Social processes in validation: Comments on Grant (1962) and Roberts and Pashler (2000)

Comments as part of Symposium on Model Fitting and Parameter Estimation at the ACT-R Workshop

Frank E. Ritter
19 July 2003

Grant (1962) argued that there were two important aspects for a model, that (a) it was worth taking seriously, and (b) you could see where it was wrong so that you could improve it. This is consistent with Newell's (1990, p. 507) view of UTCs: what is the current bar (standards), does my theory raise it, and what are the further regularities to be included in the future? I like this approach as it lets me make progress, or at least be happy. I have seen others trying to prove their model, and they are not and cannot be happy.

Taking a model seriously depends on what other models are available and what you want to do with it. What is the current bar? If you look at current theory/data comparisons of task performance models (e.g., ACT-R), the models can match several types of data and do so in different ways. For example, the data may include reaction time means, standard deviations of reaction times, the sequence of task actions, groupings of task actions into strategies, error rates and types, and trends in all of these. Summarizing the match across these sets of regularities can be done in multiple ways. For example, John (1996) has used a type of bar chart across a set of different types of behaviour being matched. Other comparisons you will see here will just report performance (it can do the task) and correlation between predictions and data, which Simon and Grant both recommend, and often which leads me to take models seriously.

With multiple types of data with multiple values and multiple displays, how can one compare theories? I currently think that Grant's question, of is a theory worth taking seriously, can be seen at least partly as a social process. What is currently an interesting theory or an impressive fit will vary on a large number of difficult to reduce dimensions. I like one model here because it does not touch the simulation and offers new worlds to models (but it has a crappy fit). I like another model because it opens up new areas of data to be included into ACT-R. I like a third because it offers a way to fairly test millions of combinations of parameters to do docking (Burton, 1998) — to compute what types of human characteristics best fit a dataset (that is, what develops in children?). In each case, the judgment of "is this model interesting" is based on other models, how well the model fits the data, what use I have for the theory, how easy the theory is to use, and a host of other factors, such as future applicability (science, like politics, is the art of the possible, said Newell, and I rather strongly agree). That means that I take models that I can download and include with my model much more seriously than those that I cannot inspect and cost $1,000.

A recent set of comments (Roberts & Pashler, 2000, 2002; Rodgers & Rowe, 2002) argue that a reader needs to know more than the fit, particularly, readers need to know what kind of data that the theory cannot fit, the variability of the data, and the likelihood of fitting data. Roberts and Pashler's stance appears to be consistent basically with Grant's two step process, but they ask for more details. The details they ask for appear to me to be more relevant for simple models covering well trod but narrow ground rather than broad theories. Roberts and Pashler do request a standard that is worth striving for, but they also

appear to be more ready to bite fingers than to look at the direction the fingers are pointing. I also find much of psychology data surprising, which they do not, and I know of several theories that do not predict smooth curves and the data matches, and they think smooth curves are the almost universal norm. Finally, I believe that task performance is much more important than fits. Most authors have not credited a model for performing the task as much as I think they should.

So, (a) I recommend that you tell us about your model's predictions, what the data look like, and how the match goes in detail, enough so that we can see that the model is worth taking seriously. There is no a priori quality required. You might also note its other virtues, such as ease of use, consistency but not yet correlation with large swaths of behaviour.

There are also reasons to dismiss a model. If the model would fit any data, then it is not worth taking seriously (but only if such data already exist, hypothesized data need not apply). If I cannot understand the model; if it is a hack; or if I believe it will not generalize to other data; and I would add now, if it is not part of a UTC, I am less interested. I think most models here are worth taking seriously.

I also recommend that you (b) Note where the model can be improved. This does not mean 40 pages of comments in reviews, or a laundry list of data that your model does not yet cover because you ran out of time (Law 4b. Good intentions are far more difficult to cope with than malicious behavior and Law 8, No amount of genius can overcome a preoccupation with detail. *Levy's Ten Laws of the Disillusionment of the True Liberal*). I would include just enough for readers to know that you know where the holes are, and know enough to improve your model, but not to apologize for tasks it cannot yet do.

Acknowledgment

Preparation of this has been supported by a grant from the Office of Naval Research, N00014-03-1-0248.

References

Burton, R. (1998). Validating and docking: An overview, summary and challenge. In M. Prietula, K. Carley, & L. Gasser (Eds.), *Dynamics of organizations.* 215-228. Menlo Park, CA: AAAI.

Grant, D. A. (1962). Testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review, 69*(1), 54-61.

John, B. E. (1996). TYPIST: A theory of performance in skilled typing. Human-Computer Interaction, 11(4), 321-355.

Newell, A. (1990). Unified theories of cognition. Cambridge, MA: Harvard University Press.

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

Roberts, S., & Pashler, H. (2002). Reply to Rodgers and Rowe (2002). Psychological Review, 109(3), 605-607.

Rodgers, J. L., & Rowe, D. C. (2002). Theory development should begin (but not end) with good empirical fits: A comment on Roberts and Pashler (200). Psychological Review, 109(3), 599-604.

Also see acs.ist.psu.edu/papers/ (email me if you need a password)