Distinct Errors Arising From a Single Misconception

Ryan S. Baker (rsbaker@cmu.edu)
Albert T. Corbett (corbett@cmu.edu)
Kenneth R. Koedinger (koedinger@cmu.edu)
Human-Computer Interaction Institute, Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213 USA

Data and Introduction

We present an account of the mistakes students have been observed to make when generating scatterplots, and give evidence that somewhat disparate error behaviors can be traced to the same strategic decision.

In two prior studies (Baker, Corbett, & Koedinger 2001, 2002), we observed middle-school students attempting to generate scatterplots (which have a quantitative variable on each axis) but making two conceptually similar errors. When given both categorical and quantitative variables but no advice on which to place in their graph, 15% made what we call the *choice* error, incorrectly choosing a categorical variable for the X -- 0% used the correct variables. Naming the variables to use in the question did not eliminate this error, but 77% used the correct variables. 13% of those students, however, then made what we term the *representation* error: treating the values of the quantitative X variable as if they were categorical. They wrote the variable's values along the axis in the order they appeared in the data table, rather than numerical order, e.g., placing "22 20 23 25 24 19 23" along the axis rather than "19 20 21 22 23 24 25". Labeling the axis variables for the student did not significantly reduce the representation error's frequency.

Our conjecture was that students made both errors due to a substantially greater familiarity with bar graphs, leading to the belief that graphs always have a categorical x-axis. A possible alternate conjecture for the representation error is that students don't understand the difference between categorical and quantitative variables. However, this conflicts with their comparatively greater ability to represent the Y axis quantitatively, as is done in a bar graph.

Modeling

In drawing each axis of a scatterplot, there are two key decisions for the student to make -- which variable to graph and how to represent its values. We developed an ACT-R (Lebiere & Anderson, 1998) model of these decisions and fit it to the students' behavior displayed in Table 1. In alternate fits we modeled the two correct decisions (correct variable choice and quantitative value representation) as two different productions (if-then rules) or the same production. Similarly, we modeled the two errors (categorical variable choice and categorical representation of a quantitative variable) as two productions or the same production. Modeling the two correct decisions as different productions produces a significantly better fit than modeling them as a single production ($F(1,30)=39.853$, $p<0.001$). The model where the two errors stem from the same strategic production has equal fit but superior parsimony to the model where they stem from different strategic productions. (BiC (same)=77.00, BiC(different)=79.28) The former model achieves an excellent fit to the overall pattern of data from the two experiments. ($r=0.990$, mean absolute dev =.057)

This model is generally consistent with but clarifies our early conjectures. As shown in the top row of the table, students never pick a quantitative variable for the x-axis unless one is suggested. This implies that in these studies correct selection of a quantitative variable just reflects the ability to follow directions; treating a quantitative variable as quantitative, on the other hand, is modeled as active knowledge of the difference between quantitative and categorical variables. The two errors in this account, selecting a categorical variable and treating a quantitative variable as categorical, reflect a single misconception. These students know the difference between categorical and quantitative variables, but are biased to make the X axis categorical in any way possible, consistent with their prior experience with bar graphs.

Thus, in this domain multiple error behaviors arise from a single misconception, an overgeneralization of their prior knowledge of bar graphs.

Table 1: Percent occurrence of behaviors

	No prompts	No labels	X label	Y label	Both label
Correct	0	53	59	62	61
Choice error	15	27	9	27	8
Rep error , X axis only	0	10	14	12	10
Rep error, Y axis only	0	0	0	0	0
Rep error, both axes	0	3	4	0	6
Other/ Give Up	85	7	14	0	15

References

Anderson, J. R. & Lebiere, C. (1998). The atomic components of thought. Mahwah, NJ: Erlbaum.

Baker R.S., Corbett A.T., & Koedinger K.R. (2001) Toward a Model of Learning Data Representations. Proceedings of the Cognitive Science Society Conference. pp. 45-50

Baker R.S., Corbett A.T., & Koedinger K.R. (2002) The Resilience of Overgeneralization of Knowledge about Data Representations. Presented at American Educational Research Association Conference. New Orleans, LA.