Comparing the Comparisons

Panel Chair
Walter Warwick
MA&D Operation, Alion Science and Technology
4949 Pearl East Circle, Suite #300
Boulder, CO 80301
303-442-6947
wwarwick@alionscience.com

Model comparison is becoming an increasingly common method in computational cognitive modeling. The methodology is seemingly straightforward: model comparisons invite the independent development of distinct computational approaches to simulate human performance on a well-defined task. Typically, the benchmarks of the comparison are goodness-of-fit measures to human data that are calculated for the various models. Although the quantitative measures might suggest that model comparisons produce "winners," the real focus of model comparison is, or at least should be, on understanding in some detail how the different modeling "architectures" have been applied to the common task. And in this respect, the seemingly straightforward method of model comparison becomes more complicated.

The idea that a model comparison might be used to pick a winning approach resonates with common intuitions about model validation, namely, that a good fit is good evidence for the theory the model implements. But to the extent that model comparisons seek to illuminate general features of computational approaches to cognition rather than to validate a single theory of cognition, they depart from the familiar mode of good fit, good theory. Instead, a model comparison forces us to think about the science of modeling. A good fit is thus relegated to a minimum requirement for participation in a model comparison, rather than an end in itself, and the focus shifts toward a more qualitative understanding of the modeling approaches themselves. This shift brings into focus a host of new questions having to do with the relationship between model and architecture, theory and implementation, the relative contributions of the modeler and the architecture to the final model, the role of parameter estimation in model development, the suitability of the simulated task to exercise features of the various architectures, the extensibility of the simulated task and the practical considerations that go into integrating disparate approaches within a common simulation environment. In the past, it might have been enough to address such questions in a one-off or ad-hoc fashion but with model comparisons becoming increasingly common we feel it is time to formulate more general answers to these kinds of questions and ultimately evolve a formal methodology to ensure the soundness of future efforts.

We will begin our discussion by way of example, announcing a new model comparison effort. We will briefly describe the task to be modeled, our motivation for selecting that task and what we expect the comparison to reveal. Next, we describe the programmatic details of the comparison, including a quick survey of the requirements for accessing, downloading and connecting different models to the simulated task environment. We hope to solicit audience input on this comparison and ways it might be improved. Moreover, we hope to encourage BRIMS attendees to participate in the comparison.

We will then turn the discussion to the panel, asking the members to reflect on their direct experience with model comparison. These efforts include the AFOSR AMBR modeling comparison (Gluck & Pew, 2005) and the NASA Human Error Modeling comparison (Foyle & Hooey, 2008). Our panelists have also entered cognitive models into multi-agent competitions (Billings, 2000) and organized symposia featuring competition between cognitive models as well as mixed human-model competitions (Lebiere & Bothell, 2004; Warwick, Allender, Strater and Yen, 2008). We will also ask panelists to describe how the current effort has been or should be shaped by these experiences. Finally, we will turn to our commentator to reflect on the important themes and issues that have been raised by the discussion. Given these general insights into the structure of successful model comparisons we will conclude by discussing the possibility of model comparison as a persistent activity at future BRIMS conferences.

Panelists

Walter Warwick (Chair) – Alion Science
Christian Lebiere – Carnegie Mellon University
Coty Gonzalez – Carnegie Mellon University
Kevin Gluck (Commentator) **–** Air Force Research Laboratory

References

Billings, D. (2000). The First International RoShamBo Programming Competition. *ICGA Journal*, Vol. 23, No. 1, pp. 42-50.

Foyle, D. & Hooey, B. (2008). *Human Performance Modeling in Aviation*. Mahwah, NJ: Erlbaum.

Gluck, K, & Pew, R. (2005). *Modeling Human Behavior with Integrated Cognitive Architectures*. Mahwah, NJ: Erlbaum.

Lebiere, C., & Bothell, D. (2004). Competitive Modeling Symposium: PokerBot World Series. In *Proceedings of the Sixth International Conference on Cognitive Modeling*, Pp. 32.

Warwick, W., Allender, L., Strater, L., & Yen, J. (2008). *AMBR Redux: Another Take on Model Comparison*. Symposium given at the Seventeenth Conference on Behavior Representation in Modeling and Simulation. Providence, RI.