Validating Human Behavior Representations: Moving Beyond "Preaching" To "Practice"

Gwendolyn E. Campbell, Ph.D.
NAVAIR Orlando TSD
Orlando, Florida
Gwendolyn.Campbell@navy.mil

Amy E. Bolton
Strategic Analysis, Inc.
Arlington, Virginia
abolton@sainc.com

ABSTRACT

While it is widely agreed that human behavior representations (HBRs) must be validated before they are incorporated into military simulations, there is much less agreement on what activities and evidence satisfy validation requirements. In this paper we will begin by discussing psychological taxonomies of theory and measurement validity, identifying some insights that the Department of Defense Modeling and Simulation community might gain from these well-established paradigms. This discussion will include brief descriptions of a variety of techniques for collecting validity evidence gleaned from the psychological research literature. While qualitative evidence will be mentioned, special emphasis will be placed on quantitative techniques for assessing HBR validity. A number of relevant issues, such as appropriate and inappropriate statistical tests, overfitting data, and model complexity, will be addressed. Next, we will discuss some limitations of the psychological perspective in general and for our community in particular. Finally, we will expand on Defense Modeling and Simulation Office's definition of validity and illustrate how this definition provides guidance for additional HBR assessment measures and processes that are highly appropriate for the military user community.

ABOUT THE AUTHORS

Gwendolyn E. Campbell is a Senior Research Psychologist at NAVAIR Orlando, Training Systems Division. She holds an M.S. and Ph.D. in Experimental Psychology from the University of South Florida and a B.A. in Mathematics from Youngstown State University. Her research interests include the application of human performance modeling techniques within training systems and the development of a cognitively based science of instruction.

Amy E. Bolton is a Staff Research Psychologist at Strategic Analysis, Incorporated. She holds a M.S. in Cognitive Human Factors and is a doctoral candidate in the Applied Experimental and Human Factors Psychology Ph.D. program at the University of Central Florida. Her research interests include the application of intelligent agents in individual and team training for dynamic tasks. She is the Principal Investigator on a project investigating the appropriate timing of feedback delivery during scenario-based training.

Validating Human Behavior Representations: Moving Beyond "Preaching" To "Practice"

Gwendolyn E. Campbell, Ph.D.
NAVAIR Orlando TSD
Orlando, Florida
Gwendolyn.Campbell@navy.mil

Amy E. Bolton
Strategic Analysis, Inc.
Arlington, Virginia
abolton@sainc.com

INTRODUCTION

It is generally acknowledged that human behavior representations (HBRs) should be validated before they are widely integrated into other Department of Defense (DoD) military simulations (e.g., U.S. Department of Defense, 2001a; Committee on Technology for Future Naval Forces, National Research Counsel [NRC], 2003). It is also agreed that validation is difficult, costly and rarely done well (e.g., Ritter & Larkin, 1994; U.S. Department of Defense, 2001b; Harmon, Hoffman, Gonzalez, Knauf, & Barr, 1999).

There is less agreement, however, on the nature of the activities and evidence that are sufficient to justify statements about the validity of an HBR. This paper is divided into two main sections. In the first section, we investigate the potential lessons learned about model validation from an established and relevant academic community. In the second section, we move beyond this academic and theoretical perspective on validity, and present a more practical approach to collecting validity evidence in the applied military modeling and simulation (M&S) community. A more extensive treatment of this topic can be found in Campbell and Bolton (2005).

PSYCHOLOGICAL PERSPECTIVE ON VALIDATION

HBRs come in many forms – ranging from mathematical functions to symbolic, rule-based software programs. While it is not always the case, HBRs are often thought of as executable mini-theories of human behavior. This analogy suggests that insight into techniques for validating HBRs may be sought among the validation taxonomies and techniques applied in the psychological community to its theoretical models.

Validation Taxonomies

There are two classic psychological taxonomies of validity. The first, initially proposed by Cronbach and Meehl (1955), focused on the validation of psychological tests. Cronbach and Meehl distinguished three different ways in which a test score could have meaning. If a person's score on one test (such as an entrance exam) predicts that person's performance on some other test (such as G.P.A.), then that test is said to have criterion validity. If it is possible to demonstrate that a test includes questions from every area of a domain, then the test is said to have content validity. Finally, if a test can be shown to provide accurate information about some underlying psychological trait or characteristic of the test taker, then the test is said to have construct validity.

The second classic psychological validation taxonomy, described by Cook and Campbell (1979), focused on the validation of an experiment or experimental design. Cook and Campbell distinguished different dimensions along which the validity of an experiment could be assessed. An experiment must meet certain requirements for its statistical tests to be appropriate, such as having a sufficiently large sample size, and Cook and Campbell referred to this form of validity as statistical conclusion validity. An experiment's internal validity refers to the extent to which the researcher can draw accurate conclusions about the causal relationships among the manipulated and observed variables, and the experiment's external validity refers to the extent to which those causal relationships generalize across populations and environmental conditions. Finally, the extent to which the causal relationships among the manipulated and observed variables accurately represent causal relationships among underlying psychological traits and processes is referred to as an experiment's construct validity.

One thing that is interesting to note about these two taxonomies is that both include a category called construct validity. Both definitions of construct validity assume that humans have underlying psychological traits and processes that cannot be directly observed, but can only be inferred based on observable behavior. The more information a test or experiment provides about those underlying traits and processes, the more confidence we have that the test or experiment is construct valid. The extension to HBRs

is obvious. The more closely an HBR corresponds to underlying human traits and processes, the more confidence we may have in the validity of that HBR. This extension suggests that the types of evidence useful for demonstrating the construct validity of a test or experiment could also be useful for demonstrating the validity of an HBR.

Qualitative Evidence for Validity

The phrase "qualitative evidence" typically implies asking someone – hopefully someone with relevant credentials – for his or her opinion about the validity of a measure or model. In the psychological community, this type of evidence is not thought to provide strong support for construct validity and, instead, is thought to address a measure or model's face validity. One of the key issues that limits the confidence that psychologists have in qualitative evidence is that human judgments are prone to a number of well-documented limitations and biases (Gilovich, 1993; Tversky & Kahneman, 1974). As just one example, once a person has formed an opinion, that person is unconsciously predisposed to only seek evidence that confirms that opinion and will often discount any discrepant evidence (Gilovich, 1993). Thus, face validity is typically considered to be the weakest form of validity.

This does not mean that qualitative evidence is useless for the DoD M&S community. Certainly collecting SME assessments of an HBR encourages buy-in from the user community. In addition, if collected carefully, this type of evidence can also fulfill the powerful role of helping to identify specific ways in which to improve the HBR. What constitutes a "careful" collection procedure? First, the process should be standardized, objective, systematic and repeatable. This can be accomplished by developing a formal procedure involving questionnaires, checklists or structured interviews in advance. Then this procedure would be applied identically to each of several SMEs who are selected for their specific knowledge of the subject matter.

Second, the process should be independent. There are two requirements for achieving a desirable level of independence. First, the SMEs who were used to help develop the model should not be the only SMEs involved in the evaluation of the model, as they will have a vested interest in the outcome of the validation process. Second, several SMEs should make their judgments independently of each other before the results are compared and combined. Qualitative evidence for validity can be attributed to those aspects of the model that elicit positive feedback independently from all the SMEs, but not to model components that elicit disagreement or consistent concern. Above all else, the key is to avoid the BOGSAT (bunch of guys sitting around a table) process (Committee on Technology for Future Naval Forces, NRC, 2003).

Unfortunately, no matter how carefully qualitative evidence is collected, SME judgments alone are insufficient to establish the validity of an HBR.

Quantitative Evidence for Validity

An obvious alternative to collecting qualitative evidence in the form of human judgments of the apparent reasonableness of an HBR's behavior is to collect quantitative evidence in the form of a statistical assessment of the similarity between an HBR's behavior and a human's behavior. In fact, there is a long-standing tradition of comparing model predictions to empirical data. Roberts and Pashler (2000) cite examples from the psychological literature going back over 60 years.

Statistical tests

So, what statistical test(s) should be used to compare samples of human behavior and model performance to support a claim of model validity? Unfortunately, the traditional statistical approach of hypothesis testing (using tests such as the Student's t-test and the F-test for analysis of variance) cannot be applied to compare these two samples of data (Grant, 1962). These tests were designed to support traditional experiments, in which the researcher is hoping to demonstrate that a treatment or intervention of some sort has had an impact and thus the two sets of data (from the control condition and the experimental condition) are fundamentally different (or, in statistical terms, came from different underlying populations). When evaluating the validity of an HBR, however, the researcher is hoping to show that the model's performance is just like a person's behavior. Statistical hypothesis testing simply cannot be applied to support a claim that two sets of data (from the human and the model) are fundamentally similar.

Instead, the statistical techniques that are appropriate in this case are called goodness-of-fit tests. There are a large number of goodness-of-fit tests available. Recently, Schunn and Wallach (2001) proposed that the goodness-of-fit between a model's predictions and empirical data should be assessed along two dimensions: trend consistency and exact match. They discussed the strengths and weaknesses of a number of statistical options for assessing these aspects, and ultimately recommended, when possible, calculating r^2 to assess trend consistency and the Root Mean Squared Scaled Deviation (RMSSD) to assess the exact match.

Demonstrating high goodness-of-fit statistics between a model's predictions and a set of human performance data can be powerful evidence in an argument for the construct validity of that model. Unfortunately, this is still not considered sufficient evidence to support a conclusive claim of construct validity (Roberts & Pashler, 2000). There are a number of reasons for this, and we shall briefly address two of those reasons next.

Overfitting

Measurement theory tells us that the set of human behavioral data being used to evaluate a model's performance always has a component of error variance. Mathematical theory tells us that many modeling techniques are so powerful that they may not only capture some of the systematic variance in data, but they may also capture some of the error variance. This condition is known as overfitting, and the problem with overfitting is that a model that overfits one set of data (as evidenced by high goodness-of-fit statistics) will not generalize to any other set of data. A model that can only match one set of data is unlikely to accurately capture the underlying psychological traits and processes that produced that set of data. It would be comparable to memorizing a set of specific responses without any understanding of how those responses were generated in the first place.

There are techniques that are routinely used in the mathematical modeling community to assess the extent to which a model is overfitting a data set. A common technique, cross-validation, is to divide the empirical data set into two subsets, and use one subset to build (or "train") the model and the other subset to evaluate (or "test") the model. The best indicator of a model that has only captured the systematic variance in empirical data is a model that demonstrates similar goodness-of-fit values on the two subsets of data. A model that fits the training data well, but does not demonstrate good fit to the testing data, has probably been overfit, and is unlikely to be valid. There is an important extension of this technique, called bootstrapping, which produces an even more reliable indicator of the extent to which a modeling technique is overfitting the training data. A nice illustration of this process can be found in Dorsey & Coovert (2003).

Model complexity

A second factor that can undermine the meaningfulness of a high goodness-of-fit statistic between human behavioral data and a model's performance is the complexity or powerfulness of model. Mathematicians have long known that there are models that are sufficiently powerful that they could fit any set of data, even data that could not conceivably have been produced by a group of human participants. Obviously, that type of model would not be a valid representation of underlying psychological traits and processes. The components of a model that contribute to its level of power include the number of free parameters in the model and its functional form. Taken together, these are often referred to as a model's level of **complexity**. This suggests that a model's goodness-of-fit to data should be interpreted in light of the model's level of complexity, and more "credit" given to a less complex model that is able to fit empirical data.

There are a number of quantitative techniques for adjusting goodness-of-fit measures to take model complexity into account, and a recent study (Pitt, Myung & Zhang, 2002) compared the effectiveness of four. They found that one technique, the minimum description length (MDL), was the least likely of the four to be "fooled" into giving the highest goodness-of-fit scores to an unnecessarily complex and powerful model. Further work needs to be done, especially in the area of extending these ideas to symbolic and rule-based HBRs that are not easily represented in a closed mathematical form.

Moving beyond simple goodness-of-fit tests

There are at least two, non-mathematical ways to modify the process of using one or two goodness-of-fit statistics to compare one set of human behavioral data to one set of model data that increase our ability to draw conclusions about the validity of that model. One is to increase the scope of the comparison by including a multivariate data set. This approach has been called the pattern matching perspective (Trochim, 1985, 1989). An impressive application of a multivariate pattern approach can be found in the AMBR program. As described in Pew and Gluck (2005), each model was assessed on its ability to fit and/or predict: (a) a diverse set of performance measures, including reaction time and performance accuracy on primary and secondary tasks and self-reported workload, (b) performance data at multiple levels of aggregation, and (c) performance under a wide variety of conditions, including manipulations of the system interface, task load, and the cognitive complexity of category learning task. Needless to say, the more complex a pattern of human data that can be fit with an HBR, the stronger the evidence for the validity of the HBR.

A second, non-mathematical way to increase our ability to draw conclusions about a model's validity from quantitative data is to assess the model's ability to make a priori predictions about human behavior under new conditions, before the model developer has access to any data collected under these conditions.

Recall that a typical cross-validation approach is to take a single coherent set of human behavioral data and split it into two subsets, one for training and one for testing. In other words, the model is only being asked to "predict" behavioral data that is likely to be highly similar to the data used to build the model. The testing data may come from different people, but it was collected at the same time and under the same circumstances as the training data. Obviously, accurately predicting human performance under different conditions is a much more difficult and stringent test, and thus would provide more compelling evidence of an HBR's validity.

As with the pattern matching perspective, an example of this evaluation approach can be found in the AMBR program (Pew & Gluck, 2005). As described in Pew and Gluck, in the final phase of the AMBR project, modelers were provided with human performance data collected during a category-learning task, asked to develop models of this task, and then those models were then evaluated on their ability to fit those data. However, human participants also completed a transfer task, during which they were asked to categorize (without receiving any feedback) stimuli with values that they had never seen before. The AMBR models were asked to predict the participants' responses to these novel stimuli without being provided with any representative performance data from this stage of the experiment. The difficulty of this challenge can be seen by comparing the ability of the models to fit the category learning data to their ability to predict the transfer data. Our confidence in the validity of an HBR is correlated with the difficulty of the test that the HBR is able to pass.

Limitations of the Psychological Approach

While there is clearly a great deal that the DoD M&S community can learn from the psychological community in reference to validating HBRs, there is a limit to the appropriateness of the psychological notion of construct validity in our context. In particular, within the psychological community, "construct validity" means that a theory embodies an accurate description of the actual underlying processes that explain human behavior, and that alternative theories can be disregarded.

We propose that this basic goal is fundamentally inappropriate for our community. As Anderson (1993) explains, the process of developing a computational model that represents a particular theory requires the modeler to make a large number of implementation decisions that are irrelevant from the perspective of the theory. Not only are these details theoretically irrelevant, there are many different implementations capable of producing the same model output or behavior. This means that the goal of finding the "single correct" computational model is, quite simply, misguided.

This leaves us with a question: what is an appropriate goal for the developers and users of HBRs in applied, military settings? What does it mean to say that our models must be validated? In the next section, we will begin by reminding the reader of the DMSO definition of validity, and go on to explain how unpacking this definition will provide insight into the measures and processes for assessing HBRs that are most appropriate for this community.

DMSO PERSPECTIVE ON VALIDATION

Definition

The Defense Modeling and Simulation Office (DMSO) defines validation as "…the degree to which a model or simulation is a faithful representation of the real world from the perspective of the intended uses of that model or simulation" (U.S. Department of Defense, 2001c). It is easy to focus on the first part of the definition, "faithful representation of the real world," which sounds quite a bit like the psychological notion of construct validity. In fact, the second part of the definition, "from the perspective of the intended uses," is a critical qualifier with the potential to provide significant guidance into the processes, metrics and requirements associated with assessing the validity of an HBR.

General Implications for Assessment

The first step in unpacking this definition is to identify the "intended use" of an HBR in a military community. At a high level, the military has a straightforward and pragmatic goal, which is to improve military capability, by, among other things, improving human performance. Different sub-communities will use HBRs in different ways in order to serve this higher-order goal. The military training community, for example, uses HBRs as synthetic adversaries in training simulators, to increase the effectiveness of the training activities, and thus improve performance of the trainees. The military acquisition community, on the other hand, uses HBRs to evaluate candidate system designs and identify those designs that are likely to lead to the best human-system integration, and thus improve human performance. The point is that an HBR can be assessed directly against its ability to support a particular intended use.

Interestingly, there is at least some evidence that improving human performance does not necessarily require a construct valid model. For example, in a recent training study (Bolton, Buff & Campbell, 2003), researchers compared the capability of three different models of expert performance to provide the basis for effective feedback. The models were all generated from the same expert performance, but represented fundamentally different reasoning strategies, so, at best, only one of them could have been "construct valid" in the psychological sense. However, all of the models supported the development of effective feedback – feedback that led to a statistically significant improvement in trainee performance. In other words, even the models that were not construct valid were capable of meeting the goals of this particular application. This suggests that a model's capability to serve an applied goal (DMSO's definition of validity) is not necessarily equivalent to its construct validity. In order to distinguish these two types of validity, we will use the term "application validity" to capture DMSO's meaning.

We are not trying to imply that assessing a model's application validity will ever be a trivial undertaking. But we contend that taking an "intended use" perspective serves two purposes: (a) it will bound the scope of the validation problem, and (b) it may provide insight into the activities, metrics and measurement paradigm that could be used to demonstrate application validity. In the next section, we will expand on this contention within the context of the military training community's use of HBRs.

Training Application Example

Military training, when designed correctly, is probably the application that provides the most "built-in" support for HBR validation efforts. First, most training systems will have mechanisms in place to assess student learning and performance. A simple reinterpretation of these measures provides insight into the effectiveness of the training system itself. In addition, the ultimate goal of these training systems, to improve human performance, will already have been cast in terms of a number of well-defined learning objectives, learning activities will have been planned, the conditions under which performance will be assessed will have been established and performance criteria will have been set. All of this existing infrastructure can be leveraged when validating HBRs developed to be incorporated into training simulators.

Demonstrating application validity of an HBR within a training community would require demonstrating that the incorporation of an HBR into a training system leads to some benefit, with improved human performance being the most obvious. There are many possible ways to measure this. Some common approaches include: (a) the average performance of a group of students increases, (b) the number of students who fail to meet some minimum criterion decreases, and (c) the amount of time it takes the average student to reach a criterion decreases. All of these measures are probably already being collected. In some cases, the original training system (sans HBR) can serve as the control condition and baseline data may even exist, reducing the burden associated with demonstrating an HBR's application validity even further.

SUMMARY

HBR validation may be a costly, difficult and time consuming process, but the risk of drawing erroneous conclusions from unvalidated models is simply unacceptable (U.S. Department of Defense, 2001a). We can learn a lot about model validation from the psychological community, including techniques to improve the quality of the qualitative data we collect from SMEs, the strengths and weaknesses of statistical tests for assessing the goodness-of-fit between an HBR's performance and a set of human behavioral data, and non-mathematical techniques to increase the strength of those quantitative comparisons. Ultimately, however, the goals of the academic psychological community the DoD M&S community are not the same, and we need to consider our own goals when determining how to assess an HBR's validity. We have proposed, in fact, that a careful analysis of the intended use of an HBR will help bound the validation problem and make it more tractable. We can't afford to skip the validation process because it is too difficult or costly, and we can't afford to conduct a weak or meaningless assessment by following a flawed validation procedure. It is time to practice what we preach, allowing our application goals to guide us.

REFERENCES

Anderson, J. (1993). Rules of the mind. Hillsdale, NJ: Erlbaum.

Bolton, A. E., Buff, W. L., & Campbell, G. E. (May, 2003). Faster, cheaper, and "just as good"? A comparison of the instructional effectiveness of three HBRs that vary in development requirements. Presented at the 12th Conference on Behavior Representation in Modeling and Simulation, Scottsdale, AZ.

Campbell, G. E, & Bolton, A. E. (2005). HBR validation: Integrating lessons learned from multiple academic disciplines, applied communities and the AMBR project. In R.W. Pew & K. Gluck (Eds.), *Modeling Human Behavior with Integrated Cognitive Architectures: Comparison, Evaluation, and Validation*. Lawrence Erlbaum. Pp. 365-397.

Committee on Technology for Future Naval Forces, National Research Counsel (NRC) (2003). *Technology for the United States Navy and Marine Corps, 2000-2035 becoming a 21st-century force: Volume 9: Modeling and simulation.* National Academies Press. Retrieved April 1, 2004 from http://books.nap.edu/catalog/5869.html

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton Mifflin.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

Dorsey, D. W., & Coovert, M. D. (2003). Mathematical modeling of decision making: A soft and fuzzy approach to capturing hard decisions [Special issue]. *Human Factors*, *45*(1), 117-135.

Gilovich, T. (1993). *How we know what isn't so: The fallibility of human reason in every day life.* New York: Free Press.

Grant, D. A. 1962. Testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review, 69*, 54-61.

Harmon, S. Y., Hoffman, C. W. D., Gonzalez, A. J., Knauf, R., & Barr, V. B. (1999). *Validation of human behavior representations.* Retrieved April 1, 2004 from https://www.dmso.mil/public/library/projects/vva/found_02/sess_papers/b3_harmon.pdf

Pew, R.W., & Gluck, K. (Eds.) (2005). *Modeling Human Behavior with Integrated Cognitive Architectures: Comparison, Evaluation, and Validation*. Lawrence Erlbaum.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review, 109*(3), 472-491.

Ritter, F. E. and Larkin, J. H. (1994). Using process models to summarize sequences of human actions. *Human-Computer Interaction*, *9*(3), 345-383.

Roberts, S. & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review, 107*(2), 358-367.

Schunn, C. D. & Wallach, D. (2001). *Evaluating goodness-of-fit in comparison of models to data*. Online manuscript. Retrieved April 1, 2004 from http://www.lrdc.pitt.edu/schunn/gof/GOF.doc

Trochim, W. (1985). Pattern matching, validity, and conceptualization in program evaluation. *Evaluation Review, 9*(5), 575-604.

Trochim, W. (1989). Outcome pattern matching and program theory. *Evaluation and Program Planning, 12*, 355-366.

Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124-1131.

U.S. Department of Defense (2001a). *A Practitioner's Perspective on Simulation Validation: RPG Reference Document.* Washington, DC: Defense Modeling and Simulation Office. Retrieved April 1, 2004 from http://vva.dmso.mil/Ref_Docs/Val_Lawref/Val-LawRef-pr.pdf

U.S. Department of Defense (2001b). Validation of Human Behavior Representations: RPG Special Topic. Washington, DC: Defense Modeling and Simulation Office. Retrieved April 1, 2004 from http://vva.dmso.mil/Special_Topics/HBR-Validation/nbr-validation-pr.pdf

U.S. Department of Defense (2001c). VV&A Recommended Practices Guide Glossary. Washington, DC: Defense Modeling and Simulation Office. Retrieved April 1, 2004 from http://vva.dmso.mil/Glossary/Glossary-pr.pdf