Results and Lessons Learned from the 2009 DSF Model Comparison Challenge

Panel Chair
Walter Warwick
MA&D Operation, Alion Science and Technology
4949 Pearl East Circle, Suite #200
Boulder, CO 80301
303-442-6947
wwarwick@alionscience.com

At the 2009 BRIMS conference, we announced a model comparison challenge (Warwick, 2009; Lebiere, Gonzalez & Warwick, 2009). The challenge was based on modeling human performance on the dynamic stocks and flows (DSF), a generic control task that captures many of the complexities of dynamic decision making (Dutt & Gonzalez, 2007; Gonzalez & Dutt, 2007). The DSF was designed to be as simple and accessible as possible to computational modelers while focusing on two key ubiquitous components of general intelligence: the control of dynamical systems and the prediction of future events. A general call for participation was submitted to invite independent developers, of distinct computational approaches, to simulate human performance on the DSF task.

Nine different individuals or teams chose to participate in the challenge by developing computational models to simulate human performance on the DSF task in a variety of conditions. All participants were provided a description of the DSF task and samples of detailed human data that had been collected and reported in previous studies (Dutt & Gonzalez, 2007). In addition, sample software was provided to facilitate a socket-based connection between the models and the DSF simulation environment. The stated goal of the comparison challenge was to reproduce human behavior, including learning, mistakes, and limitations in a way that their models would generalize to *new* conditions of the task undisclosed to the participants. Results from three of the models were selected for presentation at the 2009 International Conference on Cognitive Modeling (Lebiere, Gonzalez, Dutt & Warwick, 2009). In addition, after the challenge was complete, we issued a call for papers for a special issue of the Journal of Artificial General Intelligence devoted to the challenge and its implications for advancing cognitive science and Artificial General Intelligence. The human performance data and the output from each model under every condition are available on the challenge web site:

<http://www.cmu.edu/ddmlab/modeldsf>

The goal of this panel discussion is to present our experiences in conducting the DSF comparison challenge and to reflect on the enterprises of model comparisons and modeling challenges in general. Walter Warwick will begin by discussing the motivation for this challenge and some of the issues faced in organizing it.

Next, we will turn to Varun Dutt of Carnegie Mellon University who will briefly review the DSF task itself, touching on both human performance in the laboratory and how he extended the experimental software to allow participants to link any model to the task environment supporting model comparison. He will also describe some of the challenges we faced, as organizers, in understanding the human performance and drawing meaningful comparisons among models. It became clear only after the fact that traditional measures of fit would not illuminate important performance differences among models on the DSF task.

The third panelist will be Kevin Gluck of the Air Force Research Laboratory. Gluck served in the role of Commentator in the previous panel on the DSF Comparison Challenge at BRIMS 2009. In that role, he recommended systematic exploration of the relative contributions of key mechanisms in all of the models that would be submitted, in order to establish the necessity of those mechanisms for predicting the transfer data. For any of several understandable reasons this did not happen as part of the standard process within the DSF Comparison Challenge. However, Gluck and colleagues at AFRL took on this and more as an independent set of supplementary analyses, exploring the complex interactions among architectural mechanisms, knowledge-level strategy variants, and task conditions. The general point motivating these efforts and to be summarized in Gluck's panel presentation is that the behavioral and cognitive modeling communities may reap greater scientific return on research investments – may achieve an improved understanding of architectures and models – if there is more emphasis on systematic sensitivity and necessity analyses during system development, evaluation, and comparison.

Finally, we will offer the first-hand experience of one of the participants. David Reitter, of Carnegie Mellon

University, submitted a cognitive model to the DSF challenge that generalized to yield the most accurate predictions of unseen data in novel conditions. He will report on the insights gained from his participation and from Gluck's subsequent parameter optimization and comparison with a competing model, pointing out three aspects of desirable progress in model evaluation: 1) generalization through prediction as opposed to post-hoc evaluation; 2) goodness-of-fit measures in numeric spaces other than the direct empirical measures obtained, yet; 3) the undesirable effect of "teaching the test" in competitions in other fields, such as Automatic Document Summarization and Machine Translation.

Although many of the issues we broach will be familiar to members of the BRIMS community, the challenges they present are no less urgent. In particular, this panel will provide a concrete, first-hand context for discussing questions about the representation of human variability in model performance, the need for task-specific quantitative measures of fit, the difficulty in expressing model content, the role of architectures in model development and the challenge in capturing the human cognitive ability to adapt to entirely new experiences. But more important than addressing those specific issues, we hope that the discussion will help us understand as a community what is needed to transition model comparison from an occasional and idiosyncratic exercise to a foundational research enterprise. Indeed, we see organized modeling comparisons and challenges as essential activities for advancing the science of human behavior representation but we cannot realize this vision without widespread community engagement.

Panelists

Walter Warwick (Chair) – Alion Science
Varun Dutt – Carnegie Mellon University
Kevin Gluck – Air Force Research Laboratory
David Reitter – Carnegie Mellon University

References

Dutt, V., & Gonzalez, C. (2007). Slope of inflow impacts dynamic decision making. In Proceedings of the 25th International Conference of the System Dynamics Society (pp. 79). Boston, MA: System Dynamics Society.

Gonzalez, C., & Dutt, V. (2007). Learning to control a dynamic task: A system dynamics cognitive model of the slope effect. In Lewis, Polk, & Laird (Eds.), 8th International Conference on Cognitive Modeling (pp. 61-66). Ann Arbor, MI.

Lebiere, C., Gonzalez, C., Dutt, V. & Warwick, W. (2009). Increasing Generalization Requirements for Cognitive Models: Comparing Models of Open-ended Behavior in Dynamic Decision-Making. In Proceedings of the 9th International Conference on Cognitive Modeling. Manchester, England.

Lebiere, C., Gonzalez, C., & Warwick, W. (2009). A Comparative Approach to Understanding General Intelligence: Predicting Cognitive Performance in an Open-ended Dynamic Task. In *Proceedings of the Second Artificial General Intelligence Conference (AGI-09).* Amsterdam-Paris: Atlantis Press.

Warwick, W. (2009) Comparing the Comparisons. In Proceedings of the Behavior Representation in Modeling and Simulation (BRIMS-09). Sundance, Utah.