International Journal of Approximate Reasoning 50 (2009) 450–460

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

journal homepage: www.elsevier.com/locate/ijar

Bayesian networks: A teacher's view

Russell G. Almond *,[1], Valerie J. Shute [2], Jody S. Underwood, Juan-Diego Zapata-Rivera

Florida State University, ETS, Princeton, NJ, United States

ARTICLE INFO

Article history:
Received 1 February 2007
Received in revised form 29 October 2007
Accepted 21 April 2008
Available online 27 June 2008

Keywords:
Bayesian networks
Computer graphics
Probabilities
Aggregation

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.

Published by Elsevier Inc.

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 [1]). 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 of 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 (or school or district) 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? Other classes in the same school or district? Classes in other districts with similar characteristics (where similar characteristics will depend on the purpose)?
- How do previously identified groups within the classroom differ?

* Corresponding author.
E-mail addresses: ralmond@ets.org (R.G. Almond), vshute@fsu.edu (V.J. Shute), junderwood@ets.org (J.S. Underwood), dzapata@ets.org (J.-D. Zapata-Rivera).
URL: http://ralmond.net (R.G. Almond).

[1] Several people have contributed data, ideas and suggestions which have improved this paper, particularly, Aurora Graf and Eric Hansen.

[2] Val Shute is now at Florida State University.

0888-613X/$ - see front matter Published by Elsevier Inc.
doi:10.1016/j.ijar.2008.04.011

- What are the typical patterns of skill acquisition (this is a more advanced question that might be asked when doing curriculum design or professional development, but not for routine classroom teaching)?
- 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? What should I do differently?

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 [1].

2. ACED

ACED (Adaptive content with evidence-based diagnosis [1,2]) is a computer-based assessment-for-learning system covering the topic 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 [3] algorithm to select the next task in an assessment, (b) the use of targeted diagnostic feedback, and (c) the use of technological solutions to make the assessment accessible to students with visual disabilities.

Graf [4] 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 proficiency variables given in Fig. 1. Each variable can take on one of three proficiency levels: low, medium, and high.

The model was constructed through expert (Graf) judgment about the correlation between the variables and their parents in the hierarchy. 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 tasks are based on the National Council of Teachers of Mathematics standards which in turn form the basis of the standards of all 50 US states. Although firmly based on those standards, true alignment is difficult to achieve because (a) ACED has a finer level of detail in its (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 cut point set in any given state.

The data used in the graphs below comes from an evaluation of ACED [2]. It consists of data from 157 students who received the adaptive[3] version of ACED. Roughly half the students received diagnostic feedback designed to help them understand their mistakes; 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 a functional[4] of that posterior distribution—can be used as a score. In practice, the marginal distributions for each of the proficiency variables in Fig. 1 were recorded for each participating student.

Let $S_{i0},\ldots,S_{ik}$ be the proficiency variables for student i, with S_{i0} representing the special overall proficiency variable (SOLVE GEOMETRIC PROBLEMS in ACED). Each variable can take on the values low, medium, and high. Given a body of evidence, $\mathbf{X}_i$ the Bayes net can efficiently calculate $p(S_{ik}|\mathbf{X}_i)$, the conditional distribution of S_{ik} given the observed outcomes. Four statistics of this marginal distribution are of particular interest.

Margin: The marginal distribution of the proficiency, $p(S_{ik}|\mathbf{X}_i)$. This has the disadvantage of being three numbers (summing to 1.0) so it is not compact to display or simple to interpret.

Cut: If one of the states has special meaning, e.g., students at the medium level or above are considered to be "proficient" on some set of standards, then the probability that the student is proficient, $P(S_{ik} \geqslant \text{medium}|\mathbf{X}_i)$, is just a sum of

[3] The adaptive version of ACED used the Bayes net to select the item sequence based on the pattern of scores observed so far, and hence Bayes net scores were more readily available [2].

[4] A *functional* is an operator which maps a function (e.g., a probability distribution) to a scalar quantity.

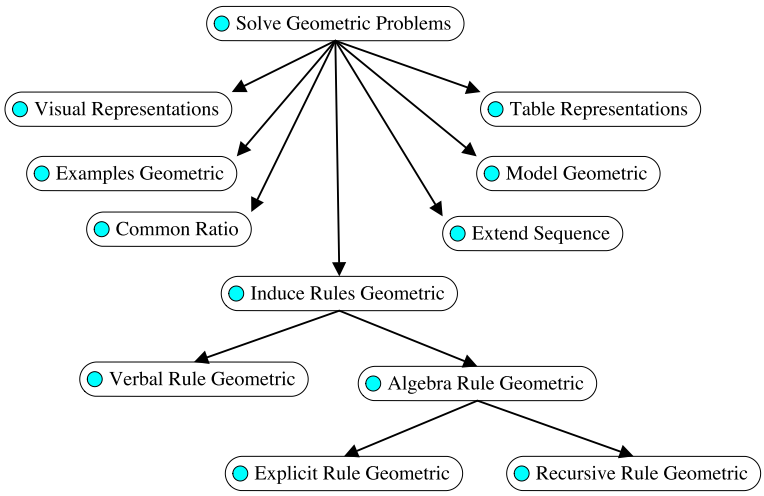


Fig. 1. Proficiency model for ACED (geometric branch only).

selected marginal probabilities. Note that in a dynamic presentation environment, the cut score could be set to different values (medium or high) by the person viewing the report to provide different views on the data.

Mode: The value of m which maximizes $P(S_{ik} = m|\mathbf{X}_i)$. This is easy to interpret, but ignores information about our certainty of the classification. It could be further refined through two improvements. First, students for whom the modal probability differed from the next highest (or lowest) 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 reporting system should identify those individuals as ones about which it has a lot of uncertainty.

EAP: By assigning numbers to the states, high = 1, medium = 0, low = -1, one can take an expectation over the posterior, the *expected a posteriori* (EAP) score. $1*\Pr(S_{ik} = \text{high}|\mathbf{X}_i) + 0 *\Pr(S_{ik} = \text{medium}|\mathbf{X}_i) + -1* \Pr(S_{ik} = \text{low}|\mathbf{X}_i)$, which in the case of three level reduces to $\Pr(S_{ik} = \text{high}|\mathbf{X}_i) - \Pr(S_{ik} = \text{low}|\mathbf{X}_i)$. The EAP score has a monotonic relationship to the score created through another method, item response theory (IRT), which is commonly used to score high-stakes assessments [5].

Fortunately, all four statistics are relatively simple to aggregate to form summaries for classes and other groups of interest.

Margin: Summing the marginal distributions across students, $\sum_i p(S_{ik}|\mathbf{X}_i)$ produces expected numbers of students in eachproficiency level. Thus, it has a straightforward interpretation; however, "fractional students" can be difficult to explain to some audiences (rounding alleviates this problem somewhat at the cost of precision). It also can be divided by the class size to produce an average proficiency for the class.

Cut: The average cut score is the expected proportion of "proficient" students in the class.

Mode: To aggregate the modal scores, we simply count the number of students assigned to each category. Alternatively, we can divide by the class size to produce a percentage. However, the uncertainties about the classifications also accumulate with this statistic, giving the aggregate statistic a large standard error.

EAP: The average of the EAP scores has a straightforward interpretation as the average ability level in the class. The standard deviation of the EAP scores gives an indication of the variability of proficiency in the classroom, however, many teachers would require additional training to effectively use the standard deviation.

Note that the potential audience for classroom level reports could include individuals with a wide range of statistical abilities [6,7]. Groups of teachers often analyze the results from classroom assessments together as a professional development activity. Some teachers learn to interpret well, and some teachers still have problems interpreting [7]. Reviewing the results is more productive when the relationship between the scores and the curriculum is clear, and when the scores are presented using clear and focused representations.

Consequently, it is worth carefully considering how the scores are scaled and how that impacts their interpretation. Counts and expected counts are probably the easiest for lay audiences to understand, followed by percentages and probabilities (with percentages being more familiar than probabilities), followed by the EAP score. One could also use any

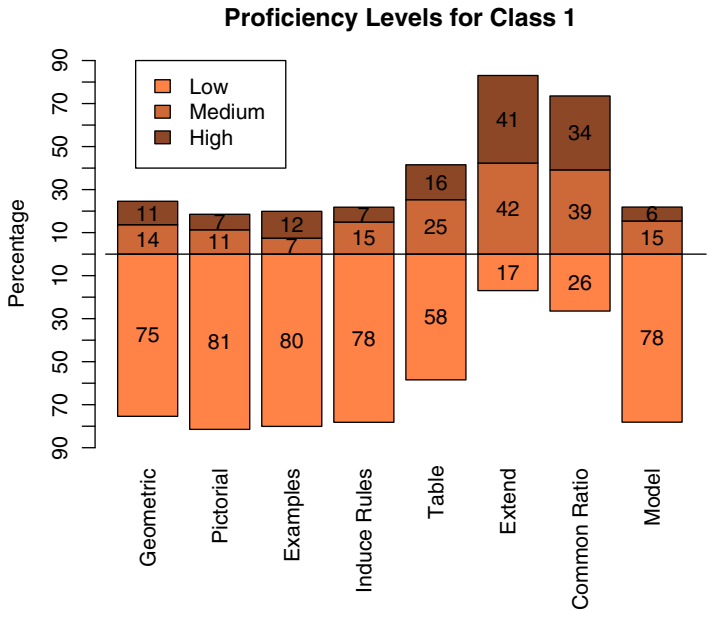


Fig. 2. Bar plot for a "classroom" of 25 students. The shading in the bars gives the expected proportion of students in the high (dark), medium, and low (light) ability groups for each skill. The numbers in the bars give the expected proportion as a percentage. The bars are offset so that the percentage of students above the low ability is the height of the bar above the reference line.

monotonic transformation (linear or non-linear) of those scales. In choosing the reporting scale for various statistics, the designers must take into account how much training the teachers will need in order to interpret the plot.

Another possibility is to use a fixed set of words to describe the probability. For example, the report might state, "it is *likely* ($p = 0.67$) that this student has reached the medium proficiency level". Mosteller and Youtz [8] summarize a body of research on general probability levels associated with common expressions which might form the basis for generating such a vocabulary. In our experience, it is best to include graphical, numerical and natural language information on score reports as various members of the audience will work best with different modes of presentation.

The margin and cut statistics present another difficulty. If a teacher learns that 30% of the students are not proficient, the natural next question is which ones? This cannot be simply answered with a list of names, as many students were fractionally counted to make up the aggregated scores. One answer to this question is an ordered list of individual student scores (see Section 5).

4. Group-level plots

Teachers often want to look at the average performance for a group of students. A group of 25 students were selected randomly from the ACED data to form a "class" and the average marginal distributions were calculated for this class. Fig. 2 shows how these data might be depicted.

We use a hanging bar plot to represent the probabilities because humans are typically better at judging position against a common scale than judging length which is in turn better than judging angles [9] (pie charts, which are often used for conveying proportions, require the teacher to use the perceptually more difficult skill of comparing angles). The hanging bar chart makes the perceptual task of comparing scores for different nodes at two different probability levels (proportion of students at the low level and proportion of students at the medium levels) one of judging position rather than length. One of the proficiency levels (e.g., probability of medium) is chosen as a cut point for the anchor line (if the plot was presented in a dynamic environment, then the viewer could select different cut points 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 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.[5] Fig. 2 meets these goals by varying the colors primarily on the saturation scale in the standard hue saturation value color model (see [10] for a more thorough discussion of color models).

This plot poses some difficulties in interpretation for the intended audience. For example, 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

[5] In this case, the journal is printed in black and white. Color versions of the figures can be obtained by writing the authors.

coming from the Bayes net is a probability for each student that they do not make the cut. Although the number of students classified at the low state is likely to be similar to the numbers from the expected values, it can differ due to rounding errors. In this example, the number of students 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 Fig. 2 could be produced, with the advantage that it could be annotated with actual counts rather than percentages.

5. Individual level plots

As mentioned above, an almost immediate question of a teacher when confronted with a plot like Fig. 2 is "which students are below the line (representing minimal proficiency)?" The answer to this question is complex, as the bar below the line is made up of small fractions of all of the students. But the deeper question concerns the identification of students in need of additional help. This can be answered by simply plotting all of the scores for all of the students. Fig. 3 is one realization of that idea.

Fig. 3 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 (a measure of association [11]) 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.

Fig. 3 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 students at a similar overall ability level with tabular representations and student S258 seems to do better than others with a similar overall ability with pictorial representations. This might suggest specific instructional strategies for those students. Also, from this plot it appears that the EXTEND SERIES and COMMON RATIO proficiencies are acquired before (for students at lower levels on the overall proficiency scale) the others. This might suggest general instructional strategies (see Section 7). Automatically generating such natural language descriptions is an interesting task beyond the scope of this paper.

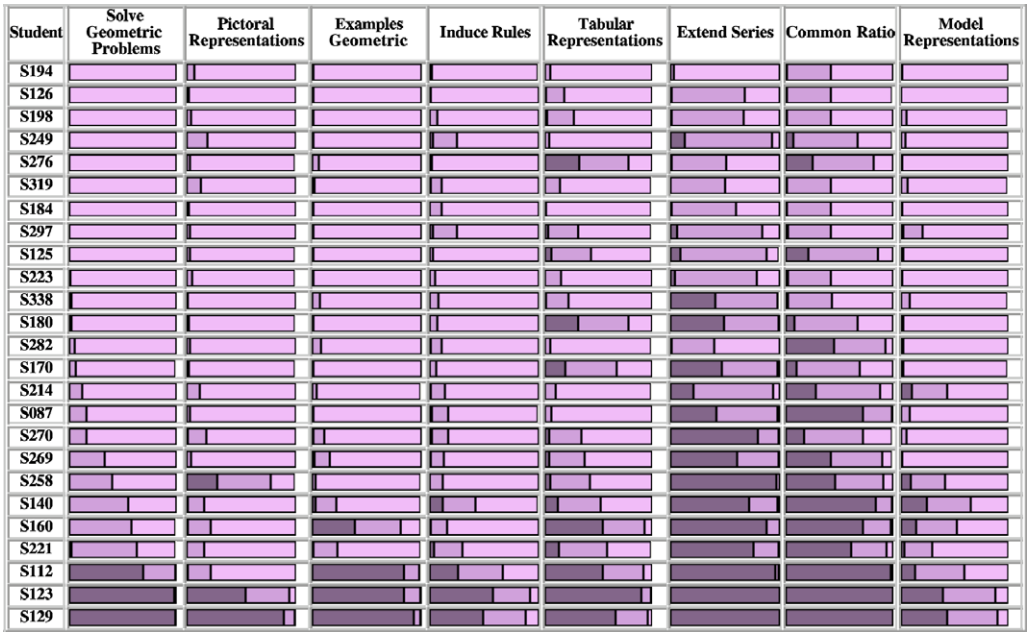


Fig. 3. Score profiles for 25 ACED students. The rows in this bar plot matrix correspond to students and the columns correspond to the individual proficiencies. The bar plots are drawn horizontally to facilitate comparisons among students. Dark bars represent the high category and light bars the low categories. The rows are sorted by the probability that the overall ability is high, columns are sorted by mutual information of that variable with the overall ability variable.

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 stacked bars can be placed side by side for easy comparison. Fig. 4 shows an example. The same graphical display could be used to compare an individual student to the class (in this case, the numbers would be interpreted as the probability of being at that level for the student, and the average probability of being at that level across the classroom. This is a slightly different interpretation and may require additional training for the teachers).

Note that the "School" in Fig. 4 is the entire sample collected by [2], 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,

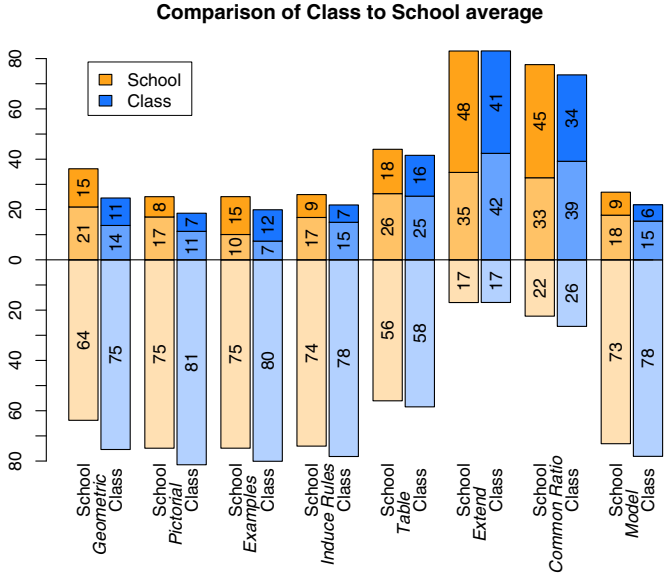
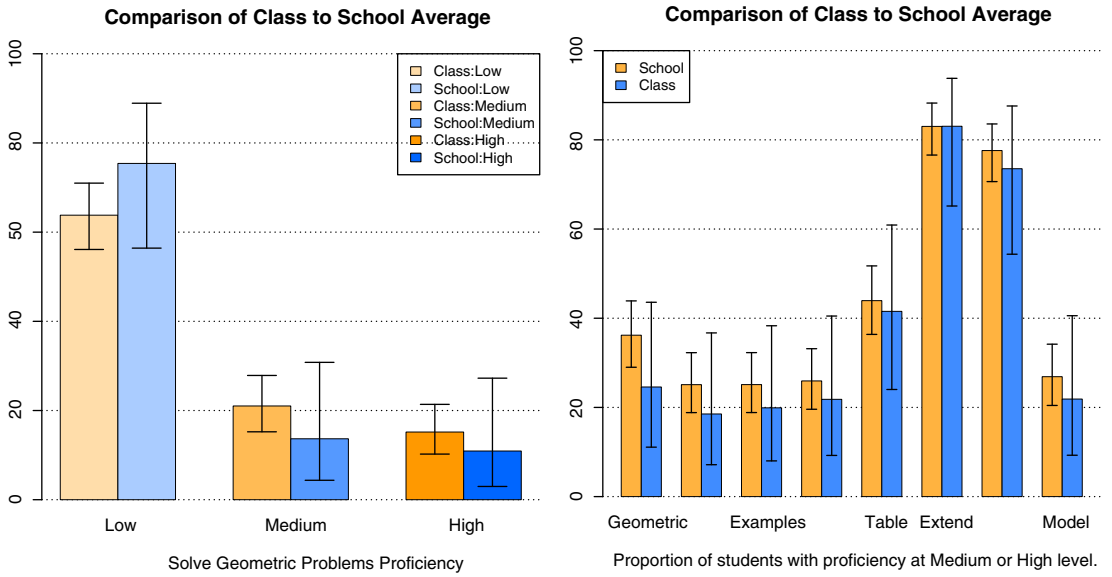


Fig. 4. Comparison of class to school. This plot compares all the students in the ACED evaluation ("school", bars on the left of each pair) to the randomly selected "classroom" full of students (bars on the right of each pair). Each pair of bars corresponds to a proficiency variable.



(a) Concentrate on one proficiency (b) Concentrate on one skill level.

Fig. 5. Two approaches to adding error bars to the comparison barplot.

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," that is, by the assignment of student into classes.

One possibility for indicating which comparisons are important is to add error bars to the plot. An ad hoc procedure for doing this is to set up a beta prior (say, a Jeffreys prior of (0.5,0.5)) for the proportion of students falling into a category and use the expected counts as data. As the beta distribution whose parameters are any positive real numbers is well defined, it is no problem to produce a posterior distribution even with fractional counts in the data. Thus a highest posterior density credibility interval [12] could be calculated for the class proportion of students at each level (note that this is not a particularly realistic model, but rather a heuristic for getting some idea of the sampling variability. A better approach would be to do some kind of hierarchical modeling along the lines of [13]). Adding error bars to a hanging barplot requires some care. It is not the probability of each category but the sum of the probabilities from the anchor line which is of interest. Thus the three probabilities we need estimates for are: $\Pr(S_k = \text{low})$, $\Pr(S_k = \text{medium})$ and $\Pr(S_k \in \{\text{medium,high}\})$.

Unfortunately, adding all of those error bars to Fig. 4 produces a plot which is too busy to be readily interpreted. Two simplifications are possible. First, one could look only at the comparison of just one proficiency at a time. Fig. 5a shows the comparison for the SOLVE GEOMETRIC PROBLEMS proficiency (note that this plot requires some training to interpret as the probabilities of low and high run in opposite directions—one is bad, the other is good. One possible remedy is to have the low bar hang below the x-axis). Second, one could choose a cut point in the proficiency scale and only look at scores above that cut point. Fig. 5b shows the comparison for all proficiencies at the medium level or above. In a dynamic reporting environment, these could be selected by the viewer from the overall comparison graph (Fig. 4) by clicking on an an appropriate place in the plot. In that case it would be good to have some visual indication that a particular contrast is worth looking at (say by highlighting the labels of the bars when the mean of the larger group falls outside of the credibility interval for the smaller group).

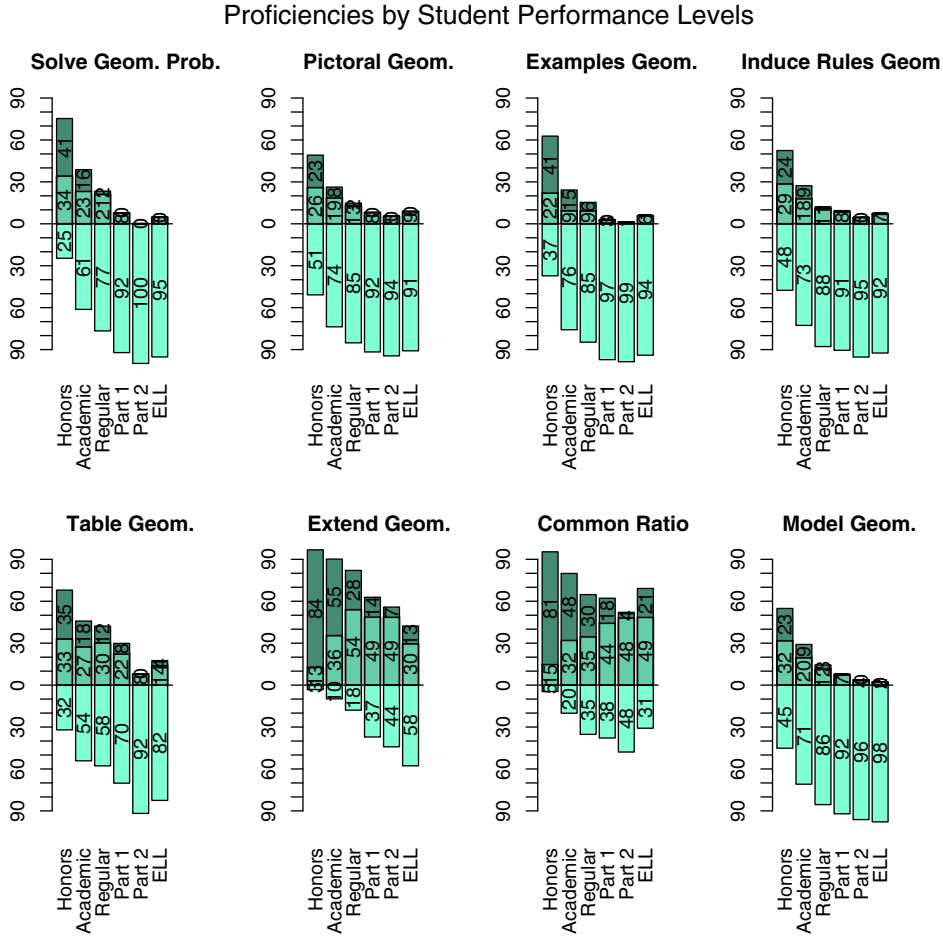


Fig. 6. Comparisons among six student ability groups. Each plot compares six different ability groups for one of the proficiency variables. The groups are based on student performance level and include: honors, academic, regular, part I (special education students who are mainstreamed), part II (special education students who are sheltered), and ELL (English language learners).

The expected proportions for the class and the marginal probabilities for an individual student have very similar structures. Thus, it should be possible to produce a similar plot comparing an individual student to a class. However, studying such plots for all 25 students is a fair amount of work. A better solution might be to add some kind of flag to a class list, such as the one presented in Fig. 3, to highlight students which the reporting system thinks show unusually large differences from the class average.

In many cases, teachers or administrators would like to compare more than two groups. However, as the number of groups increases, plots like Fig. 4 grow increasingly crowded. One way of getting around this problem is to generate separate plots for each of the proficiency variables. Fig. 6 shows one possible realization of this idea, comparing students by their mathematical performance level, as assigned by their school.

One way to improve Fig. 6 would be to add an indication of the size of the group. For example, there is no indication in the current design that there are 88 students in the academic track and only five in the Part 2 track. This could be done through a legend in the plot or by varying the width of the bars. In the latter case, reducing the number of plots per page would increase legibility.

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 (Fig. 7) plots all possible pairs of proficiency variables. It is intended to resemble a correlation matrix where each cell give the correlation, as a scatterplot, between two of the 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). For example, all of the plots in the first column have EAP (SOLVE GEOMETRIC PROBLEMS) on the *x*-axis and all of the plots in the

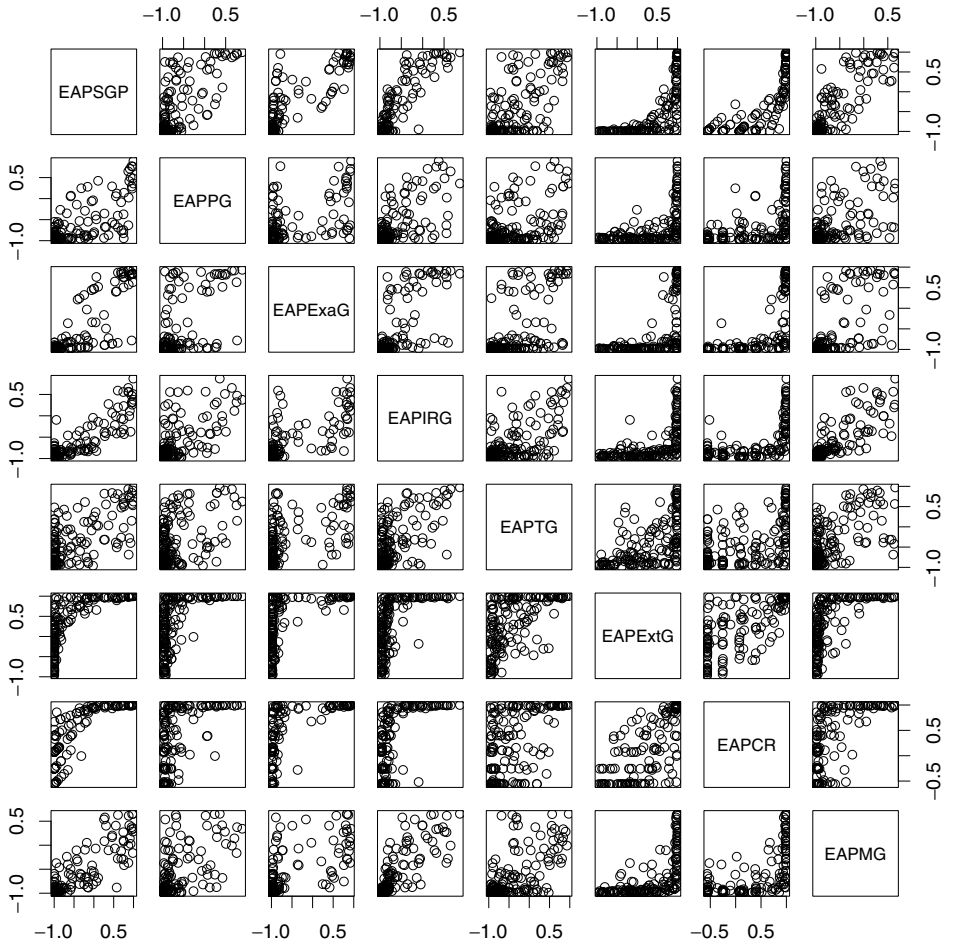


Fig. 7. Scatterplot matrix for all students. This scatterplot matrix shows all possible pairings of proficiency variables. Each cell in the matrix plots the EAP score for one proficiency variable against another using the scores for all the students in the study as the data points. The labels in the diagonal indicate which variables are in the corresponding column (*x*-axis) and row (*y*-axis).

seventh row have EAP (COMMON RATIO) as their *y*-axis. Like a correlation matrix, it is symmetric along the main diagonal, so that the plot in the first row and seventh column is a mirror image of the one in the seventh row and first column.

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 most of the other variables. This indicates that these skills are usually acquired before the others. A similar pattern can be seen in Fig. 3 where the columns associated with these variables have substantially more dark values (associated with the high state) than do the other columns. When these result were shown to the expert who constructed the original model, she thought that it matches her intuition about how math skills are acquired (Aurora Graf, private communication, July, 2006). However, these are observational findings and it would be good to verify them with specifically targeted experiments.

In many respects, however, this is a display more suited to the researcher than the teacher. A teacher, with sufficient training, might find it of use for curriculum design or instructional planning, but it is not likely to be of immediate use in day-to-day decision making in the classroom.

8. Reliability and validity

A teacher will typically have many sources of information about a student. When integrating the information from ACED, the teacher needs to know both about the amount of measurement error in the assessment, its *reliability*, and degree to which it suits the purpose for which it is being used, its *validity*. Both of these have specific meanings in psychometrics.

The reliability of an assessment is based on a definition from classical test theory which states:

$$\text{ObservedScore} = \text{TrueScore} + \text{Error}. \tag{1}$$

The *reliability* of the assessment is defined as the correlation of the OBSERVEDSCORE with the TRUESCORE. This is obviously related to the signal-to-noise ratio concept used in signal processing.

Table 1 gives the reliabilities for the EAP scores from ACED [2]. These numbers are respectably high. In the case of ACED this is due mainly to the fact that geometric sequences is a narrow and quite focused domain. Shute et al. (op cit.) note that the reliability of a number right score from all ACED items matches the reliability of EAP score for the overall SOLVEGEOMETRIC-PROBLEMS up to rounding error.

At first glance, the equal reliabilities seem to indicate that the Bayes net scoring does not offer significant improvements over the number right scoring. However, note that the reliability of the subscores is almost as high as the reliability of the overall test. The same is not true if we use simple number right subscores. For example, there are six COMMONRATIO tasks. Applying the Spearman–Brown formula for adjusting reliability for test length with a ratio of 6/63 yields a reliability of 0.41 for the corresponding number right score. It appears that the Bayes net estimates are stabilizing the subscores by borrowing strength from the reliability of the overall score. In this respect, the Bayes net score is similar to other proposed methods for augmenting the strength of subscores [14–16]. The major difference is that the Bayesian network approach starts with an expert constructed model of how the proficiencies interact, while the other methods use observed correlations between the scores on various subtests.

While reliability looks at the internal consistency of an assessment, validity looks at its relationship to some external criteria. Messick [17] defines two ways to establish validity: internal construct validity and predictive validity. Internal construct validity is related to documenting the basis for why each task provides evidence for the construct to be measured. Ref. [1,2] document ACED's construct validity.

Predictive validity is usually documented by correlating the test score with some external measure of interest. Ref. [2] studied the validity of ACED when used to predict the scores on a post-test of a student's ability to solve geometric sequence problems. The correlation between the post-test score and EAP (SOLVEGEOMETRICPROBLEMS) score was 0.65, which is very close to the reliability of the post-test (a practical upper bound).

Table 1
Reliability for ACED EAP scores by proficiency variable

Proficiency variable	Reliability
Solve geometric problems	0.88
+Visual representations	0.82
+Examples	0.92
+Table representation	0.82
+Model geometric	0.80
+Common ratio	0.90
+Extend sequence	0.86
+Induce rules	0.78
++Verbal rule	0.67
++Algebra rule	0.76
+++Explicit rule	0.62
+++Recursive rule	0.76

Variables have been sorted by mutual information with the SOLVE GEOMETRIC PROBLEMS variables, and the number of '+' signs in front of the name indicates the distance of that variable from the top of the tree in the proficiency model.

Note that validity is very much a function of the purpose for which the teacher and/or administrator wants to use the assessment. Although the high correlation with the post-test imply that ACED is a highly valid assessment of geometric sequences, it is much weaker evidence for its use as an assessment of algebra or pre-algebra. Frequently the intent of the test designers does not exactly match the needs of the test users. The educators using the assessment must review the validity and reliability evidence for a particular assessment to see if it is suitable for their particular purpose.

The standards on educational and psychological testing [18] call for publishing score interpretation guides that include information about reliability and validity studies that have been performed. However, the authors anecdotal evidence from focus groups is that teachers and administrators often do not pay much attention to this information. Although teachers and administrators should look at information about these factors when selecting an assessment, the reliability does not present the information about test scores in a form that is well suited for the typical tasks faced when interpreting an individual score or comparing two scores. Here the standard error of measurement (which is related to the reliability through Eq. (1)) is more immediately useful to the problem at hand.

Although training in educational measurement should be part of the professional development of both teachers and administrators, they are often unable or uninterested in using information about the measurement error. The situation is worse with lay audiences; parents, students, journalists, and politicians often have not received even minimal training in interpreting test scores. It is obvious the psychometricians place much more importance on reporting standard errors than do the score users.

One solution is to try and add error bars to the graphics along with just in time training for their use. However, as mentioned above, error bars add to the visual clutter of plots. When the displays are incorporated into a dynamic environment, we recommend including the error bars to a detailed plot that focuses on a particular contrast of interest. In such an environment, the overall plot should contain some kind of visual indication of which contrasts are significant and hence worthy of closer inspection.

When looking at group-level scores, the reliability of the assessment is only one contributing factor to the observed variability. In particular, if the students have been assigned to classrooms by some arbitrary mechanism we expect some measure of variability in the group scores due to that assignment process as well. This was apparent in Fig. 4 when the class was a randomly generated subset of students. Conveying the expected magnitude of differences due to class assignment is an important role of error bars on the plots, however, such error bars can get in the way of interpretation.

The discussion here has just tapped the surface of the problem of describing sources of variability, which are critical to the interpretation of the assessment. More research into ways of presenting the variability and professional development for teachers and administrators in properly interpreting results from aggregate displays could lead to better inferences from displays such as the ones described above.

9. Unanswered questions

In many ways, the questions which have been answered by the plots shown above are the easiest and most obvious. More research is needed 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 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 evidence-centered design process (see [1,2]), 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 is the model underlying the assessment? This paper has mostly taken the position that teacher will care more about the inferences made from the model than the model that supports the inferences. As with all generalizations, this will be true for some individuals and false for others. Other authors have experimented with some success with visualization techniques that incorporate some elements of the model to help teachers gain insight of student performance. CourseVis [19], for example, produces 2D and 3D representations of student data collected by a web-based course management system. These graphical representations are used to help teachers keep track of students' social, behavioral, and cognitive aspects in a distance learning environment.

Teachers and students can use ViSMod [20] to interact with Bayesian student models. ViSMod offers several visualization techniques (e.g., color and size of nodes, link type, and animation) to represent students' cognitive and social aspects. In addition, guidance mechanisms (i.e., virtual guiding agents) are used to facilitate navigation and interaction with Bayesian student models.

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. Falmagne et al. [21] suggest that students should next be taught skills that fall in their *outer fringe* of ability—that is skills which are within capability, but just barely. Another approach is to embed the Bayesian network in a Markov decision process which would be responsible for instructional planning [22,23].

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.

Acknowledgement

ACED development and data collection was sponsored by National Science Foundation Grant No. 0313202.

References

[1] V.J. Shute, E.A. Graf, E.G. Hansen, Designing adaptive, diagnostic math assessments for individuals with and without visual disabilities, in: L.M. Pytlikzillig, R.H. Bruning, M. Bodvarsson (Eds.), Technology-based Education; Bringing Researchers and Practitioners Together, Information Age Publishing, Greenwich, CT, 2005, pp. 169–202.
[2] V.J. Shute, E.G. Hansen, R.G. Almond, An Assessment for Learning System called ACED: the Impact of Feedback and Adaptivity on Learning, Research Report RR-07-26, ETS, 2007. <http://www.ets.org/research/researcher/RR-07-26.html>.
[3] D. Madigan, R.G. Almond, Test selection strategies for belief networks, in: D. Fisher, H. Lenz (Eds.), Learning from Data: AI and Statistics V, Springer-Verlag, 1995, pp. 89–98.
[4] E.A. Graf, Designing a Proficiency Model and Associated Item Models for a Mathematics Unit on Sequences, Paper Presented at the Cross Division Math Forum (September 2003).
[5] L.A. Hemat, R.G. Almond, IRT versus Bayes Nets: Proficiency Estimates and Parameter Recovery, Research Report, Educational Testing Serivce, submitted for publication.
[6] L.R. Bettesworth, Data Analysis and Interpretation used to Inform Decisions in School, Paper Presented at the Annual Meeting of the American Educational Research Association (AERA), Chicago, IL, 2007.
[7] P.G. Solomon, R. Searson, D. Fried, M. Gajria, Data Analysis and Interpretation used to Inform Decisions in School, Paper Presented at the Annual Meeting of the American Educational Research Association (AERA), Chicago, IL, 2007.
[8] F. Mosteller, C. Youtz, Quantifying probabilistic experience (with discussion), Statistical Science 5 (1) (1990) 2–34.
[9] W. Cleveland, R. McGill, Graphical perception: the visual decoding of quantitative information on graphical displays of data, Journal of the Royal Statistical Society, Series A 150 (1987) 192–229.
[10] R. Ihaka, Colour for presentation graphs, in: K. Hornik, F. Leisch, A. Zeileis (Eds.), Proceedings of the 3rd International Workshop on Distributed Statistical Computing, 2003. <http://www.ci.tuwien.ac.at/Conferences/DSC-2003/Drafts/Ihaka.pdf>.
[11] A. Nicholson, N. Jitnah, Using mutual information to determine relevance in Bayesian networks, in: Pacific Rim International Conference on Artificial Intelligence, 1998, pp. 399–410. <http://citeseer.ist.psu.edu/nicholson98using.html>.
[12] G. Box, G. Tiao, Bayesian Inference in Statistical Analysis, John Wiley and Sons, 1973 (reprinted in Wiley Classics Library Edition, 1992).
[13] A. Gelman, J. Carlin, H. Stern, D. Rubin, Bayesian Data Analysis, second ed., Chapman and Hall, 1995/2003.
[14] W. Yen, A Bayesian/IRT Index of Objective Performance, Paper Presented at the Annual Meeting of the Psychometic Society, June 1987.
[15] H. Wainer, J.L. Veva, F. Camacho, B.B. Reeve III, K. Rosa, L. Nelson, K.A. Swygert, D. Thissen, Augmented scores – "borrowing strength" to compute scores based on a small number of items, in: D. Thissen, H. Wainer (Eds.), Test Scoring, Lawrence Earlbaum Associates, 2001, pp. 343–388.
[16] S.J. Haberman, When can Subscores have Value? Research Report RR-05-08, ETS, 2005. <http://www.ets.org/research/researcher/RR-05-08.html>.
[17] S. Messick, Validity, in: R. Linn (Ed.), Educational Measurement, third ed., American Council on Education/Macmillan, 1989, pp. 13–103.
[18] AERA, APA, NCME (Eds.), Standards for Educational and Psychological Testing, Joint Publication of American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999. <http://www.apa.org/science/standards.html>.
[19] R. Mazza, V. Dimitrova, Coursevis: exernalizing student information to facilitate instructors in distance learning, in: U. Hoppe, F. Verdejo, J. Kay (Eds.), Proceedings of AIED2003, 2003, pp. 279–286.
[20] J.D. Zapata-Rivera, J.E. Greer, Interacting with inspectable bayesian student models, International Journal of Artificial Intelligence in Education 14 (2) (2004) 127–163.
[21] J.-C. Falmagne, E. Cosyn, C. Doble, N. Thiéry, H. Uzun, Assessing Mathematical Knowledge in a Learning Space: Validity and/or Reliability, Paper Presented at the Annual Meeting of the American Educational Research Association (AERA), 2007.
[22] R.G. Almond, Cognitive modeling to represent growth (learning) using markov decision processes, Technology, Instruction, Cognition and Learning (TICL) 5 (2007) 313–324. http://www.oldcitypublishing.com/TICL/TICL.html.
[23] A. Dekhtyar, J. Goldsmith, B. Goldstein, K.K. Mathias, C. Isenhour, Planning for success: the interdisciplinary approach to building bayesian models, International Journal of Approximate Reasoning Appears 50 (3) (2009) 416–428.