 4 would be to add indication of the size of the groups. For example, there is no indication in the current design that there are 88 students in the `Academic` track and only 5 in the `Part 2` track. This could be done through a legend in the plot or by varying the width of the bars.

7 Associations Among Scores

A fair amount of understanding of the relationships among the variables can be found simply by plotting the EAP scores for various nodes against one another. The scatterplot matrix (Figure 5) plots all possible pairs of proficiency variables. The central column gives the abbreviated names of the variables that are plotted on the x-axis (that column) or y-axis (that row).

Some interesting patterns emerge from this graph. Note the strong inverted L-shape in the plots involving either "EAPCR" (Common Ratio proficiency) or "EAPExtG" (Extend Geometric proficiency) and one of the other variables. This indicates that these skills

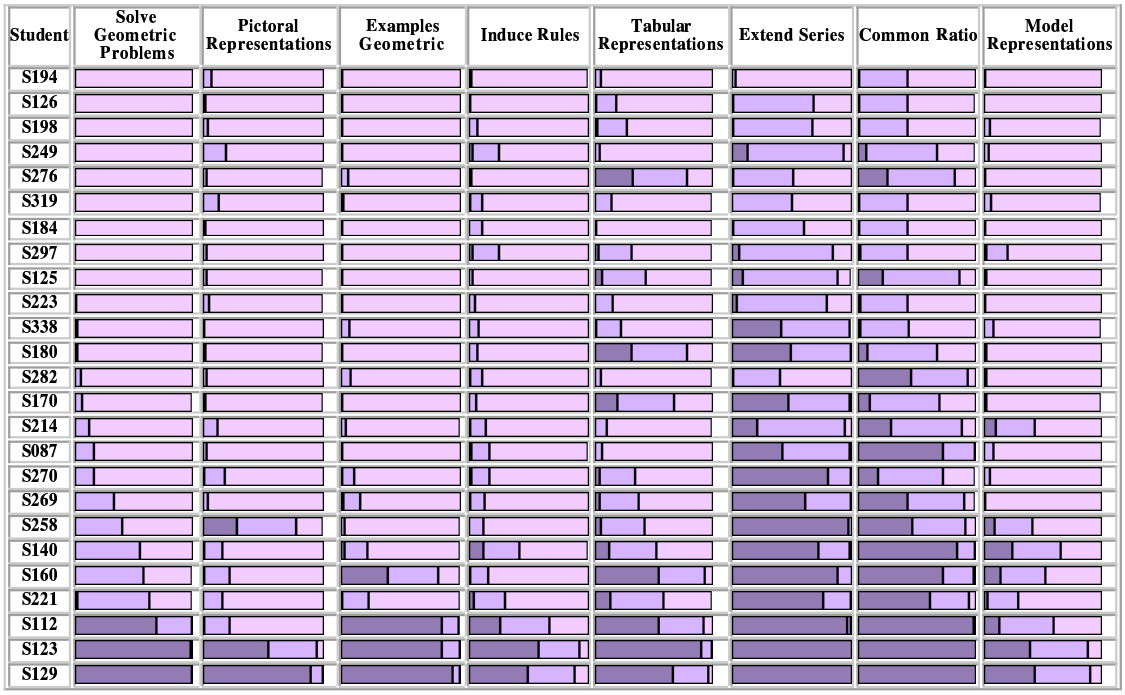

Figure 2: Score Profiles for 25 ACED students.
The rows in this bar plot matrix correspond to students and the columns correspond to the individual proficiencies. The bar plots are draw horizontally to facilitate comparisons among students. Dark bars represent the `High` *category and light bars the* `Low` *categories. The rows are sorted by the probability that the overall ability is high, columns are sorted by mutual information with overall ability.*

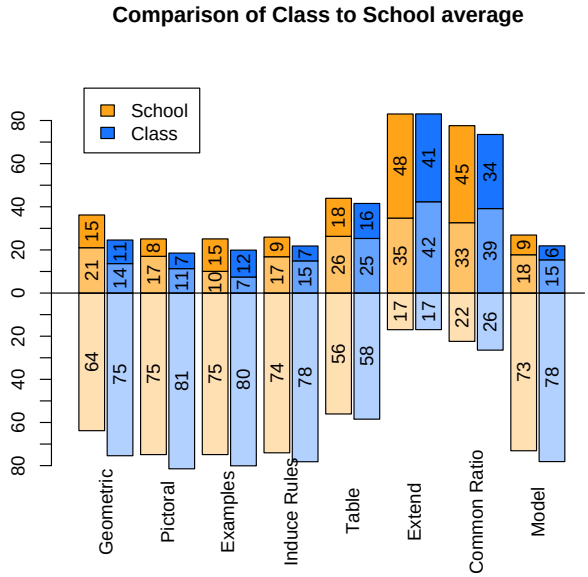

Figure 3: Comparison of Class to School.
This plot compares all the students in the ACED evaluation ("School", orange bars on the left) to the randomly selected "Classroom" full of students (blue bars on the right). Each pair of bars corresponds to a proficiency variable.

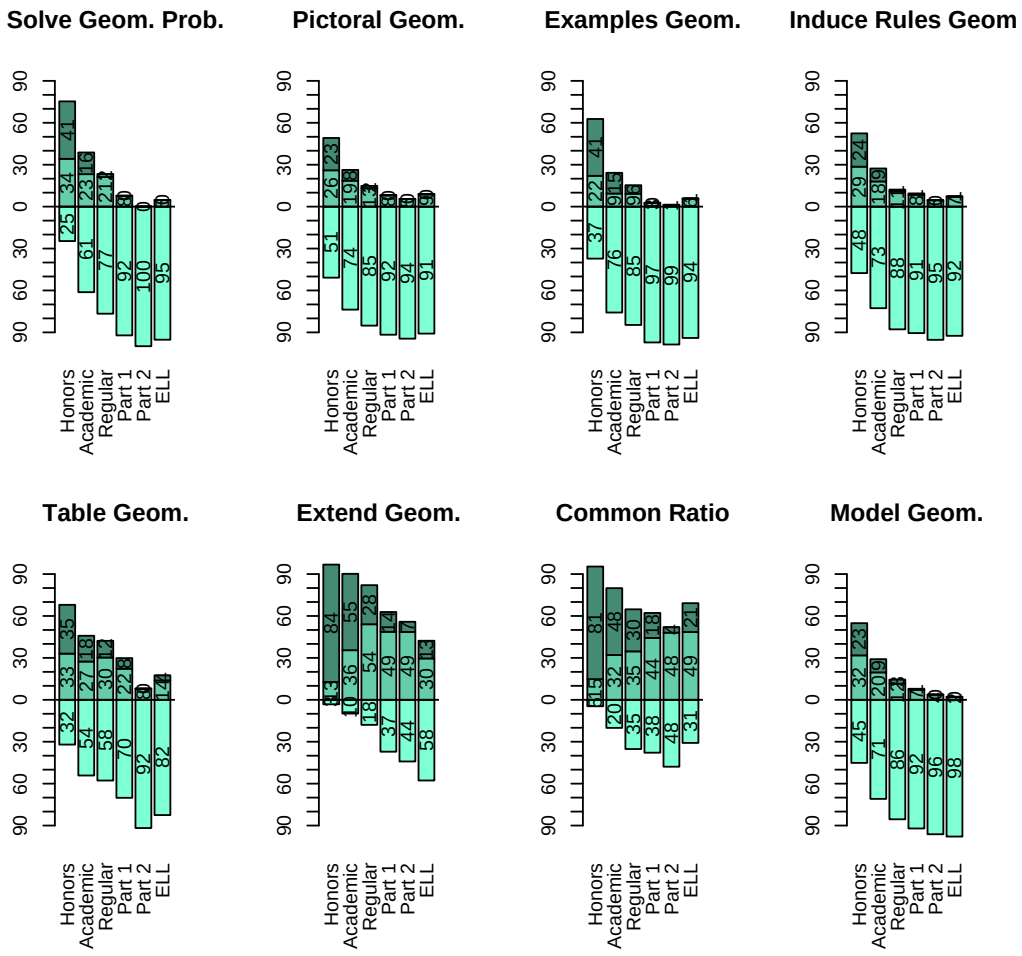

Figure 4: Comparisons Among Six Student Ability Groups.
Each plot compares six different ability groups for one of the proficiency variables. The groups are based on Student Performance Level and include: Honors, Academic, Regular, Part I (Special Education Students who are mainstreamed), Part II (Special Education Students who are sheltered), and ELL (English Language Learners).

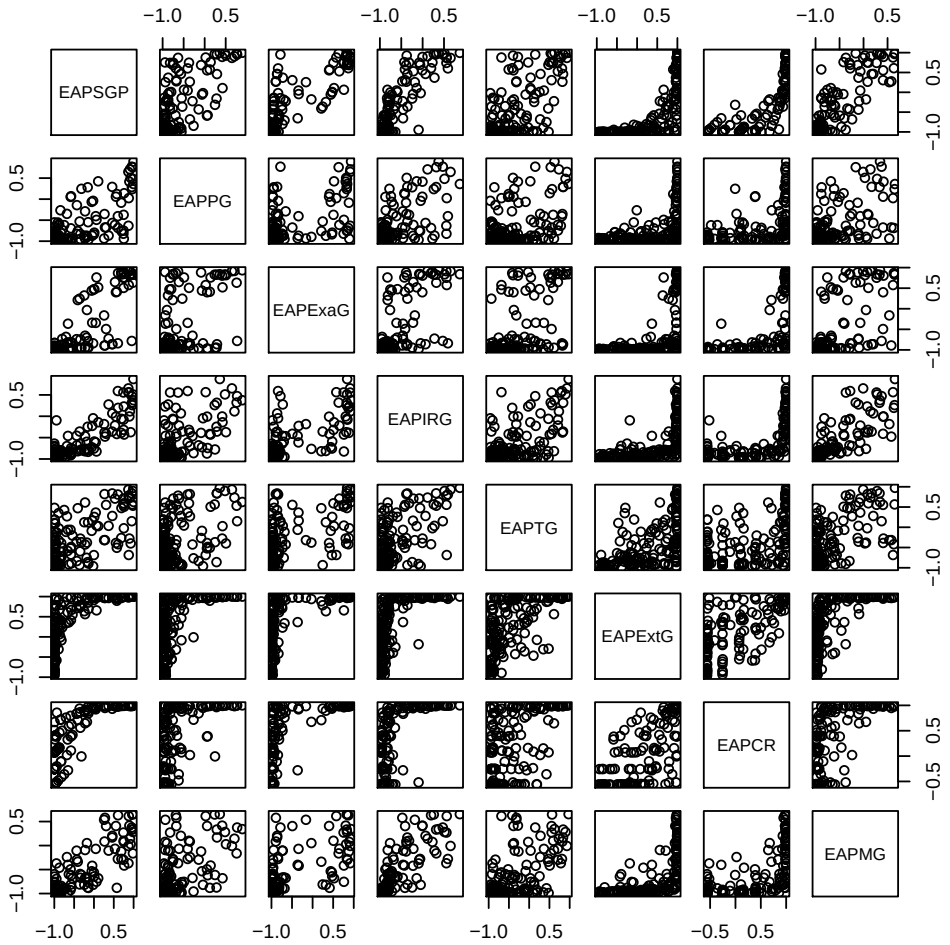

Figure 5: Scatterplot Matrix for All Students.
This scatterplot matrix shows all possible pairings of proficiency variables. Each cell in the matrix plots the EAP score for one proficiency variable against another using the scores for all the students in the study as the data points. The labels in the diagonal indicate which variables is in the corresponding column (x-axis) and row (y-axis).

are usually acquired before the others (this fits well with the intuition of the math expert on the team). In many respects, however, this is a display more suited to the researcher than the teacher. There is much going on in this graph that requires further study to validate and clarify the findings.

8 Unanswered Questions

In many ways, the questions which have been answered by the plots shown above are the easiest and most obvious. More research needs to be done into how to best answer the more complex questions. A few ideas are presented below:

Are there individual differences in the acquisition of subskills that deserve attention? and *Are there individuals with atypical patterns which require special attention?* The scatterplot matrix actually does a good job of helping to identify groups which seem to be behaving similarly and differently. One possible method for identifying individuals with unusual patterns of skills might be to look for outliers in the natural regression models for each skill. Some care is needed as many of the relationships appear to be non-linear.

How much credence should I put in the estimates from the Bayes nets relative to other sources of information? This is again a difficult question to answer. ACED shows good predictive validity for a post-test on geometric sequences (Shute, 2006), but the reliability and validity will vary with the length of the assessment. The situation is more complex when the comparisons are group comparisons as both variability due to the assessment instrument and due to the groups sizes need to be expressed. Another substantial problem is how to express the reliability of the assessment in the display without getting in the way of the primary comparisons.

How should performance levels be interpreted and validated? This is obviously a key question when fielding an assessment using a Bayesian network based scoring engine. In the ECD process, proficiency variables are defined through claims that are made about students at that proficiency level. Those claims must be validated both through internal design constraints on the assessment and through studies that link them to their intended use. The biggest challenge is how to help the viewer understand whether or not the use to which they intend to put the scores is supported by the current validity evidence.

What should I do next? This is a hard question to answer. In the sample class, "Tabular Representations" seems like a good candidate (as it has the highest probability among all non-mastered skills), but there might be good pedagogical reasons for teaching another skill first. Ideally we should be able to use the inferences from the Bayes net as input to a planning system to help suggest next steps for the teacher.

Our next steps are clear. We should put these graphs in front of teachers and perform a usability study of the representations. This will identify ways to improve the graphical displays and needs of the teachers which are not met by current graphics.

References

Cleveland, W. and McGill, R. (1987). Graphical perception: The visual decoding of quantitative information on graphical displays of data. *Journal of the Royal Statistical Society, Series A*, 150:192–229.

Graf, E. A. (2003). Designing a proficiency model and associated item models for a mathematics unit on sequences. In *Paper presented at the Cross Division Math Forum*, Princeton, NJ.

Hemat, L. A. and Almond, R. G. (Under Review). Irt versus bayes nets: Proficiency estimates and parameter recovery. Research report, Educational Testing Serivce. Under Review for Publication.

Madigan, D. and Almond, R. (1995). Test selection strategies for belief networks. In Fisher, D. and Lenz, H., editors, *Learning from Data: AI and Statistics V*, pages 89–98. Springer-Verlag.

Nicholson, A. and Jitnah, N. (1998). Using mutual information to determine relevance in Bayesian networks. In *Pacific Rim International Conference on Artificial Intelligence*, pages 399–410.

Shute, V. J. (2006). Asessments for learning: Great idea, but do they work? In *Paper presented at the annual meeting of the American Educational Research Association (AERA)*.

Shute, V. J., Graf, E. A., and Hansen, E. G. (2005). Designing adaptive, diagnostic math assessments for individuals with and without visual disabilities. In Pytlikzillig, L. M., Bruning, R. H., and Bodvarsson, M., editors, *Technology-based education; bringing researchers and practitioners together*, pages 169–202. Information Age Publishing, Greenwich, CT.

Acknowledgments

Several people have contributed data, ideas and suggestions which have improved this paper, particularly, Aurora Graf and Eric Hansen. ACED development and data collection was sponsored by National Science Foundation Grant No. 0313202.