Global Model Analysis by Landscaping

Daniel J. Navarro, In Jae Myung, Mark A. Pitt and Woojae Kim
{navarro.20, myung.1, pitt.2, kim.1124}@osu.edu
Department of Psychology
Ohio State University
1827 Neil Avenue Mall
Columbus, OH 43210, USA

Abstract

How well do you know your favorite computational model of cognition? Most likely your knowledge of its behavior has accrued from tests of its ability to mimic human data, what we call local analyses because performance is assessed in a specific testing situation. Global model analysis by landscaping is an approach that "sketches" out the performance of a model at all of its parameter values, creating a landscape of how the relative performance abilities of the model and a competing model. We demonstrate the usefulness of landscaping by comparing two models of information integration (Fuzzy Logic Model of Perception and the Linear Integration Model). The results show that model distinguishability is akin to power, and may be improved by increasing the sample size, using better statistics, or redesigning the experiment. We show how landscaping can be used to measure this improvement.

Introduction

The development and testing of theories is one of the most important aspects of scientific inquiry. Theories provide us with the tools we need to understand the world (Kuhn, 1962), and frequently spark new and exciting experimental work (Estes, 2002). When we develop a mathematical or computational model of a cognitive process, it is generally an instantiation of a few fundamental properties of a verbal theory. This translation process requires some degrees of freedom because data sets can vary in many ways, and still be consistent with the qualitative ideas in the model. For instance, while forgetting curves generally look like decreasing exponential or power functions (e.g., Rubin & Wenzel, 1996), some people remember items more easily than others. We capture this idea by proposing models that have a number of free parameters which may be fine-tuned to fit the data. This idea has been widely adopted, and has led to successful models of a wide variety of cognitive phenomena.

One potential drawback with this approach is that, as the models we propose become more elaborate, the task of understanding the model itself becomes increasingly difficult. When assessing the performance of a model in light of some observed data, we generally try to find a single set of parameters which allow the model to fit the data best. Maximum likelihood estimation and least sum of squares methods are examples of this approach. Because a model is evaluated by identifying the single, best-fitting parameter set, these methods are examples of *local model analysis*. Like an iceberg, the vast majority of the model is hidden beneath the waves: only those few parameter sets that provide best fits ever allow the model to come to the surface and show us how it behaves. As a result, our experience of the model is limited to a few, possibly unrepresentative cases.

As modelers, we want to learn something about how our models behave in general, not just at the few specific points that come to light when we use local methods such as fitting data sets generated in an experiment. The limitations of such methods leave a number of interesting and important questions unanswered. We may be concerned that our model makes so many predictions that it could provide a good fit to almost any data (model complexity), that many different models make essentially the same predictions (model distinguishability), or that the predictions do not conform to the original qualitative theory (model faithfulness). Questions such as these can be referred to as issues of *global model analysis*. Global model analysis, as we conceive it, refers to the task of discovering what a model can and cannot do, particularly in light of empirical data and other models. In this paper, we introduce a simple global analysis method that we call landscaping.

Sketching a Landscape

The idea underlying landscaping is remarkably simple. Find all the things that a model can do, compare it to the things that other models can do, and see how these things relate to empirical data. In one sense it is the very opposite of parameter optimization (i.e., finding the best fit): rather than look for a single set of parameters (and a single prediction), we look at them all. When we do this, we obtain the full range of predictions made by the model, which we call the landscape.

Landscaping is a modeling tool, not a statistic, and can be adapted to answer many questions. In

this paper we are concerned primarily with model distinguishability and the relationship to experimental design.

The Domain of Application

Models of information integration are concerned primarily with stimulus identification. Given some potentially ambiguous information from multiple sources regarding a stimulus, what is the stimulus most likely to be? The classic example is phonemic identification, in which visual and auditory cues are combined in order to make a decision (e.g. was it /ba/ or /da/?). In this study we compared the landscapes of two models of information integration: Oden and Massaro's (1978) Fuzzy Logic Model of Perception (FLMP) and Anderson's (1981) Linear Integration Model (LIM).

To briefly summarize an extensive literature: FLMP provides a remarkably good account of a wide range of phenomena, including some that look rather LIM-like and many that do not. There are a number of local analyses of both these models. It would be useful to unify them in some way to understand better how their performance is related. Landscaping does just this. In the process it provides answers to related questions. For instance, are there large numbers of FLMP patterns that are never observed in experimental data? Are there LIM patterns that FLMP cannot mimic, or is FLMP flexible enough to fit all LIM-like patterns? Is experimental design important? In short, we want to find out what lies beneath the surface of all these local analyses.

Furthermore, we want to answer these questions with respect to the kind of small sample sizes that characterize real experiments. With that in mind, we approach the comparison as experimentalists. We have in mind a particular experiment that we wish to conduct, and have a number of questions about the relationships between FLMP, LIM, and experimental data. For example, what kinds of data sets are consistent with each of the two models? Are there some kinds of data that are consistent with both models? How successfully can our experiment tell the two models apart? What statistics will we need to do so? These questions can be very difficult to answer using local methods, but readily fall out of a landscaping analysis.

Experiment One

The experiment that we have in mind is a two-choice identification task (i.e. choose A or B) with a 2×8 design. In other words, there are two different levels of one source (e.g., visual) , $i \in (i_1, i_2)$, and eight different levels of the other source (e.g., auditory) $j \in (j_1, \ldots, j_8)$. In total, there are 16 stimuli that may be produced by combining the two evidence sources[1]. Furthermore, we anticipate a sample size of $N = 24$ (not unusual in psychological experiments). Letting p_{ij} denote the probability of responding "A" when presented with the i-th level of one source and the j-th level of the other, FLMP is characterized by the equation,

$$p_{ij} = \frac{\theta_i \lambda_j}{\theta_i \lambda_j + (1-\theta_i)(1-\lambda_j)},$$

whereas LIM predicts that

$$p_{ij} = \frac{\theta_i + \lambda_j}{2}.$$

In both cases we assume that $\theta_i \leq \theta_{i+1}$ and $\lambda_j \leq \lambda_{j+1}$ for all i and j.

The Landscape of Model Fits

Our landscaping analysis consists of generating a large number "experimental" data sets from each model: that is, the kind of data the we would expect to observe if FLMP (or LIM) were the true model of human performance. The comparison between the models is then based on how well each model fits all of these data sets. The results reveal the distinguishability of the models and their similarities and differences.

Generating hypothetical experimental data is simple. In a two-choice task, both FLMP and LIM assume that the sampling error follows a binomial distribution (with $N = 24$ in this case). To sketch a landscape of FLMP data, we randomly generated a large number of FLMP parameter sets (10,000 in this case), found the p_{ij} values, and added sampling error[2]. This was then repeated using LIM.

Because each of these data sets represents the potential outcome of an information integration experiment, by fitting both FLMP and LIM to them (using maximum likelihood estimation) we can determine how effectively an experiment of this kind will discriminate between the two models. Intuitively, one expects the generating model to fit its own data better than the competing model, but due to the joint

[1] It is well known that FLMP is non-identifiable for this experimental design, but that we may fix one parameter value (say, θ_1) without loss of generality (Crowther, Batchelder & Hu 1995). Although this technique does not work for LIM, LIM may be reparametrized as the identifiable model, $p_{ij} = \alpha_i + \beta_j + c$, where $\alpha_{\min} = \beta_{\min} = 0$, $\alpha_i \in [0, \frac{1}{2}]$, $\beta_j \in [0, \frac{1}{2}]$, and $c \in [0, 1 - \alpha_{\max} - \beta_{\max}]$.

[2] Clearly, the data sets depend on the distribution from which one samples. In this case we sampled from Jeffreys' distribution (see Robert 2001, for instance), corresponding to the assumption of maximum uncertainty about the data. However, Jeffreys' distribution is difficult to sample from in many situations, and one may wish to specify a different distribution. Another principled choice is the uniform distribution, which corresponds to maximum uncertainty about the parameters, and is trivial to sample from.

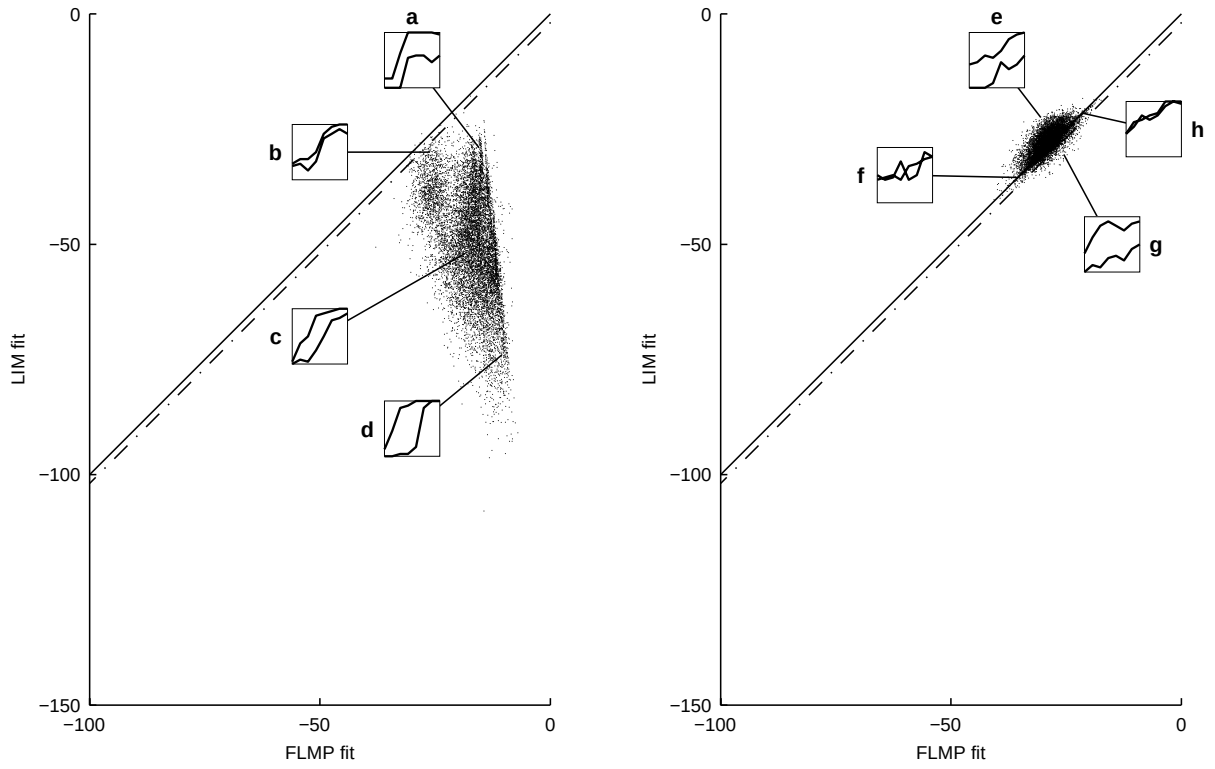

Figure 1: Scatterplots of the (logarithm of) maximum likelihood estimates for 10,000 data sets generated by FLMP (left panel) and LIM (right panel). Values on the x-axis denote the fit of FLMP to the data, and y-values denote LIM fit. Data come from a 2×8 experimental design with $N = 24$. The inset panels (a) through (h) display typical data sets sampled from different regions. The solid line depicts the ML decision threshold, and the broken line is the MDL threshold.

effects of sampling error and differences between the models, this is not always the case.

The Lay of the Land

Figure 1 displays the 10,000 data sets generated by FLMP (left panel) and LIM (right panel), plotted as a function of the maximum likelihood estimate for each model. The solid line marks the decision boundary for the maximum likelihood (ML) criterion: LIM provides a better fit to all data sets that lie above the solid line, whereas FLMP provides the better fit to the points below. The inset panels display p_{ij} values (on the y-axis) as a function of the 8 levels of j (on the x-axis). The two lines correspond to the two levels of i (the upper line represents i_2).

These plots are remarkably informative because the the relative performance (i.e., fits) of the two models across then entire range of data patterns for each model is visible. Inspection of the left panel reveals that data generated by FLMP are almost always better fit by FLMP than by LIM. In fact, there are only 6 points (of 10,000) above the solid line. In other words, if we used ML as a method to guess which model generated the data, we would only be incorrect in a tiny proportion of cases. Not only that, the vast majority of FLMP patterns are nowhere near the solid line, indicating that in most cases, the decision is clear-cut: FLMP provides the better fit. Interestingly, we note that the scatterplot tapers in the lower righthand area: when the LIM fit is exceptionally poor, the FLMP fit is especially good. In short, FLMP will almost never be confused for LIM.

On another note, it appears that the variability in this scatterplot is interpretable in terms of human performance. Even a cursory examination of the types of response patterns that fall in different regions of the FLMP landscape is informative. Inset (d) in Figure 1 shows a sample data set drawn from the lower tail of FLMP distribution, which displays a pattern that is typical of those observed in such experiments. The sigmoidal curves in (c), and to some extent (b), are not atypical of human data, though the step-function curves in (a) are not characteristic of human performance.

The right panel tells a different tale, in which LIM data sets tend to cluster in a tight region near the

decision boundary. In fact, a total of 3,130 of the 10,000 data sets fall on the wrong side of it, meaning that FLMP can fit LIM data better than LIM itself almost one third of the time. Worse yet, the LIM data sets cluster in a direction parallel to the decision boundaries. This means that when LIM fits the data well, so does FLMP, even though the data sets came from LIM. In fact, since the data sets rarely fall far away from the solid line, it is clear that FLMP is capable of mimicking LIM all the way across LIM's parameter space. The models are highly confusable.

A cursory sweep through the LIM data is consistent with these findings. Since LIM is such a simple model, most response patterns look alike (parallel lines), and the major source of variation is sampling error. When both models fit well, as in inset (h), the data tend to look like parallel lines. Both models fit poorly when the noise heavily distorts the response pattern as in (f). Occasionally, as in (e), sampling error is more damaging to the FLMP fit than the LIM fit. On the other hand, as in (g), it sometimes allows FLMP to fit better than LIM.

The major implication of the landscape analysis is this: if FLMP is the correct model, then it should be easy to perform 2×8 design experiments that support it over LIM. However, if LIM is the correct model, such an experiment will not be very effective in distinguishing it from FLMP, and another test will need to be devised in order to do so. Looked at another way, the inability to distinguish between two models is an issue of power, of determining how effective an experiment can possibly be. This insight makes it clear that power is asymmetric in the current experiment: It is easy to distinguish FLMP from LIM, but difficult to distinguish LIM from FLMP.

Remedying the Asymmetry

There are at least three ways of increasing the power of the experiment to overcome the asymmetry and make the comparison a more balanced test of the two models. The standard remedy is also the simplest: increase the sample size. By increasing the sample size, we decrease the amount of sampling error, and should therefore be better able to discriminate between the two. However, this approach can suffer from pragmatic and theoretical difficulties. The pragmatic problem is that it may not be feasible to increase the sample size, as in clinical studies for instance, where one may have limited opportunities to collect data. From a theoretical point of view, it is possible that increasing the sample size will yield limited returns. If FLMP can produce response patterns that look LIM-like even without sampling error (i.e. as $N \to \infty$), then reducing the error may not help.

The second solution is to use more powerful statistics. Athough ML is useful for measuring fits to data, it is outperformed in small samples by a great many other statistics. One of the more accurate of these is Rissanen's (1996) Minimum Description Length (MDL) criterion, which has recently been employed with some success in psychological modeling (e.g., Lee 2002; Navarro & Lee 2002), and is more effective at discriminating between FLMP and LIM (Pitt, Kim & Myung, in press). ML and MDL differ only by a constant "geometric complexity" (GC) term (Pitt, Myung & Zhang 2002):

$$\text{MDL} = -\ln(\text{ML}) + \text{GC}.$$

In this case, the geometric complexity of FLMP is greater than that of LIM by only 1.9, which can seem like a small difference in view of the variability in Figure 1. Nevertheless, when the MDL criterion is applied (shown by the broken lines) instead of ML, the asymmetry greatly diminishes. By introducing a complexity penalty, MDL makes a few more misclassifications for FLMP data, incorrectly choosing LIM 28 times out of 10,000. However, the major difference occurs for the LIM data, in which the error rate falls ten-fold, from 3,130 to 356 out of 10,000. Because the LIM data sets cluster in such a tight region in the scatterplot, this small correction produces a massive improvement in classification: the overall error rate across the 20,000 data sets drops from 15.7% to 2.3%. On the basis of these results, it is tempting to simply recommend the use of MDL over ML, since it is the better statistic in general (see Grünwald 2000). However, the GC term in the MDL criterion can be very difficult to evaluate even for simple models due to an often-intractable integral term. For nonlinear models with many parameters, it can be nearly impossible.

The third remedy relies on Lord Rutherford's assertion to the effect that "if your experiment needs statistics, you should have done a better experiment". It might be that, with only minor alterations, we could perform an experiment that would be more likely to distinguish between the models without requiring elaborate statistical inference or enormous sample sizes. Of course, inventing new experimental designs requires the kind of insight on behalf of experimenters for which no methodology can substitute. On the other hand, once we have thought of a new experimental setup, it is simple enough to use landscaping to see if the new design is likely to be more sucessful than the first. It is this possibility that we now examine.

Experiment Two

One of the difficulties with the original 2×8 design is that the experiment does not directly measure θ_i and λ_j. In other words, it does not ask how would people respond if only one source of evidence (auditory or visual) were provided. FLMP and LIM make different predictions in this regard: LIM predicts $p_i = \theta_i$, whereas FLMP predicts that

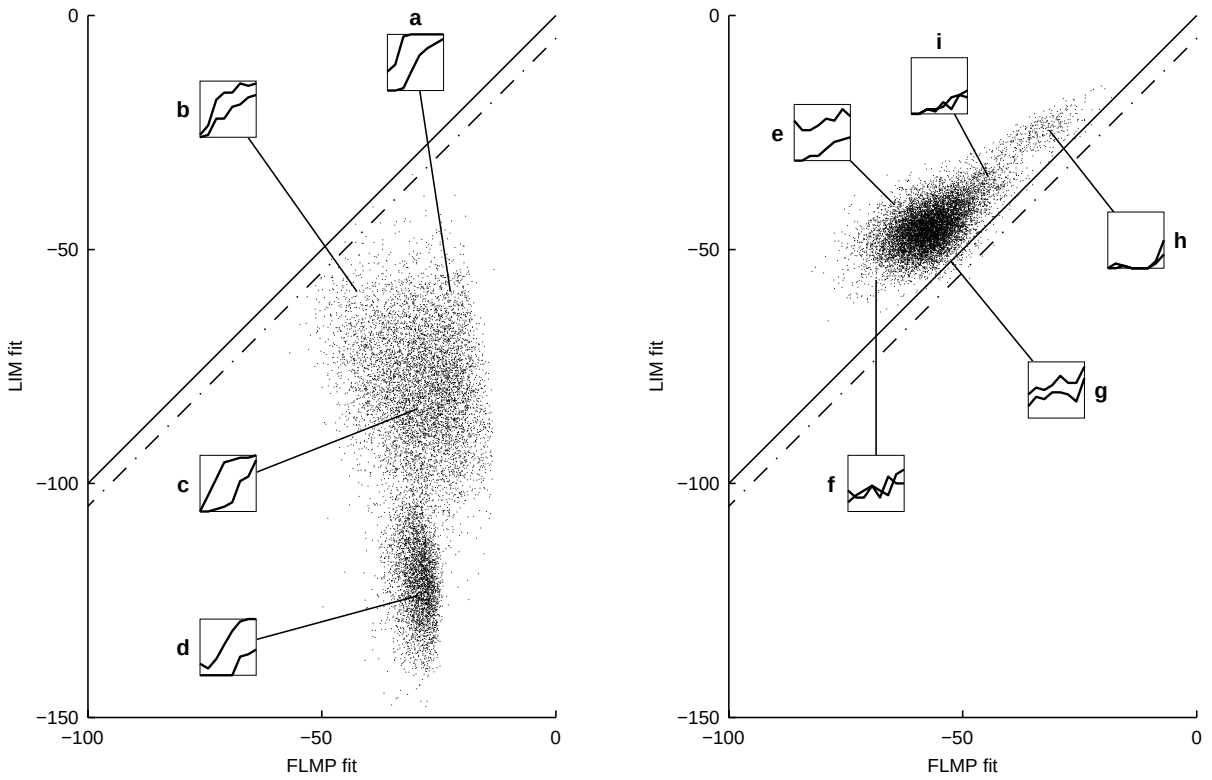

Figure 2: Scatterplots of FLMP fit versus LIM fit (again on a logarithmic scale) for the expanded 2×8 experimental design with unimodal conditions and $N = 24$. As before, the solid line represents equal fit, the broken line represents equal MDL values, and the inset panels display sample data sets.

$p_i = \theta_i/(1-\theta_i)$. This suggests an elegant alteration to the original design, by including the 10 extra "unimodal" stimuli as additional conditions in the design. This redesign accomplishes three objectives. Firstly, the non-identifiability problem that we alluded to in footnote 1 vanishes. Secondly, the recovered parameter values are more easily interpreted as measurements of the evidence provided by each source. Thirdly, the power of the experiment is increased, as we show below.

Performing the same landscaping analysis with 10,000 data sets on our new 2×8 (plus unimodal) design shows the effect of adding these conditions, displayed in Figure 2. Given that the shapes of the scatterplots are quite similar to those in Figure 1, it seems likely that there are no substantial qualitative differences between the models across experiments. Rather, the change in design has constrained their behavior (i.e., data-fitting ability) to regions in the landscape in which LIM is distinguishable from FLMP and vice versa.

As before, the FLMP data sets are generally quite distant from the decision thresholds, and both ML and MDL are very successful in selecting FLMP as the correct model: ML makes no misclassifications at all, whereas MDL makes only 18 errors. An informal scan across the scatterplot supports our intuition that the qualitative features of FLMP are unchanged. The lower tail of the FLMP distribution still contains the classic FLMP-like data sets, illustrated in panels (c) and (d), whereas the patterns closest to the decision boundaries, as in (b), are much more linear. It is not clear, however, if the pattern in (a) represents a difference from its counterpart in Figure 1.

Inspection of the panel on the right reveals that the expanded design has allowed the LIM data to distinguish itself from FLMP. Although the distribution of data sets is still parallel to the decision thresholds, indicating that FLMP can still mimic LIM, they are shifted away from the decision criteria, indicating that the extent of the mimickability has been substantially reduced. In this experimental design, ML makes far fewer mistakes, only 100, and MDL makes only a single misclassificiation. That is, by adopting this expanded design, the ML error rate drops from 15.7% to 0.5%, and the MDL error rate drops from 2.3% to 0.1%. Again, a brief survey of the landscape shows that the patterns illustrated by (e), (f) and (g) match their counterparts in Fig-

ure 1. However, since LIM has developed a long tail, we display two patterns (h) and (i), both of which display a resemblance to panel (h) in Figure 1.

General Discussion

Landscaping is a simple and powerful tool for model evaluation and comparison. It is a method for viewing the relationship between two models across the space of all possible data patterns that the models can generate within a particular experimental design. FLMP encompasses a larger range of data sets than LIM, a range that includes patterns produced by participants. Furthermore, these patterns fall in the main body of the FLMP scatterplot, and appear to be as representative of FLMP as they are unrepresentative of LIM.

Secondly, by plotting the LIM data sets, we became aware of the potential difficulty of distinguishing between FLMP and LIM. To do so required the use of very powerful statistical methods (MDL) or an expanded experimental design.

Thirdly, despite the change in experimental design, the shape and composition of the model landscapes seem to be more or less invariant. We speculate that, although data fits and model distinguishability vary substantially across experimental designs, the qualitative flavor of the landscape is constant.

Among the strengths of landscaping is its adaptability. It is a tool that can and should be modified to suit the circumstances. For instance, in this paper we sampled parameter values (with some pain) from a Jeffreys' distribution. In many cases a simple uniform distribution is appropriate, particularly if the parameters are assumed to correspond to real psychological variables. Similarly, few modeling situations require 10,000 data sets. If the aim is only to estimate the power of an experimental design, a few hundred would likely suffice, since the fine detail of the landscape is irrelevant. If we are interested in looking at the types of response patterns predicted by the models (rather than the data sets that we would expect to observe), there is no need to add sampling error.

In general, we suspect that landscaping analyses on the scale that we have undertaken here will rarely be required, and smaller, simpler evaluations will suffice. Even a little landscaping may go a long way. If model indistinguishability is unavoidable, we are alerted to the necessity of tools such as MDL. On the other hand, high distinguishability suggests that smaller samples, simpler designs, or more convenient statistics will be adequate.

If local analyses such as maximum likelihood show us the tip of the iceberg, then global methods such as landscaping allow us to look beneath the surface to the model below. We hope that in doing so, global methods may actually simplify the work required to distinguish models. Landscaping allows one to quickly "sketch out" all the possible outcomes of an experiment that we are thinking about conducting. Should it reveal a problem such as indistinguishable models, landscaping can be used to estimate the effectiveness of a proposed solution to increase the power of the experiment. Every remedy requires something extra to be added, be it statistical machinery, participants, or experimental conditions. Perhaps this is unavoidable. Even so, while there may be no free lunches in model evaluation and testing, we can often choose a preferred method of payment.

Acknowledgments

This research was supported by research grant R01 MH57472 from the National Institute of Mental Health. We thank Nancy Briggs, Michael Lee and Yong Su for many helpful comments and discussions, and an anonymous reviewer for suggesting some interesting avenues for further development.

References

Anderson, N. H. (1981). *Foundations of Information Integration Theory.* New York: Academic Press

Crowther, C. S., Batchelder, W. H., and Hu, X. (1995). A measurement-theoretic analysis of the fuzzy logic model of perception. *Psychological Review. 1995, 102*, 396-408.

Estes, W. K. (2002). Traps in the route to models of memory and decision. *Psychonomic Bulletin & Review, 9*, 3-25.

Grünwald, P. (2000). Model selection based on minimum description length. *Journal of Mathematical Psychology, 44*, 133-152.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions.* Chicago: University of Chicago Press.

Lee, M. D. (2002). Generating additive clustering models with limited stochastic complexity. *Journal of Classification, 19*(1), 69-85.

Navarro, D. J. & Lee, M. D. (2002). Commonalities and distinctions in featural stimulus representations. In: W. G. Gray, and C. D. Schunn (Eds.) *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, pp. 685-690, Mahwah, NJ: Lawrence Erlbaum.

Oden, G. C., & Massaro, D. W. (1978). Integration of Featural Information in Speech Perception. *Psychological Review, 85,* 172-191.

Pitt, M. A., Kim, W. and Myung, I. J. (in press). Flexibility versus generalizability in model selection. *Psychonomic Bulletin and Review.*

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

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory, 42*, 40-47.

Robert, C. P. (2001). *The Bayesian Choice* (2nd ed.). New York: Springer.

Rubin, D. C. & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review, 103*, 734-760.