Learning of Categories Composed of Rules and Exceptions

Michael A. Erickson and **John K. Kruschke**
Department of Psychology and Cognitive Science Program
Indiana University, Bloomington IN 47405
miericks@indiana.edu

Many formal and folk psychological theories conceive of the mind as being composed of quasi-independent modules. From Freud to Fodor the mind has been decomposed into constituent parts. Recently, a number of researchers have proposed modular theories of cognitive phenomena such as categorization (Ashby, Alfonso-Reese, & Turken, 1995; Shanks & St. John, 1994), reasoning (Sloman, 1996), automaticity (Logan, 1988), language (Pinker, 1991), and learning and memory (Squire, 1992). In general, these theories are characterized by descriptions of each module and how each serves in those tasks for which it is best suited. However, these theories do not emphasize how modules *interact* in producing responses and in learning.

We describe two human categorization experiments designed to address the three issues relevant to hybrid rule- and exemplar-based systems: the necessity of rules, the necessity of exemplar memory, and the interaction between these two sub-systems in learning and in classification performance. We account for the participants' classifications using an updated version of the hybrid rule and exemplar model described by Kruschke and Erickson (1994). This hybrid model consists of a rule module, an exemplar module, and a gating mechanism. This gating mechanism controls the influence of each module in decision-making and the extent of learning in each module. We also show that neither of these two sub-systems, acting alone, can adequately account for human behavior. This is significant inasmuch as the exemplar sub-system is a full implementation of ALCOVE (Kruschke, 1992), which has performed well in a variety of categorization tasks.

Human Learning

Three key features of the category structures used in these experiments are: (1) some stimuli could be classified according to a rule whereas other stimuli were exceptions and had to be memorized; (2) different training instances had different relative frequencies; and (3) some stimuli were never used in training and were therefore available to examine generalization.

The stimuli in both experiments varied along two dimensions. In each training trial, a stimulus was presented and participants made a classification, after which the correct label was displayed. During the initial trials, participants' responses were just guesses, but after many trials they began to learn the correct classifications. Most of the training stimuli, the *regular* stimuli, could be classified according to a simple, one-dimensional rule. Two training stimuli were exceptions to the rule, each having its own category label.

During training in Experiment 1, participants never saw the most extreme values of the two dimensions of variation; these were reserved to test generalization. Nevertheless, even when these extreme stimuli were most similar to an exception, they were classified according to the rule. Whereas the hybrid model was able to account for this phenomenon, ALCOVE was not.

Relative instance frequencies were manipulated in Experiment 2, both for rules and exceptions. Higher rule instance frequencies caused more robust generalization. Rule-based explanations that lack exemplar memory cannot account for such behavior. Moreover, ALCOVE failed to account for the S-shape of the exception learning curve shown by participants. The hybrid model did show this same pattern of learning by virtue of its interactive gating of the rule and exemplar modules.

References

Ashby, F. G., Alfonso-Reese, L., & Turken, A. U. (1995). Competition between verbal and implicit rules of category learning. Talk presented at the 36th Annual Meeting of the Psychonomic Society, Los Angeles, CA, 10 November 1995.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*, 22–44.

Kruschke, J. K., & Erickson, M. A. (1994). Learning of rules that have high-frequency exceptions: New empirical data and a hybrid connectionist model. In *The Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*, pp. 514–519 Hillsdale, NJ. Erlbaum.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, *95*, 492–527.

Pinker, S. (1991). Rules of language. *Science*, *253*, 530–535.

Shanks, D. R., & St. John, M. F. (1994). Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences*, *17*, 367–447.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*, 3–22.

Squire, L. R. (1992). Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review*, *99*, 195–231.