Toward a Model of Learning Data Representations

Ryan Shaun Baker (rsbaker@cmu.edu)
Human-Computer Interaction Institute, Carnegie Mellon University
Pittsburgh, PA 15213 USA

Albert T. Corbett (corbett+@cmu.edu)
Human-Computer Interaction Institute, Carnegie Mellon University
Pittsburgh, PA 15213 USA

Kenneth R. Koedinger (koedinger@cmu.edu)
Human-Computer Interaction Institute, Carnegie Mellon University
Pittsburgh, PA 15213 USA

Abstract

The use of graphs to represent and reason about data is of growing importance in pre-high school mathematics curricula. This study examines middle school students' skills in reasoning about three graphical representations: histograms, scatterplots and stem-and-leaf plots. Students were asked to interpret graphs, select an appropriate graph type to represent a relationship and to generate graphs. Accuracy levels varied substantially across the three tasks and three graph types. The overall pattern of results is largely explained by the varying ease of transfer of student knowledge from a simpler graph type, based on surface similarity.

Introduction

External graphical representations are of considerable importance in problem solving. Considerable research has taken place over the last two decades on the different mechanisms through which graphical representations assist their users in drawing inferences (Larkin & Simon, 1987; Stenning, Cox, and Oberlander 1989).

In this paper we take up the use of representations at a very early point – at the point when a student is just learning to generate and interpret a representation – and ask what some of the major challenges are in learning these skills. There has been growing interest in attempting to teach these skills to students as young as those in the third through eighth grades[1] (NCTM 2000), but there is considerable evidence as well that these skills have not yet been developed by many undergraduates (Tabachneck, Leonardo, and Simon 1994).

We take up this subject in the context of developing a cognitive model of how novices generate and interpret some of the simpler representations used in data analysis. This model is designed with production-rule logic, in ACT-R (Anderson 1993). In this process, we hope to follow in the footsteps of some of the successful cognitive models of novices developed in other domains such as algebra problem solving (Koedinger & MacLaren 1997).

One area which might considerably influence students' performance on these tasks is transfer of the knowledge students already have of generating and interpreting other representations. Since students are taught different sets of representations at different grade levels (NCTM 2000), it is quite plausible that an important model for learning new representations will be the representations encountered earlier. Previous research into when transfer occurs shows that transfer can happen between exercises taking place in different representations, through mechanisms such as analogy, and that transfer can occur between similar processes (Novick 1988, Novick and Holyoak 1991, Singley and Anderson 1989). Hence, we seek to find out if and how these processes extend to the very first stages of learning how to use and generate a representation.

We are interested both in positive transfer, and in overgeneralization, where knowledge is transferred inappropriately. Scanlon's (1993) research in the use of representations for physics problem-solving provides some excellent examples of overgeneralization in the interpretation of different graphical representations. Additionally, other research has shown that misconceptions in physics, arising from overgeneralization of previously learned knowledge, causes long-term difficulties in correctly learning new material. How best to deal with such misconceptions is an active question in the research literature, with some arguing for a curricular strategy which acknowledges the appropriate contexts for certain conceptions and helps students see when they are inappropriate (NRC 1999).

In this paper, we present results and analysis of a empirical study we conducted in this domain, investigating novice performance (with an eye towards transfer effects) on interpreting, generating, and selecting representations important to early data analysis.

Domain

Representations

This study focuses on three graphical representations of data: histograms, scatterplots, and stem-and-leaf plots.

A *histogram* depicts a frequency distribution, as displayed in Figure 1. A set of interval categories (as in Figure 1) are represented in the X axis, and the frequency of each category is represented by the height of the corresponding vertical bar. A *stem-and-leaf plot,* shown in Figure 2, also

[1] Between the ages of 7 and 13.

displays frequency distribution data – the frequency of occurrence in this case for values between 0 and 99. In Figure 2, a distribution of 30 values, ranging from 4 to 97, is displayed. The higher order "tens" digit of the values form 10 categories down the left side of the graph. The lower order "ones" digit of each observed value is displayed in an ordered row to the right of the associated tens digit. Finally, a *scatterplot* employs a Cartesian plane to represent the relationship between two quantitative variables, as displayed in Figure 3. Each axis represents one of the variables, and the points represent paired values of the variables.

These three representations were selected because they are featured in most middle school math curricula and to systematically vary graph characteristics. Note that histograms and stem-and-leaf plots each portray univariate frequency distributions, although their surface features are dissimilar. The stem-and-leaf plot looks more like a table, frequency is depicted horizontally rather than vertically, and the frequency count is not directly depicted. In contrast, histograms and scatterplots share some superficial similarities – each has two numerically labeled axes – but they represent very different types of information.

A fourth type of graph, which was not included in the experimental tests, will be relevant in interpreting student performance. This is a *bar graph*, as depicted in Figure 4. A bar graph displays the values of a categorical variable along its X axis, and of a related quantitative variable along its Y axis.

Teacher Predictions

In this study, students are asked to (a) interpret graphical representations, (b) select the appropriate representation for different types of data display, and (c) generate different representations. We asked the two teachers in our sample classes to predict how their students would perform. The teachers predicted that students would perform about equally well on interpretation and generation, and poorly on selection. They predicted that students would be most successful with histograms, next most successful with scatterplots (because scatterplots are more conceptually challenging) and would perform worst on stem-and-leaf plots (because they are the least familiar to students).

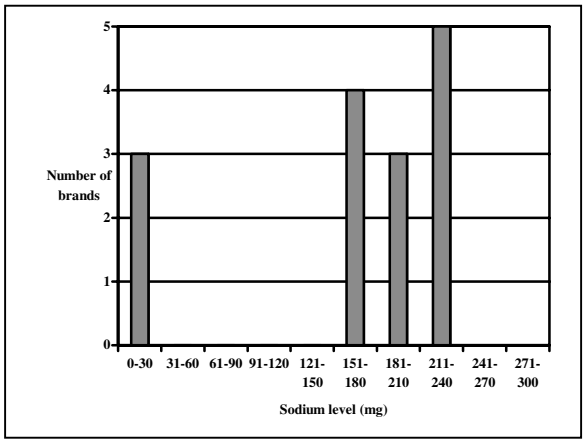

Figure 1: The histogram used in the interpretation exercises

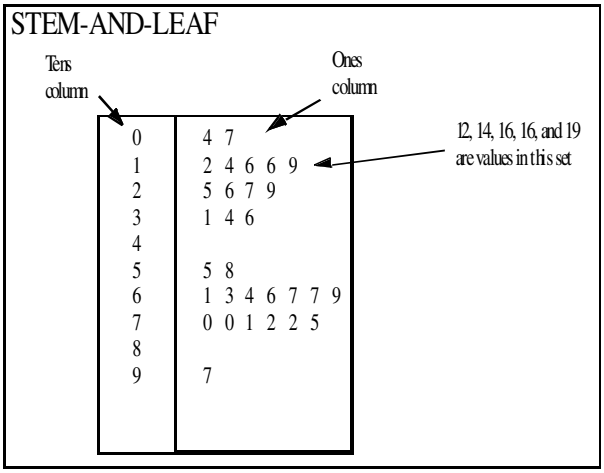

Figure 2: The drawing of a stem-and-leaf plot we used in our refresher sheet

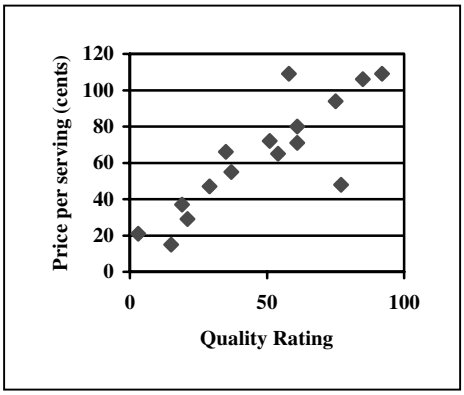

Figure 3: The scatterplot used in the interpretation exercises

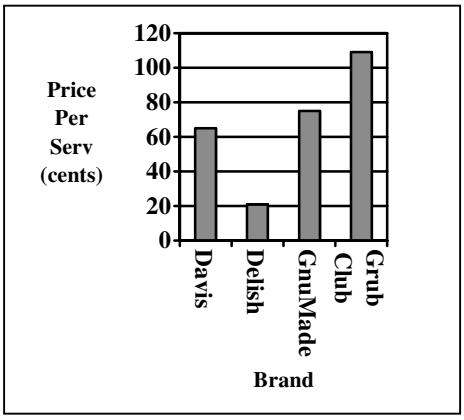

Figure 4: A Bar graph

Methods

Participants

The participants were 52 8th and 9th grade students from three mainstream pre-algebra classes in two Pittsburgh suburbs, half male and half female. The study was conducted prior to the year's data analysis unit; students had some exposure to histograms, scatterplots and stem-and-leaf plots in the last two years, and considerably greater exposure to bar graphs before that point.

Design

In the study, each of the participants completed 3-4 exercises in which they were asked to generate ("draw") a histogram, scatterplot, or stem-and-leaf plot, to answer a set of interpretation questions for one of those representations, or to select the most appropriate representation for a particular question. Four different forms were used, with questions on each form chosen such that neither the same category of task nor the same representation were assigned twice in one form. Within each form, the order of the exercises was rotated for different students to prevent order effects. These forms also included exercises involving box plots and tables, but we neither expected nor found the kind of effects we found for the representations discussed here.

The generation exercises gave the student a data set, in the form of a table, and asked them to draw the given representation of that data. The exercise statement read as follows:

> Please draw a [scatterplot, histogram, stem-and-leaf plot], showing all of the data in this table. Show all work. Feel free to use graph paper, if necessary.

The interpretation exercises gave the student a drawing of that representation and a set of questions to answer using that representation (see Figure 1and Figure 3).

The interpretation exercises had three types of questions, both multiple-choice and open-ended. The first type were straightforward questions typically asked for the target representation. These required no understanding of the representation's global properties (for the histogram, "How many brands of creamy peanut butter have between 0 and 30 mg of sodium?"). The second were also typical, but required understanding of the representation's global properties -- properties which require the student to make inferences (Stenning, Cox and Oberlander 1995, Leinhardt, Zaslovsky, and Stein 1990). "Is there a relationship between quality and price? Answer yes or no." is one such question for scatterplots. Finally, the third type were questions that are not typically asked for the target representation, but could be answered through productions more appropriate for another representation (for the scatterplot, "What is the price of the brand with a quality rating of 3?").

The representation selection questions gave the students a data set, in the form of a table, a question to answer (such as "What type of graph would be best for determining whether or not there is a relationship between price and quality?"), and four choices.

The students were also given a sheet with a picture of the four types of representations directly addressed in the exercises (histograms, scatterplots, stem-and-leaf plots, and box plots – bar graphs were not included in this sheet, nor mentioned in the study). An example from this sheet is shown in Figure 2. We did this in order to prevent the forgetting of terms from having an effect on the students' performance.

Scoring

For generation exercises, we developed rubrics for completely correct solutions (no features incorrect), and solutions that had the correct surface features (with the same general appearance as a correct solution – axes and bars or dots). For interpretation and representation choice exercises, we scored answers either completely right or wrong.

Results and Discussion

Performance accuracy in the graph interpretation, generation and interpretation tasks is displayed in Table 1. Students performed moderately well overall on graph interpretation, averaging 56% correct. However, there was large difference between performance on different representations – the 15 students interpreting histograms performed considerably better (average of 96% correct) than the 12 students interpreting scatterplots (average of 56% correct on the open-ended questions) ($t(25)=4.925, p<0.0001$). Both groups performed considerably better than the 13 students interpreting stem-and-leaf plots (average of 17% correct) (for scatterplots versus stem-and-leaf plots, $t(23)=4.109$, $p<0.001$; for histograms versus stem-and-leaf plots, $t(26)=12.191$, $p<0.0001$). In contrast, student performance on graph selection and graph generation was quite poor. Students were not better than chance accuracy (1 out of 4, 25%) in graph selection. Furthermore, they were completely unsuccessful at generating histograms and scatterplots. Performance by the 15 students who attempted to generate stem-and-leaf plots was relatively poor at 20% completely correct, but was marginally significantly better than the performance of the 12 students attempting to generate histograms (0% completely correct) and the 12 students attempting to generate scatterplots (0% completely correct), using a test of the significance of independent proportions. ($z=1.64$, $p<.11$).

The teachers accurately predicted that students would struggle with graph selection problems. Their predictions that histograms would be easiest and stem-and-leaf plots hardest corresponded with the data, but only for graph interpretation. Their expectation that students would have comparable success with generation and interpretation proved dramatically incorrect. Nathan and Koedinger (2000) report a similar result, that experienced teachers sometimes exhibit an "expert blindspot" and, in some cases,

Table 1: Average student performance in graph generation and interpretation. Percent which students would be expected to get right through guessing is placed in parentheses where appropriate.

	Histogram	Scatterplot	Stem-and -leaf plot
Generation	0%	0%	20%
Interpretation (open-ended questions)	95%	56%	17%
Selection	15% (25%)	20% (25%)	8% (25%)

consistently make inaccurate predictions about which problems will be most challenging for students.

This unanticipated decoupling between performance on graph interpretation and graph generation is striking. In addition to the large difference in overall accuracy, the relative difficulty of different representations is essentially reversed in the two tasks. Students had the most trouble in interpreting stem-and-leaf plots, but that was the only graph type they had any success in generating. A similar dissociation of interpretation and generation has been observed in other domains such as programming (Anderson, Conrad & Corbett, 1989), although not always (Pennington & Nicolich, 1991).

Histogram & Scatterplot = Bar Graph

When we examined the student graph generation solutions which were correct at least so far as having the proper surface features, we noted a characteristic error in 100% of the histograms and 28% of the scatterplots that provides strong evidence about the students' problem solving strategy. Figure 5 displays a typical histogram solution and Figure 6 displays a typical scatterplot solution. In each case the students have constructed axes that are appropriate for a bar graph. In both graphs, the x-axis represents individual values of a categorical variable (individual brands of peanut butter) and the y-axis represents values of a quantitative variable (sodium level). Each of these graphs is the informational equivalent of a bar graph. This suggests that, in this stage of learning, students are transferring existing knowledge about bar graphs, instead of using knowledge specific to the target representation. That students would already have knowledge of bar graphs is consistent with bar graphs being a simpler representation than histograms (no aggregation of data in the x-axis) and scatterplots (only one continuous variable). This hypothesis not only explains the graph generation results, but also appears to account for the overall pattern of graph interpretation and graph selection results as we discuss below.

Graph Interpretation

Figure 7 displays a set of production rules for common bar graph interpretation problems – given one of the categorical

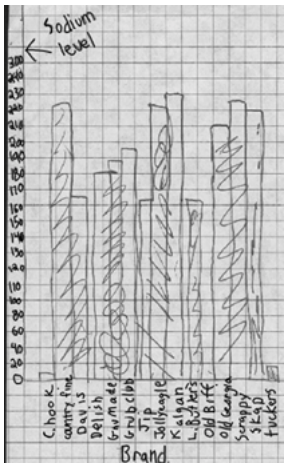

Figure 5: An example of a student-drawn "histogram". Note the axes are a categorical variable versus a quantitative variable – appropriate for a bar graph.

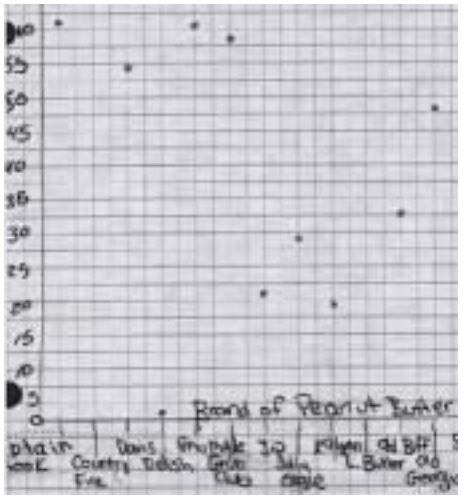

Figure 6: An example of a student-drawn "scatterplot". Note the axes are once again more appropriate for a Bar graph.

instances on the x-axis, find the associated quantitative value on the y-axis. For example, if a bar graph displays the price of several brands of peanut butter (see Figure 4) and a student is asked "What is the price of GnuMade peanut butter?", the student finds "GnuMade" on the x-axis, finds the top of the associated bar, then reads horizontally across to find GnuMade's price on the y-axis. These productions were directly applicable to all of our histogram interpretation exercises. On those exercises, the 15 students had an average accuracy rate of 95%. These productions were also applicable to several of the scatterplot questions (e.g., "What is the price of the brand with a quality rating of 3?"). Students had an average accuracy rate of 50% on these questions.

Student performance on the scatterplot questions is quite striking, when we consider the fact that these questions are not standard for scatterplots. At the same time, the students

Table 2: Average student performance in graph interpretation, for different kinds of problems. Percent which students would be expected to get right through guessing is placed in parentheses where appropriate.

	Histogram	Scatterplot	Stem-and -leaf plot
Interpret: "Bar graph Productions"/ Analogous	95%	50%	21%
Interpretation: Emergent Properties	N/A	67% (50%)	N/A
Interpretation: Overall	95%	57%	17%

had considerable difficulty with other scatterplot exercises. In fact, on the exercises where students had to interpret a scatterplot's global properties, generally considered that representation's main function, the students did not perform significantly better than chance ($p>.10$, $N=12$, using a sign test).

By contrast, the bar graph interpretation productions in Figure 7 do not transfer to stem-and-leaf plots, because the student can neither look up the given value on either axis, nor read the answer off the other axis. This is borne out by the fact that on the 4 questions which were almost word-for-word identical to the histogram interpretation questions, the 13 students performed much more poorly, averaging 21% correct. This is significantly lower than the performance on the corresponding questions for histograms ($t(26)=9.908, p<.0001$).

Hence student interpretation performance levels are more similar between histograms and scatterplots that share surface features with bar graphs, than between histograms and stem-and-leaf plots that have similar content but different surface features (cf. Chi, Feltovich, & Glaser 1981).

Graph Selection

If students are reasoning about all three representations with reference to a single familiar bar graph, they have no basis for discriminating which of the three is appropriate for representing different types of relationships and we would expect chance performance levels. This is in fact what we found, with students getting 15% on the graph selection exercises, even below the 25% accuracy choosing 1 out of 4 would predict. In the absence of understanding which representation is appropriate, an apparent bias in favor of the representation the student had drawn in another question, which by our study design was necessarily wrong, may have led to the observed below chance performance.

```
If we are trying to find a value on a graph
  And we are looking for the value
        corresponding to a value V
Then
  Set a subgoal of looking for V written on the x axis

If we are looking for V on the x axis
  And value V is written at location x* on the x axis
Then
  Set a subgoal of looking at location x*

If we are looking at location x* on the x axis
  And point P is the topmost drawn in graphic
        above location x* on the x axis
Then
  Set a subgoal of looking across from P

If we are looking across from P
  And y* is the value on the y axis horizontally
        over from P
Then
  Return y* as the value we were looking for
```

Figure 7: English-language productions for bar graph interpretation, also suitable for some histogram and scatterplot interpretation.

```
If we are trying to generate a graph G
And G's axes have not been selected
Then
  Set a subgoal for selecting G's axes

If we are attempting to select axes for graph G
  And the X axis is not selected
  And there is a variable V in our data set
  And V is a categorical variable
Then
  Select V as our X axis
```

Figure 8: An English-language production for choosing the X axis variable during generation, overgeneralized to apply inappropriately during scatterplot and histogram generation.

Conclusion and Future Work

Novice students' performance on interpretation, generation, and selection of the data representations in this study can be explained as depending upon transfer of their prior knowledge of bar graphs. Where transfer and generalization are afforded by surface similarity, they occur, whether appropriate or not. This hypothesis exposes a more integrated pattern of interpretation and generation performance than is apparent in the overall results.

Given these findings, we are working toward developing a more complete ACT-R cognitive model of learning data representations

A major subtask in developing our model will be refining our understanding of the factors which scaffold the novices in transferring their knowledge, both appropriately and inappropriately. We do not yet know which common features between representations are essential to this process – certainly it seems that surface features are more important than deeper features, a finding compatible with those in analogical transfer (cf., Novick 1988; Novick and Holyoak 1991) – but which surface features are most salient is an important question in itself. For example, stem-and-leaf plots have three large differences from histograms: flipped axes, the need to remove the tens digit, and the need to count up values. Determining which of them is most

important will have large impacts on our understanding of the generality of the productions students use.

Eventually, we hope to use the cognitive model we develop to build a Cognitive Tutor (Corbett, Koedinger, & Hadley, in press) for this domain. Already, our research has given us extensive information about some of the important difficulties novices have in learning how to generate and interpret these basic representations, including the confusion between histograms and bar graphs. Additionally, our curriculum will be strongly shaped by further research determining whether these overgeneralizations are truly misconceptions which need to be broken down, or whether they are preconceptions which can still be built upon in some way (cf., NRC 1999).

A final future direction is one that may have surprising power – rather than trying to repair misconceptions, we may get even better long-term results from addressing the possibility that we can create a curriculum where bar graphs are taught differently, and the overgeneralization never develops in the first place – where the interpretation productions still transfer, but the generation productions do not become inappropriately broad. Through research in these areas, we hope to transform students' knowledge in this domain.

Acknowledgments

We would like to thank Jay Raspat and Katy Getman for assisting us in the administration of the study discussed here. We would also like to thank Bethany Rittle-Johnson and Jack Zaientz for helpful discussions.

References

Anderson, J.R. (1993) *Rules of the Mind*. Hillsdale, NJ: Erlbaum.

Novick, L.R. (1988) Analogical Transfer, Problem Similarity, and Expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition.* 14 (3), 510-520.

Scanlon, E. (1998) How Beginning Students Use Graphs of Motion. In M.W. van Someren, P. Reimann, H.P.A. Boshuizen, T. de Jong (Eds.) *Learning With Multiple Representations.* Kidlington, OX UK: Elsevier Science.

Anderson, J.R., Conrad, F. and Corbett, A.T. (1989) Skill acquisition and the LISP Tutor. *Cognitive Science*, 13, 467-505.

Chi, M. T. H., Feltovich, P., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5,* 121-152.

Corbett, A. T., Koedinger, K. R., & Hadley, W. H. (in press). Cognitive Tutors: From the research classroom to all classrooms. In Goodman, P. S. (Ed.) *Technology Enhanced Learning: Opportunities for Change*. Mahwah, NJ: Lawrence Erlbaum Associates.

Koedinger, K.R. and MacLaren, B.A. (1997) Implicit strategies and errors in an improved model of early algebra problem solving. In Shafto, M.G. & Langley, P. (Eds.) *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society.*382-387.

Larkin, J.H. and Simon, H.A. (1987) Why a Diagram is (Sometimes) Worth Ten Thousand Words. *Cognitive Science,*11, 65-99.

Leinhardt, G, Zaslavsky, O. and Stein, M.K. (1990) Functions, Graphs, and Graphing: Tasks, Learning, and Teaching. *Review of Educational Research.* 60 (1), 1-64.

Nathan, M.J. & Koedinger, K.R. (2000) An investigation of teachers' beliefs of students' algebra development. *Cognition and Instruction.* 18 (2), 207-235.

National Council of Teachers of Mathematics. (2000) *Principles and Standards for School Mathematics.* Reston, VA: National Council of Teachers of Mathematics.

National Research Council. (1999) *How People Learn: Brain, Mind, Experience, and School.* Washington, DC: National Academy Press.

Novick, L.R. and Holyoak, K.J. (1991) Mathematical Problem Solving by Analogy. *Journal of Experimental Psychology: Learning, Memory, and Cognition.* 17 (3), 398-415.

Pennington, N. and Nicolich, R. (1991) Transfer of Training Between Programming Subtasks: Is Knowledge Really Use Specific? *Empirical Studies of Programmers: Fourth Workshop.* 156-176.

Singley, M.K. and Anderson, J.R. (1989) *The Transfer of Cognitive Skill.* Cambridge, MA: Harvard University Press.

Stenning, K., Cox, R., and Oberlander, J. (1995) Contrasting the cognitive effects of graphical and sentential logic teaching: Reasoning, representation, and individual differences. *Language and Cognitive Processes*, 10, 333-354.

Tabachneck, H.J.M., Leonardo, A.M., Simon, H.A. (1994) How Does an Expert Use a Graph? a Model of Visual and Verbal Inferencing in Economics. *Proceedings of the 16th Annual Conference of the Cognitive Science Society.*