Issues in the Conduct of Comparisons Among Human Performance Models and Live Human Performance in Complex Simulated Task Environments

Richard W. Pew

BBN Technologies, a Unit of GTE
50 Moulton Street
Cambridge, MA 02138
pew@bbn.com

Introduction

The Air Force Research Laboratory has sponsored the Agent-based Modeling and Behavior Representation (AMBR) Project to advance the state of the art in human performance modeling in general and the state of practice in cost-effective computer generated forces more specifically (Pew & Mavor, 1998). To advance human performance modeling they have created an opportunity for multiple developers to create different models of the same human operator activity and to compare the results both from model to model and with human participants performing the same task. As you will learn in a later talk, in the course of this project these model comparisons will be conducted three times, each time using different, hopefully more demanding, human modeling requirements. The first comparison is now complete and the results were reported at the International Ergonomics Society Meeting in July 2000. The project is important because it is rare to have the opportunity to validate human performance models, and even rarer to be able to compare and contrast the results of multiple model developers who use different model architectures and draw their models from different theoretical perspectives.

BBN has been assigned the responsibility of serving as moderator for these comparisons. In that capacity, for the first round, we developed a simulated, but much simplified, air traffic control environment and task scenarios that emphasized multiple task management, task priority setting and attention management. Using D-OMAR (Freeman, 1997), an agent-based simulation environment, we have the capability to interchangeably "plug in" a human operator model or a live human participant. This capability was used to collect human performance data exercising scenarios that were identical to those the model developers were required to exercise. Together with Michael Young, the project initiator, we developed the specifications the model developers were to respond to. We collected the human performance data supplied to the developers for developing their models and the data that was used for model evaluation. Together with the developers, we conducted the experimental comparisons among models, analyzed and reported the results. Four independent development teams participated.

In the course of completing these assignments we have had to address a number of challenging issues. This paper describes a sample of these issues and what we have done to resolve them. I make no claim that our solutions are the best ones. In some cases they were driven by practical considerations alone. I hope the paper will stimulate others to think about how they should be resolved.

Choice of Domain, Activity and Tasks

At the outset, we had the challenge of selecting the human performance tasks to be modeled. As I mentioned, it had been specified that the first comparisons would focus on multiple task management, dynamic priority setting and attention management. We were of two minds. We could select a task that was of practical interest, realistic complexity and required highly trained operators to be our participants. Or we could select a task that was highly abstracted, almost a video game, that anyone could be expected to learn but attempted to capture the task management requirements that were sought.

Clearly the first alternative has greater practical significance and is more challenging from a modeling perspective. However, it would require extensive knowledge acquisition on the part of each development team, an investment that would detract from the time and effort that could be put into the modeling itself. The moderator could supply that knowledge, but we all know that first hand knowledge is really required in order to address all the context-sensitive requirements. An overlay on this debate was whether we should be requiring the developers to model experienced operators or novice operators. There were strong arguments against modeling novices, because the likely variability they would produce in the data would mask the behaviors we were trying to measure.

Using the task of practical interest also had implications for the moderator team. We had limited resources for collecting data. Either we would have had to identify and recruit experienced operators from the domain under study, or we would have to invest in a very extensive period of training.

As a practical matter we opted to use the highly-abstracted version of an air traffic control task and utilized participants who had played a lot of video games, but had no previous experience with this task.. The task is probably not as representative of multiple task management requirements as we could have achieved with a more realistic task, but we obtained stable data from our human participants in four-hour sessions and the modelers were able to develop the requisite knowledge based on their own experience or that of a small set of previously untrained subjects.

However, the results were not as clear cut as we had hoped. Without going into too much detail, there were two display conditions and three workload conditions arranged in a factorial design. We set it up so that there would be two sets of scenarios and two samples of 8 participants each. The data from one set of scenarios and one sample of participants was provided to the developers for creating their models. Then at the time of evaluation the models were run on the second set of scenarios and compared to the data from the second sample of participants.. We were willing to take the risk that the participant samples would not be equivalent, but we lost our bet. For some of the performance measures the two samples produced patterns of results that were not entirely consistent. Once this was discovered, we decided to combine the data from the two samples and run the models on both sets of scenarios. This cleared up the problem and showed consistent effects of workload and display type, but left us with models that had been tuned on the basis of half the data and evaluated on a combination of the tuning and evaluation data. We concluded that, for the level of practice we provided, 8 participants were just not enough to produce reliable results. In the future we should use more participants or longer training periods or both.

What Data To Collect

We were charged with collecting human performance data not only to provide to the modelers for estimating various kinds of parameter values that might be of interest and to develop their models, but also data for validating and comparing one model result to another. What data would be useful, feasible and practical to collect?

Most applied experiments focus on outcome data, that is, data that evaluate the success of accomplishing the tasks assigned to the operators. We obviously wanted to obtain such data and considered it one important level at which to evaluate the performance of the models. Did they produce the same level of success as typical human participants? But we were also interested in encouraging models that did not simply mimic outcome data. We wanted them to simulate theoretically plausible internal mechanisms or human infrastructure that would produce the required performance. We were thinking of memory, attention, or perception mechanisms, for example, but, of course, we could not predict what mechanisms each modeler would build or assume. As a result, we had to provide what we perceived to be as generally useful data as possible. Furthermore, the data useful for developing the models may not be the same as data useful for evaluating them.

In addition, we had to think about what data it would be possible to collect from the models as well as from human participants. Each of the models was developed and run as an independent software module. As the evaluators, we only had access to the data that was explicitly provided to us through the application programming interface (API). When we ran the models with our test scenarios we could only measure data that passed through the API. We either had to insist that the model developers send certain specified data or else rely on the data they needed to send to make the model work in our environment for performance measurement. Since we could not specify in advance what data they would make available we had a problem being very specific in what data we would require.

Our solution was to provide data on penalty points, which was equivalent to outcome data in the sense that failures to perform tasks assigned or performing them late contributed to penalty points. However, we had great difficulty identifying sufficiently generic data that could be useful for evaluating the structural details of the models. We settled on collecting data on response times for the most elemental task decomposition elements for which we could reliably identify both a stimulus event and a response event. Then we also required the model developers to provide, derived from their theory and their models' performance, an estimate of workload level that we could compare with human participant subjective workload data for each condition as measured by the NASA TLX questionnaire. In addition we provided a trace of the time history of every action of each scenario for each subject in case the model developers wished to analyze it to obtain some other parameter or index. These traces also made it possible for the developers to rerun a trial as performed by a participant and watch the resultant activity on the ATC-like displays.

The penalty points did provide outcome data related to model performance, but data from all the models fell within the confidence bands established on the basis of the standard error of the human data and where the patterns were different there was no way to interpret the differences in terms of the theory that went into the models. Similarly all the models showed patterns of results for the aggregate response time data consistent with the human response to the different treatment conditions. We are now decomposing the response time data to look for more diagnostic results. The modeler's each took a different approach to modeling the subjective workload data and the concepts adopted were interesting, but all were able to adequately represent the trends in the data

collected from human participants. The bottom line is that we were not very successful in identifying or forecasting data that would discriminate among the models. It could be that our task was not demanding enough or sophisticated enough to provide such data. However, I now also believe it requires an understanding of the conceptual underpinnings of the models ahead of time in order to design the human data collection that will be truly diagnostic.

Simulation Fidelity

Finally, with respect to data comparisons, there were issues of simulation fidelity and data consistency. The human experimentation required that the external events in the simulation run in "real time" and be consistent from trial to trial. While this might seem simple, the simulation software was written in high level object-oriented languages, JAVA and LISP, for which we had limited control of statement execution timing. If "garbage collection" intervened or if statements requiring execution piled up, it could cause irregularities in simulation timing. Since a key kind of data we were collecting was timing information, we needed some assurance that the timing data we collected was not biased when we compared the human participant data with the model data and that there were no interactions between details of the scenarios and the computation delays for stimulus-response pairs. The modeler's were not restricted to run in real time. Their models could run faster or slower than real time, as necessary or practical. However, our participants , by definition ran in real time.

We adjusted parameters controlling the garbage collector to prevent significant intervention by the garbage collector during any of the human subject trials.. In order to assure ourselves, as confidently as we could, that the data were reliable and unbiased representations of real response times, we ran repeated pilot trials in a variety of scenarios representing the full range of task loads the participants would experience on ourselves, as very experienced participants, and on initially naïve participants from the same pool as our ultimate participant population,. Then we ran a simplistic model in the same conditions and examined both the repeatability across model runs and the very rough compatibility with the behavior of the human participants. We are confident that we have eliminated biases, but we never aspired to achieving millisecond level accuracy. More extensive cross calibration would be necessary to enable us to say that the response times we were measuring were accurate to a quantitative level of confidence, such as +/- 200 msec.

Conclusion

We have been advocates of running comparisons among human performance models on a common task for many years and have been delighted to be players in this modeling comparison project. However, it is not as easy as it seems at first blush. There are issues that arise that we, at least, never thought of until we found ourselves in the middle of task development. We have elaborated a sample of such issues but there are obviously many others. In our opinion the importance of undertaking such comparisons far outweighs the limitations that we have encountered in accomplishing this one. It is the most enlightened way to push the frontiers in terms of model architecture development as well as our understanding of stable, productive methodologies for moving from architecture to detailed, robust human performance models.

Acknowledgements

I wish to thank my colleagues, Yvette Tenney, Stephen Deutsch, Sandra Spector and Brett Benyo, for many contributions to this work. The project is sponsored by the U. S. Air Force Research Laboratory under Contract F33615-99-C-6002, Dr. Michael Young and Dr. Kevin Gluck, Project Monitors. Reproduction in whole or part is permitted for any purpose of the United States Government.

References

Freeman, B. (1997). *OMAR User/Programmer Manual, Version 2.0.* BBN Report No. 8181. Cambridge, MA: BBN Corporation. (http://www.he.afrl.af.mil/hes/hess/programs/omar/omardocs/index.html)

Pew, R. W. & Mavor, A. S. (1998) *Modeling Human and Organizational Behavior.* Washington DC: National Academy Press.