In: K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th Annual Conference of the Cognitive Science Society* (2005). Mahwah NJ: Lawrence Erlbaum Associates. p. 1537.

Performance *vs.* Learning: Knowing the Right Answers for the Right Reasons

Norma M. Chang (nchang@andrew.cmu.edu)
Kenneth R. Koedinger (koedinger@cmu.edu)
Marsha C. Lovett (lovett@cmu.edu)
Department of Psychology, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh PA 15213 USA

Introduction

An important dilemma to resolve in instruction is distinguishing between short-term performance and long-term learning in assessing students' progress. Conditions that appear favorable in acquisition are not always as effective at promoting subsequent retention and transfer, due to differences in the processing activities involved in training and at test; in some cases, poor performance in training produced better performance at test (Schmidt & Bjork, 1992). Since global measures of accuracy and speed are insufficient predictors of the effectiveness of training, we have used a modeling approach to draw inferences about the knowledge structures students use to solve problems both at test and in training. We will present the results from one study demonstrating the usefulness of such qualitative measures in predicting learning outcomes from training.

Method

We collected complete sets of data from 47 statistics-naïve undergraduate students in a five-day training study in which they received instruction and guided practice in solving exploratory data analysis problems. The focus of the instruction was to learn when and how to use pie charts, histograms, boxplots, scatterplots, and contingency tables to analyze a set of data. Following each lesson explaining and demonstrating how to use the new representation and method of data analysis, participants worked through a series of practice problems (30 problems in total). One group solved problems in which the problems' surface features were spuriously correlated with their deep structure (*S*-condition), while the other group solved problems whose surface features were varied across all the problem structures (*V*-condition). All participants received problems that were broken down into their individual steps, as well as correct-answer feedback on their solutions. On the final day, participants solved 25 new problems without any scaffolding or feedback. (See Chang, Koedinger, & Lovett, 2003, for a fuller description of a similar procedure.)

Results and Discussion

Consistent with the claim that good performance in training does not guarantee good performance at test, *V*- participants demonstrated a slight disadvantage at training but superior performance at test, in terms of their accuracies and latencies in selecting the appropriate representation type for analyzing the dataset given in the problem. Examining participants' actual answers revealed that *S*-participants' errors were not merely random, but reflected negative transfer from the surface features that had been incorporated into their training.

To assess the extent to which their answers were driven by surface features or by problem structure, we developed a model of participants' knowledge that specified the different possible features they could be using to choose the appropriate statistical display to answer each question. This model was fit to participants' data by adjusting the parameters indicating the degree to which different features were used. The best-fitting models indicated that at test, *S*-condition participants tended to derive their answers from surface features rather than deep structure, whereas *V*-condition participants made greater use of deep structure than surface features.

Analyzing the training data using the same modeling methodology showed that even during the learning phase, *V*-participants demonstrated stronger knowledge of deep structure, whereas *S*-participants exhibited a stronger influence from surface features. The contrast between the apparent performance of *S*- and *V*-participants according to the two different methods of assessment underscores the importance of measuring the target skills that students are intended to learn. Examining the accuracy data alone would suggest that the *V*-participants were performing more poorly than the *S*-participants, with average scores about half a standard deviation lower. However, examining the reasons why participants chose the answers they did reveals more sophisticated understanding in the *V*-condition. Revising our assessments of students' learning to reflect their knowledge representation, rather than relying merely on accuracy scores, may better inform instructional design by distinguishing more clearly between learning and performance.

Acknowledgments

This research was supported by a Department of Education Jacob K. Javits fellowship and NSF grant no. 0087632.

References

Chang, N.M., Koedinger, K.R., & Lovett, M.C. (2003). Learning spurious correlations instead of deeper relations. In R. Alterman & D. Kirsh (Eds.), *Proceedings of the 25th Cognitive Science Society*. Boston, MA: Cognitive Science Society.

Schmidt, R.A., & Bjork, R.A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, 3, 207-217.