# The Resilience of Overgeneralization of Knowledge about Data Representations

Ryan Shaun Baker, Albert T. Corbett, Kenneth R. Koedinger
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, PA 15213 USA
rsbaker@cmu.edu, corbett+@cmu.edu, koedinger@cmu.edu

## Introduction and Prior Work

Data analysis has become a topic of increasing emphasis within middle school mathematics in the last few years, especially in the recent recommendations by the National Council of Teachers of Mathematics (NCTM 2000). In order to better inform efforts to expand data analysis's role in middle school curricula, we have begun the development of a cognitive model of student thinking in this domain. Proper representation of data is an essential part of the process of data analysis (Larkin and Simon 1987) -- therefore, we have focused on modeling how students learn to generate and interpret some of the important and widely-used representations of data, such as histograms and scatterplots.

Students bring a large number of preconceptions and sources of prior knowledge to the process of learning new representations. Researching and modeling these preconceptions is highly important, since with knowledge of these factors, we can craft instructional interventions that build upon prior student knowledge rather than conflicting with it (Bransford, Brown, and Cocking 1999). One way that students use prior knowledge is by using the first and simplest representations of data they learn, even in situations where those representations are not appropriate, such as using Venn Diagrams instead of scatterplots to answer correlation questions (Hancock, Kaput, and Goldsmith 1992).

Beyond just choosing more familiar representations, though, students have been observed to attempt to transfer in knowledge about more familiar representations when attempting to interpret or generate a newer representation (Baker, Corbett, and Koedinger 2001). Specifically, this study found that middle school students attempt to apply their extensive prior knowledge of how to generate and interpret standard bar graphs when attempting to generate and interpret scatterplots and histograms (here defined as having an X axis broken into intervals of a quantitative variable, and a Y axis with the frequency of each interval). This attempt has both positive and negative effects.

Positively, these generalized bar graph skills enabled students to interpret

some properties of scatterplots -- such as identifying characteristics of specific points -- even when they could not use scatterplots in the ways for which this representation is usually used, such as identifying trends.

More negatively, students could not properly generate either histograms or scatterplots, typically choosing axis variables more appropriate for a standard bar graph than either of these representations. An example of this is shown in Figure 1. In this study, none of the 13 students who attempted to draw scatterplots or the 12 students who attempted to draw histograms succeeded in choosing the correct axis variables, making correct generation impossible.

One question about our results was whether this phenomenon was limited solely to the choice of axis variables, or whether students' mis-transfer was more deeply rooted. Evidence that the second might be the case came from an account of a classroom where students had difficulty drawing interval variables. These students collapsed spaces where there were no observations, drawing "Monday Tuesday Wednesday Thursday Friday Monday" with equal spacing when there were no observations for Saturday or Sunday (Lehrer and Schauble in press). By doing this, they effectively produced an ordered nominal, halfway between what they would have drawn for a standard bar graph and the correct solution.

In this paper, we present a study that we conducted to explore the depth of this transfer and to investigate methods for reducing its occurrence in generation, where it is clearly inappropriate. Given the strength of the student desire to choose axes appropriate to a bar graph when generating scatterplots and histograms, we decided to investigate whether there were ways to induce these students to instead draw the correct axes, and if there was further mis-transfer of bar graph knowledge once the students had the correct axes, as the effect in Lehrer et al suggests. Therefore, we chose the general intervention of drawing the students' attention to the variables, in order to attempt to express each of these factors.

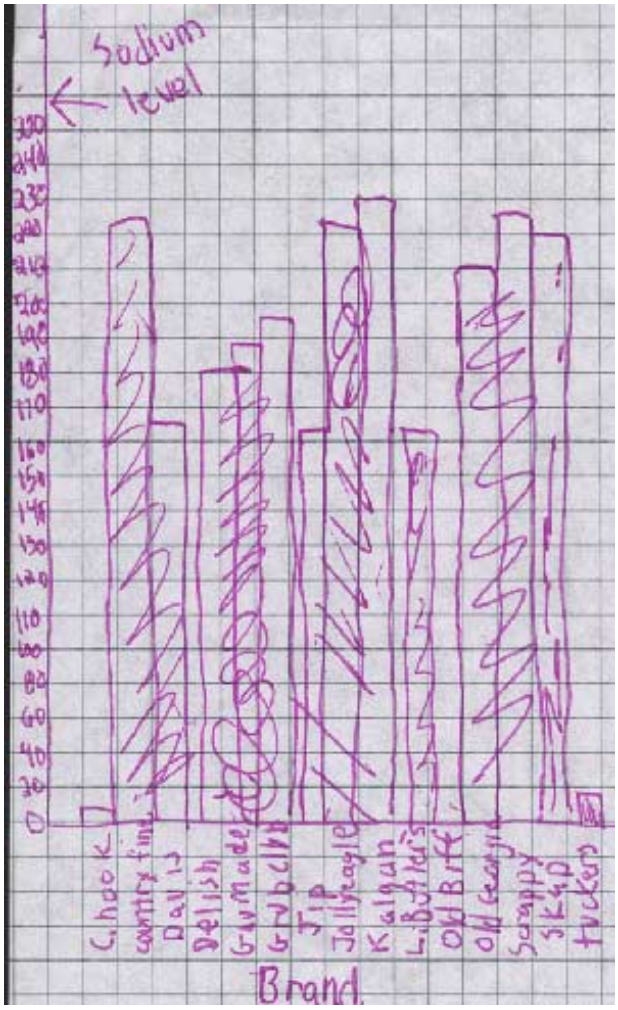

Figure 1: The student was asked to draw a histogram but chose variables more appropriate to a bar graph (a nominal X and continuous Y, instead of bins of a continuous variable on the X and frequency of each bin on the Y).

## Design

We had as participants 119 8th and 9th grade students in 5 mainstream (neither gifted nor special needs) classes in 3 schools, in both inner-city and suburban Pittsburgh. Each student completed one exercise where they generated a histogram, and one exercise where they generated a scatterplot. In order to focus solely on prior knowledge, the part of the study concerning scatterplots was administered before any data analysis was discussed in class, and the part of the study concerning histograms was administered after scatterplot interpretation was discussed but before any other data analysis topics were discussed.

In order to determine whether drawing the students' attention to the variables to use affected performance, we examined two possible scaffolds: the first, a simple and non-intrusive intervention, was to tell them which

variables to use within the question. The second intervention was to directly label the axes for the students. We created four conditions where students were told the variables to use within the question: no labels, X axis labeled, Y axis labeled, both axes labeled. We compared these to the control data from the previous study where the students were not explicitly given the variables to use. In all exercises, we gave the students a data set with all of the necessary variables, plus one distractor nominal variable.

## Results -- Scatterplots

**Conditions**

| | Variables not given (Baker et al 2001) | No labels | X labeled | Y labeled | Both labeled |
|---|---|---|---|---|---|
| Scatterplot: Completely correct | 0% | 53% | 59% | 62% | 61% |
| Scatterplot: Correct axes | 0% | 73% | 85% | 77% | 82% |
| Scatterplot: Nominalization | 0% | 20% | 26% | 15% | 21% |

Table 1: Scatterplot results, showing the frequency of each of the common results for each of the different prompts given.

The manipulation of including the variable names in the directions had a powerful effect on the students' ability to choose the correct axis variables for scatterplots. As shown in Table 1, 0% of the students in the control condition selected the correct axes, whereas 73%-85% of the students in the other four conditions did.

Although students were successful at choosing the correct axis variables, there is evidence that many of the students who correctly identified the variables still mis-transferred knowledge from bar graphs. 28% of the students who chose the correct axis variables (20% of all of the students) nominalized the X axis. Instead of plotting the values of the variable in numerical order with appropriate intervals and no space between identical values, they plotted the individual values of the variable, often in the exact same order as they were found in the table: 22,20,23,25,24,19,23. An example where a student made this error is shown in Figure 2.

This error means that even though the students placed the correct variables on the axes of the graph, the graph they produced had one nominal

variable and one quantitative variable rather than two quantitative variables. The representations these students have drawn are therefore informationally equivalent to a bar graph, despite having the appropriate choice of axis variables.

Unlike the phenomena we observed with scatterplot generation, explicitly cueing students with the correct variables had little effect on their ability to choose the correct axis variables for histograms, as shown in Table 2 (0% in the control condition, 4% in the no label condition, no better than 7% in any of the treatment conditions). This is perhaps most surprising in the both labeled condition, where both axes had already been labeled for them with the correct variables.

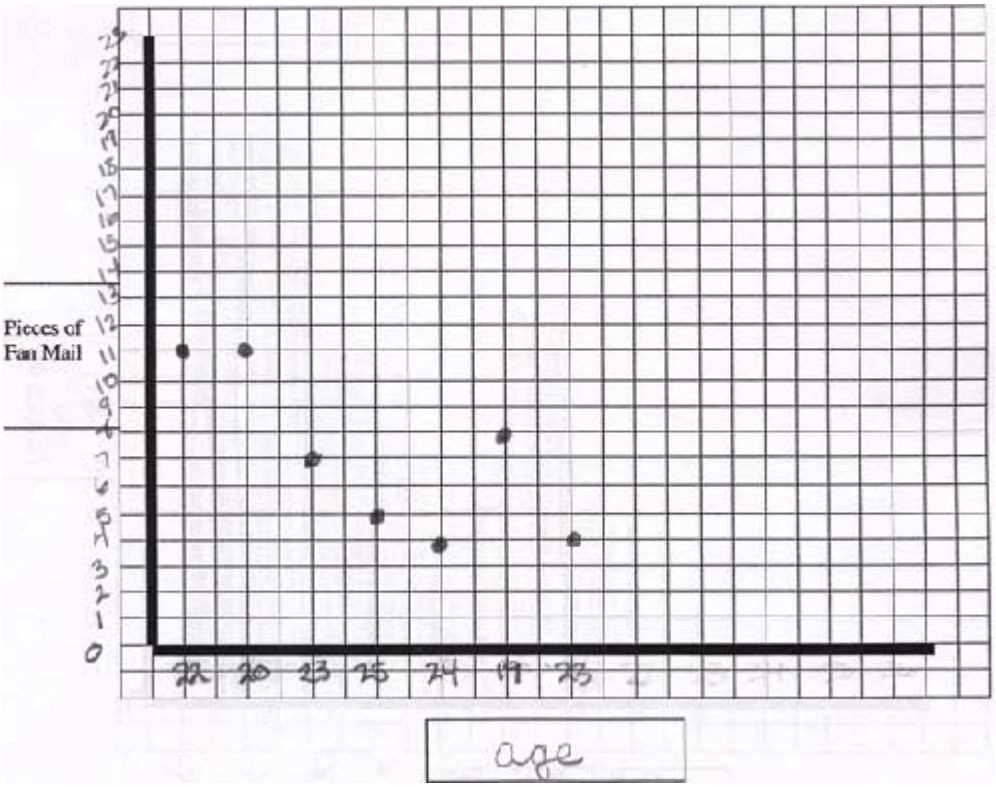


Figure 2: The student has the correct axis variables but has nominalized the X axis, drawing the values of that variable individually and unordered, consistent with over-generalized bar graph knowledge.

## Results -- Histograms

**Conditions**

| | Variables not given (Baker et al 2001) | No labels | X labeled | Y labeled | Both labeled |
|---|---|---|---|---|---|
| Histogram: Completely correct | 0% | 0% | 0% | 0% | 7% |
| Histogram: Correct axes | 0% | 4% | 0% | 0% | 7% |
| Histogram: Traditional bar graph axes | 66% | 86% | 28% | 37% | 15% |
| Histogram: Flipped bar graph axes | 0% | 7% | 59% | 41% | 54% |
| Histogram: Total bar graph axes | 66% | 93% | 87% | 78% | 69% |

Table 2 -- Histogram Results, showing the frequency of each of the common results for each of the different prompts given. .

Looking for the cause of this low performance, we found that 93% of our participants in the no label condition had drawn a nominal variable on the X axis and had drawn on the Y axis as a continuous variable the variable that they should have separated into bin categories on the X axis, giving them a bar graph. Between 69% and 87% of the students in the other three conditions made this or a similar mistake, with the error least common in the both-labeled condition (significantly so -- between both-labeled and no-labeled, $Z=2.35$, $p<0.02$ for a test of the difference of independent proportions).

Interestingly, many of the solutions in the X-labeled, Y-labeled, and both-labeled conditions contained a similar but subtly different error, shown in Figure 3. Although these students' solutions were still informationally equivalent to a bar graph, the graph in their solutions was flipped 90 degrees from a traditional bar graph, with a nominal variable on the Y axis, and with the proper variable on the X axis, but drawn as a continuous variable rather than as bin categories. 41%-59% of the students in the labeled conditions drew this bar graph with flipped axes, as compared to 7% in the no labels

condition -- a significant difference in all three comparisons. (the smallest difference was $Z=3.55$, $p<0.0002$ for a test of the difference of independent proportions)

We hypothesize that this behavior occurred because the labeled axes made it seem that the Y axis was a more appropriate place for the nominal variable. Seeing the X axis already labeled with "Pieces of Fan Mail", "Musician" (the other base variable) would seem a wholly inappropriate substitution. On the other hand, with the Y axis already labeled as "Frequency (Number of Musicians)", the student would observe that it doesn't quite match but does contain the word "Musician" (within "Musicians"), and therefore the student could more reasonably choose "Musician" in the absence of a more solid understanding of histogram axes.

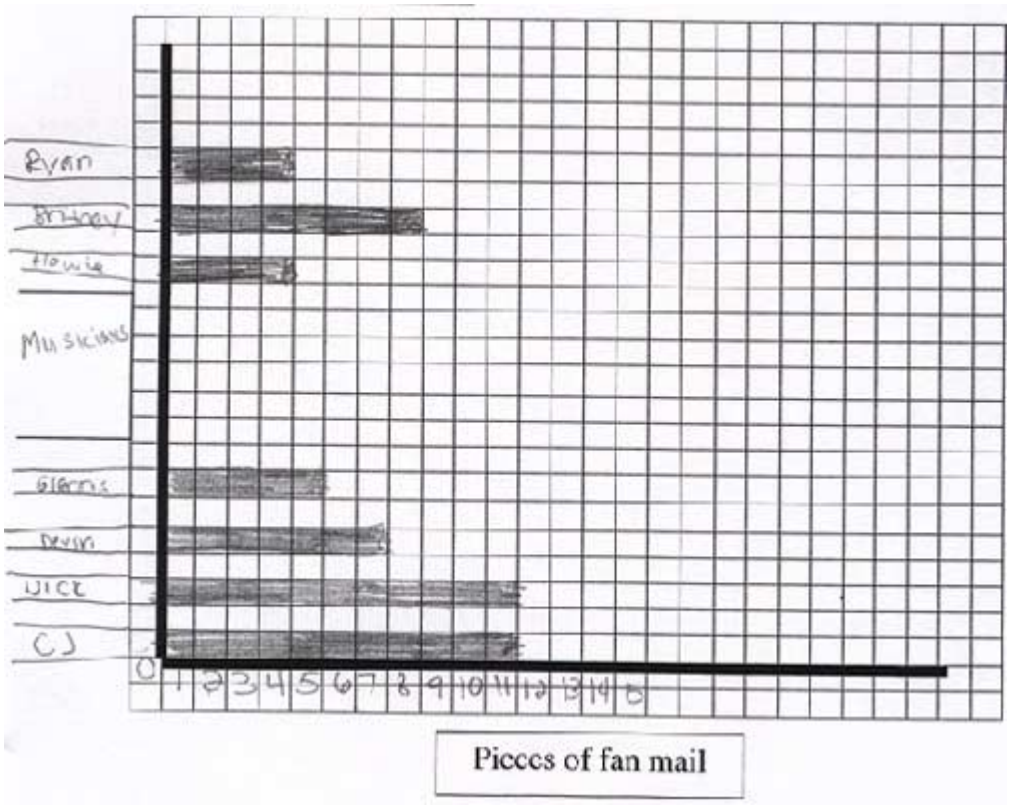


Figure 3: When asked to draw a histogram, and prompted with the x-axis, many students drew bar graphs flipped 90 degrees.

## Discussion

In summation, then, we have found that the previously documented behavior of transferring knowledge about how to generate scatterplots and histograms from prior knowledge of how to generate bar graphs is fairly robust and resilient, at least to the types of simple scaffolds used in this study. Naming the variables to place on the axes in the directions caused more

students to place the correct variables on the axes for scatterplots, but a large number of these students then processed the X axis variable as a nominal variable instead of a quantitative, resulting in a graph which was still informationally equivalent to a bar graph. For histograms, on the other hand, neither naming variables nor actually labeling the axes for the students had a significant effect on how frequently students got the axis variables correct, although being given the axis labels (especially the X axis) caused many students to draw a 90-degree flipped bar graph instead of a standard one.

It seems clear that these misconceptions about the relationship between bar graphs, histograms, and scatterplots are fairly deep-seated, and the tendency of these students is to find ways to use their knowledge of bar graphs when generating these representations, whether they do so directly, via choosing a nominal variable as an axis, or more indirectly, through nominalizing a quantitative variable. We believe that instruction and exercises should be developed that teach students when to use their prior knowledge and when it is inappropriate, in order to promote deep conceptual understanding of the interrelationships and differences of the different representations. To this end, caution should be taken when giving students scaffolds that bypass some of the most challenging aspects of representation generation, such as labeling the axis scale for students in advance. Such scaffolds may serve only to conceal robust misconceptions even from the students' teachers -- with attention to these factors, on the other hand, and curricula which make it salient to teachers when their students have these misconceptions, we believe that the process of data analysis can be made more accessible to all students.

## References

### Works cited

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

Bransford, J.D., Brown, A.L., & Cocking, R.R. (1999). How people learn: Brain, mind, experience and school. Washington, D.C.: National Academy Press.

Hancock, C., Kaput, J.J., & Goldsmith, L.T. (1992) Authentic Inquiry With Data: Critical Barriers to Classroom Implementation. Education Psychologist, 27(3), 337-364.

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

Lehrer, R. and Schauble, L. (2001, Eds.). Investigating real data in the

classroom: Expanding children's understanding of math and science. New York: Teachers College Press.

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