Navarro, D. J., Pitt, M. A. and Myung, I. J. (2004). Assessing the distinguishability of models and the informativeness of data. Cognitive Psychology, 49, 47-84.
http://dx.doi.org/10.1016/j.cogpsych.2003.11.001

Assessing the distinguishability of models and the informativeness of data

Danielle J. Navarro, Mark A. Pitt, and In Jae Myung
Department of Psychology
Ohio State University

Abstract

A difficulty in the development and testing of psychological models is that they are typically evaluated solely on their ability to fit experimental data, with little consideration given to their ability to fit other possible data patterns. By examining how well model A fits data generated by model B, and vice versa (a technique that we call *landscaping*), much safer inferences can be made about the meaning of a model's fit to data. We demonstrate the landscaping technique using four models of retention and 77 historical data sets, and show how the method can be used to (1) evaluate the distinguishability of models, (2) evaluate the informativeness of data in distinguishing between models, and (3) suggest new ways to distinguish between models. The generality of the method is demonstrated in two other research areas (information integration and categorization), and its relationship to the important notion of model complexity is discussed.

Keywords: Model distinguishability, retention models, landscaping, data informativeness

The development and testing of theories is one of the most important aspects of scientific inquiry. As psychology has become increasingly reliant on quantitatively instantiated theories, it has become possible to directly test models against data. The precision inherent in these models affords the opportunity to study their inner workings. Ideally, this precision should lead to very explicit tests, and data that are clearly explained best by one model. Nevertheless, just the opposite occurs all too often, making it harder and harder to discriminate between competing models because they provide similarly good fits to data and equally plausible explanations of the phenomenon being modeled. This is true of models of categorization (e.g., Minda & J. Smith, 2002), and reaction time (Ratcliff & P. Smith, in press), for instance.

In this paper, we introduce a new method of analyzing data that is intended to provide insight into the causes of this congestion and possible ways to alleviate it. Called *landscaping*, it provides a visual and quantitative description of the distinguishability of statistical models. Instead of comparing models on their ability to fit a single data set, landscaping steps back from a single data set and takes a birds-eye view of a pair of models. From this vantage point, one can easily assess their potential distinguishability and the informativeness of a data set in deciding between them.

1. Retention Data and Models

We consider retention functions in order to illustrate the methodology, since there is a large literature with many data sets, as well as models that are relatively simple and highly competetive. Furthermore, interest in the form of the retention function has resurfaced in recent years (Rubin, Hinton, & Wenzel, 1999; Sikström, 2002; Wixted & Ebbesson, 1991). We begin by summarizing recent work comparing models of retention.

1.1 Four Models of Retention

Rubin and Wenzel (1996) attempted to provide some clarity on the form of the retention function by performing a data-fitting meta-analysis. Some 105 two-parameter functions were fit to 210 data sets that were collected throughout the 20th century. The inconclusiveness of this review prompted Rubin, Hinton, and Wenzel (1999) to collect several data that would yield a decisive answer. Each included 100 participants, measured retention at many time intervals, and was calibrated so that responses would span the full range of the dependent measure (proportion correct, which falls between 0 and 1). Our choice of models was largely based on these papers.

We first chose the power-exponent (PE) model, based on Wickelgren's (1972) "strength-resistance" theory, and used the full 3-parameter version, where the power is treated as a free parameter, given by $y = a \exp(-b\ t^c)$, where a and c lie between 0 and 1, $b \geq 0$, and t represents the elapsed time since stimulus representation.Given the success of a (slightly different) hyperbolic model in Rubin and Wenzel's review, we also included a hyperbolic model (HY), $y = a/(a + t^b)$ that corresponds to the assumption that the *odds* of retention decline according to a power law (J. Anderson & Schooler 1991; Wixted & Ebbeson 1997). Additionally, we included the exponential model (EX) $y = a \exp(-b\ t)$ due to its historical interest. For both the EX and HY models, a must lie between 0 and 1, and b must be non-negative. Finally, we included Rubin et al.'s sum of exponentials (SE) model, $y = a_1 \exp(-b_1\ t) + a_2 \exp(-b_2\ t) + a_3$. All parameters are non-negative, and (for identifiability reasons) $b_1 \geq b_2$. The three main parameters in the model, a_1 a_2 and a_3, are interpreted as three mutually exclusive memory stores, the first two of which are most important. The faster-decaying store, a_1, corresponds to something like working memory and a_2 corresponds to a longer-term memory store, while a_3 represents very long-term residual storage.

It is worth noting that in Rubin et al.'s experiments y is an estimate of the probability of correct recall, $p(C)$, which of course cannot exceed 1 at any time, nor can it drop below 0, irrespective of whether retention was actually measured at that time. Notice that when t=0, the SE model reduces to $a_1+a_2+a_3$. Unless this sum is 1 or less, the SE model violates this constraint, and the value of y makes no sense, because it predicts that more items were

correctly recalled than were presented. Rubin et al. did not restrict the parameter range to prevent this from happening, which can (and does) result in inflated fits to the data. Accordingly we distinguish between two versions of the SE model, the unbounded SE model (SE-U) used by Rubin et al., and the bounded SE model (SE-B), in which the constraint $0 \leq a_1 + a_2 + a_3 \leq 1$ is added. It is this latter version that we consider in this paper.

1.2 Methodological Issues

In this section, we discuss three important methodological issues. The first issue regards the dependent measure. While the two most commonly used measures in retention are *p(C)* and *d'*, we restrict our discussion to studies that used *p(C)*. The main reason for this is that the same retention function yields different statistical models when the dependent measure is *d'* than when it is *p(C)*, so the data sets are not directly comparable to one another. In this initial investigation we chose to use the *p(C)* data sets due to their abundance.

The second issue is how one should calculate the goodness of fit. The commonly-used r^2 measure assumes that error variation in the observed data is normally distributed, which is often reasonable, but can be a problem for a bounded discrete measure like *p(C)*. At extreme values (e.g., $p(C) > .97$), the error distribution for *p(C)* is highly skewed, making a normal distribution inappropriate and leading to distorted values of r^2. A better method in the current situation is maximum likelihood[1]. Instead of minimizing the distance between the observed data points and the model's predictions, as in r^2, it seeks to make the data seem as unremarkable as possible. This approach uses the model to assign a probability to the observed data, and finding the parameter values that maximize this probability. The value of this maximized probability is called the maximum likelihood (ML; see Myung 2003) The ML method permits the specification of an appropriate error distribution, thereby avoiding the skewness problem. For models that predict that *p(C)* data arise as an average proportion of success across a series of independent (Bernoulli) trials, the desired error distribution is the binomial.

The third concern pertains to the data sets themselves. A total of 77 data sets from 16 studies were used in the landscaping analyses. With the exception of five data sets from Rubin et al. (1999), they are a subset of those used by Rubin and Wenzel's (1996) review (Some of the methodological details from the studies are presented in Appendix A.). Studies were selected that met three criteria, which were intended to reduce the heterogeneity of the database. (1) The experiments must not have tested autobiographical memory. (2) Humans must have been the participants. (3) The dependent measure was *p(C)*, which allows the use of ML with binomial error. While these criteria provide some degree of "quality control" for the data, there are two important issues that they do not address. Firstly, they do not ensure that the data were gathered in a methodologically rigorous fashion. Thus it is possible that many data sets are "contaminated" by systematic error. In addition, in most of the experiments, data were averaged across participants, which can distort the underlying structure in substantial ways (e.g. Brown & Heathcote, 2003; Myung, C. Kim & Pitt, 2000). In this paper, we take the data sets at face value.

1.3 Fitting the Models to Data

As a first examination of these models, Table 1 displays their ln(ML) fits to three recall and two recognition data sets (the old-new and the recognition + know data sets[2]) from the Rubin et al. (1999) study. The SE model provides the best fit, although the HY model fits are comparable for the recognition experiments. What is most interesting, however, is that the fits to the recall data are substantially worse than the fits to recognition data.

Insert Table 1 about here.

Next we fit all four models to the remaining 72 data sets, shown in the bottom row of Table 1. Again, the SE model provides the best mean fit, but just as Rubin and Wenzel found, none of the models emerges as the undisputed winner. The differences between -31, -36, and -38 on a log-odds scale are

moderate to strong, but less than convincing in light of the difference in the number of parameters among models[3]. In short, this analysis does not obviously discriminate between SE, PE, or HY. However, it does suggest that the EX model, with a substantially inferior average fit of -60, is distinctly less impressive than the other three. This large-scale data-ftting exercise is not particularly helpful in choosing between candidate retention functions. In the sections that follow we shed some light on the causes of this impasse and what can be done about it.

2. Introduction to Landscaping

This section introduces the basic ideas underlying landscaping. However, since the technique is partly motivated by the limitations of data-fitting analyses of the kind presented in the previous section, we discuss these first.

2.1 Data-Fitting: A Local Model Analysis

A common method of choosing between models is to assess how well they approximate experimental data, whether by fitting the model to the data or by simulating the phenomenon (common in connectionist modeling). This approach is informative because the mental process being modeled is reflected through the data. However, as the previous discussion shows, even large amounts of data can still fail to distinguish between models, *especially* if the data are not entirely reliable. We refer to this type of model selection as local model analysis (LMA) because of its emphasis on analyzing (local) fits to data.

There are at least three substantial difficulties with LMA. Firstly, psychological data tend to contain lots of sampling error (i.e., they are noisy), which can obscure potentially informative trends, or worse, provide genuinely misleading information about the underlying cognitive process being studied. For instance, when modeling the similarity between stimuli, it is commonplace to use the statistical technique known as multidimensional scaling. However, it has recently become clear that noisy, averaged data can distort the scaling solution in some ways (Ashby, Maddox & W. Lee, 1994; also see M. Lee &

Pope 2003). This kind of problem is particularly pronounced for nonlinear models (e.g., Brown & Heathcote, 2003; Estes 1956; Myung, C. Kim & Pitt 2000), which are increasingly common in psychology. LMA does not deal with this kind of problem.

Secondly, some experimental designs may not elicit very informative data in the first place. For example, too few variables may be measured, or a poor task might be inadvertently chosen. When trying to distinguish between retention functions, it is virtually pointless to have only three or four retention intervals, since any reasonable function will provide an excellent account of the data (Rubin & Wenzel 1996).

Thirdly, similar models can be difficult to tell apart on the basis of fit alone, even when fit to very good data, as illustrated in Table 1. Perhaps ironically, this phenomenon can be a natural consequence of good science: Competing models will become increasingly alike if they are able to provide good fits to an ever-expanding pool of data sets. This trend is evident for many types of models, including both connectionist models (e.g. the TRACE and MERGE models of speech perception; see McClelland & Elman, 1986; Norris, McQueen & Cutler 2000) and algebraic statistical models (e.g., the FLMP and LIM models of information integration; Oden & Massaro, 1978; N. Anderson, 1981; Navarro, Myung, Pitt & W. Kim, in press). When presented with two good models, the best that we can say using LMA is that both models fit the data reasonably well.

2.2 Landscaping: A Global Model Analysis

Global model analysis (GMA), in contrast, steps back from a particular data sample and focuses on the behavior of the model as a whole. One form of GMA that we have studied in prior work is model complexity (Myung, Balasubramanian & Pitt, 2000; Myung & Pitt, 1997; Pitt, Myung & Zhang, 2002). It is concerned with assessing the inherent flexibility of a model in fitting data. Not only can this property of model be quantified, but it can be integrated with a goodness-of-fit measure (ML) to improve model selection. We will briefly discuss model complexity later in the paper.

Here we introduce a complementary GMA method called *landscaping*, an early version of which was used by Pitt, W. Kim and Myung (2003). Landscaping fulfils the above desire to add more meaning to fits to data by increasing our understanding of the models and data. A related resampling technique that uses the parametric bootstrap has also been proposed by Wagenmakers, Ratcliff, Gomez and Iverson (in press), though their focus is more on local analyses of model mimicry. Similar procedures are discussed in a more explicitly Bayesian framework by Geweke (1999a, 1999b). Wagenmakers et al. provide a nice discussion of the relationship between these techniques.

At its simplest, landscaping is a graphical depiction of the relationship between two models and experimental data. It is created by fitting one model to many (e.g., 1000) data sets that were generated by that model using a particular experimental design (the data generation method is discussed in detail later). Another model is fitted to those same data sets, and the fits of the two models are plotted against one another. The scatterplot formed by these points is referred to as a landscape.

Figure 1 illustrates this schematically for two hypothetical models A and B. Data are generated from model A, and these data are fit by both models. The x axis in the figure denotes model A's fit, and the y axis denotes model B's fit. Each of the dots represents one data set. By drawing a diagonal line across the middle of the plot (at x=y), we observe that points above the line correspond to data sets that model B fits better than model A, whereas the opposite is true of points below the line. This line is referred to as a *criterion line*, or *decision threshold*. Data that both models fit very well will fall in the top right corner, whereas data that both models fit poorly will fall in the bottom left corner. By plotting a large number of data sets, we obtain a landscape of relative fits that enable us to see how closely model B can mimic model A. It would be nice if model A always provided better fits to its own data, but in practice this is not always true. When comparing fits of models A and B in a landscape plot, we denote the comparison B/A when A generates the data, and A/B when B generates the data.

Insert Figure 1 about here.

Because the co-ordinates correspond to log-likelihoods, they combine additively (see Murray & Rice 1993, p. 9-11 for a principled discussion). Consequently, the natural way to measure the distance between two points (x_1, y_1) and (x_2, y_2) in a landscape is to use the City Block distance $|x_1\text{-}x_2| + |y_1\text{-}y_2|$. In this way, the plots preserve the underlying statistical relationships between models and data. One useful by-product of this correspondence is that measuring the distance of a dot to the criterion line produces (the magnitude of) the log-odds value $\ln(ML_A)\text{-}\ln(ML_B)$. Furthermore, we can measure how much better on average one model fits the data than another, by taking the arithmetic mean of all log-odds values. Visually, this corresponds to measuring the perpendicular distance from the landscape's centroid to the criterion line.

3. Application of Landscaping to Retention Models

In this section, we apply this methodology to the EX, HY, PE, and SE models. As discussed earlier, our database included 25 unique experimental designs. We first discuss the procedure, and then use it to assess the inherent distinguishability of the models. Finally, we discuss to evaluate empirical data in a landscape.

3.1 Performing the Landscaping Analysis

Landscaping involves generating data from one model (HY, for instance) and then fitting several different models (in this case, EX, HY, PE and SE) to those data. Generating data from a model is a three step process: Randomly choose some values for the model's parameters (a and b), evaluate the model's predictions (i.e., find y), and add noise to the data. The first of these steps is the most difficult. In the real world, it is rarely if ever known in advance which parameters are most likely to be good ones. This is, after all, the very reason for the existence of free parameters. When comparing two models it is crucial to acknowledge this uncertainty. One way to do this is to specify a probability distribution over parameter values, and then sampling the parameter values

from this distribution. We used Jeffreys' (1961) distribution for this purpose (see Appendix B), for two reasons. The first is that it is *reparametrization-invariant*. In other words, if we re-write the same model using different equations, nothing changes. For instance, the functions $y = a \exp(-bt)$ and $y = r^{s-t}$ actually describe the same model, since they are related through the reparametrization $r = \exp(-b)$, $s = -(1/b) \log a$. Unfortunately, the uniform distribution on (a, b) does not correspond to a uniform distribution on (r, s), and vice versa. However, it turns out that the Jeffreys' distribution on (a, b) *does* correspond to the Jeffreys' distribution on (r, s), as discussed by Gill (2002, p. 123-125). Thus the resulting landscape does not depend on the way that the equations are written: Rather, it is an inherent property of the model.

The second reason is that Jeffreys' distribution assigns equal likelihood to every (distinguishable) distribution indexed by the model (Balasubramanian 1997), which make it a kind of "noninformative" distribution within the parameter constraints (in this case, parameter constraints ensured that the function is always decreasing, for instance). Although it may be convenient to use more tractable distributions in other applications, we have not examined the consequences of doing so. Also, in many situations a great deal of prior information about the parameters is available. In these cases it may be more useful to choose a distribution based on this information rather than use the uninformed formulation that we have adopted here.

Once a set of random parameter values (e.g., a=.1, b=2) has been chosen, it is straightforward to find the model predictions (y), by substituting these values into the model equations. This allows us to move onto the last stage, adding noise to the data. Since the error distribution for y is binomial, this is trivial. The conceptually simplest method is to simulate N hypothetical trials, where N denotes the number of trials in the experimental design. This is done by generating N uniformly distributed random numbers, and counting the number that are less than or equal to y. This count corresponds to the number of correct responses for the current time interval. Once this is done for each different time interval, a data set has been sampled from the model.

When the data have been generated, the next step is to fit the models (EX, HY, PE and SE) in order to find ln(ML) values. This process is no

different from fitting empirical data, and can be done by using standard numerical methods. In all of the current analyses, we used a combination of trust-region and Levenberg-Marquardt methods (see Nocedal and Wright 1999), repeated 10 times for each data set using different initial conditions to avoid suboptimal results. The output of this procedure is a set of four ln(ML) values for each data set, one for each model. After repeating the data-generation and data-fitting steps 1000 times to ensure that a "sufficient" range of the parameters is sampled, the fits to any pair of models (EX and HY) can be combined to produce a plot like Figure 1.

An important point to make regarding retention landscapes is that they are sensitive to those aspects of the experimental design that affect the models themselves. The time interval t appears in all of the equations describing retention functions, which means that changing the retention intervals changes the landscapes. This is also true of N, because ML is sensitive to sample size through the specification of the error distribution. However, the landscape will not be affected by changes in other design variables (e.g., stimuli, task), even though these can affect empirical data. In our collection of 77 data sets from 16 studies, there are 25 unique combinations of N and t, which meant that there were 25 conditions in which to compare the four models.

3.2 An Illustrative Case: The Burtt & Dobell (1925) Design

Figure 2 displays the matrix of landscapes for the experimental design of Burtt and Dobell (1925), and is fairly typical of the 25 designs. Data-generating models (x axis) form the rows, and competing models (y axis) form the columns. One striking aspect of the figure is the variation in the size of the landscapes. They vary mostly in their length, and provide an indication of the relative distinguishability of pairs of models. By looking across rows, the distinguishability of the data-generating model from its competitors can be seen. While EX and HY are distinguishable from each other, neither is easy to distinguish from PE or SE. PE and SE, on the other hand, are more consistently distinguishable from their competitors.

Insert Figure 2 about here.

The overall distinguishability of two models, say HY and SE, must be assessed by inspecting the landscapes generated by both models, SE/HY and HY/SE. It is clear that if HY describes the retention function, then it will be difficult to distinguish it from SE, since SE can fit most HY data fairly well (mean log odds = 4.5). Alternatively, if SE is a better description of retention, the models *are* distinguishable: There are data patterns that SE can fit well but HY cannot (mean log odds = 23.53). Of course, this only implies that they *can* be distinguished, not that they will be. In order for models to be distinguished, experimental data need to fall "in the right spot" in the landscapes.This is not guaranteed to happen in an experiment, an issue we discuss in depth later.

Two of the graphs in the upper right of the matrix look as though their landscapes are missing. The distribution of points is highly peaked and straddles the criterion line, indicating that the competing models (PE and SE) fit the EX data as well as, and sometimes better than, the data-generating model. This is as it should be because EX is nested within SE and PE. Both of these models can fit any data set generated by EX. The reverse is not true, which is why the EX/SE and EX/PE landscapes in the lower left corner are so elongated. However, comparison of these two landscapes shows that the two models are not equally discriminable from EX: Mean log odds for the EX/PE landscape is 2.21, compared with 17.32 for EX/SE. (The reason for this difference is not obvious from Figure 2, because it does not convey an impression of how dense the points are at different locations. The EX/PE landscape is in fact densely clustered near the criterion line, whereas the EX/SE landscape is more diffuse: The representativeness plots introduced later help rectify this problem).

3.3 Distinguishability as a Function of Design

Changing the number of time intervals $|t|$ or the sample size N can change the landscape, and thus the distinguishability of the models. The

experiments in our database vary enough along both dimensions to assess the effect of each fairly independently of the other. The top row of graphs in Figure 3 shows the effects of N on model distinguishability when $|t|$=6. The HY/SE landscape in the left graph is from a recall study by Krueger (1929; N = 280), and the one on the right is from a cued recall study by Runquist (1983; N = 1728). The effect of increasing N is dramatic. The distribution of points in the low-N graph is highly peaked and centered very close to the criterion (mean log odds = 3.94), providing little opportunity for finding discriminating data. The high-N distribution is much more spread out, indicating that HY fits many of these SE data sets poorly (mean log odds = 94.3).

Insert Figure 3 about here.

Variation in $|t|$ has a similar though smaller effect than N, probably due in part to its smaller range (5-15). The left landscape on the middle row of Figure 3 is from Wickelgren (1968; $|t|$=5), and the one on the right is from Strong (1913; $|t|$=13). Both are recognition studies with N =40. The models are essentially indistinguishable when $|t|$ is small (mean log odds= 0.71), but HY's fits worsens when $|t|$ is large, making the models a little more discriminable (mean log odds=8.56). Comparison of the relative location of the landscapes in the graphs shows that the fits of both models decrease as $|t|$ increases, a change that is much less evident with variation in N; note how the landscape slides down the criterion line from the left to the right graph. The bottom graph shows the landscape with the largest N and $|t|$ in our database (Squire, 1989). It provides what might be considered a likely upper bound on the distinguishability of the HY and SE models.

Although these examples illustrate that N and $|t|$ can affect model distinguishability for the four models under consideration, the effects of N are not only much more potent, but also more predictable. When the mean log odds of each of the 25 experimental designs (collapsed over the 12 model pairings in a design) were correlated with each variable, the relationship was strong for N (r = 0.85) but weak for $|t|$ (r = -0.09; N and $|t|$ are weakly

correlated, r = -0.04). We suspect that this is caused by the different effects the two variables have on data-fitting. As N increases, the models must fit the data more accurately because error variance is so small. This is also true for $|t|$, but unless N is also large, there will be enough uncertainty (i.e., error variance) in *p(C)* at each time value to render the models indistinguishable.

The top panel of Figure 4 summarizes findings from all 25 experimental designs, rank ordered by their mean log odds. The symbols represent the log odds for each of the 12 model pairings for each design. The legend for the enumerated designs on the x axis is in Appendix A, though it is not needed here. It is unlikely that any of the four models could be easily distinguished with the first twelve designs (up to #22). Not one log odds value exceeds 10. The next five designs (up to #19) offer a bit more hope, but at best only for one or two model pairs. These 17 designs are shown in the upper panel of Figure 4. Of the 8 most informative designs (bottom panel), it is noteworthy that the only one with $N < 400$ has 13 time intervals (#18), and that the three most distinguishable designs all have $N >$1000. These are the designs of Burtt and Dobell (1925), Rubin et al. (1999), and Runquist (1983).

Insert Figure 4 about here

3.4 Distinguishability as a Function of Models

The relative distinguishability of the model pairs can be seen by comparing the symbols across designs. When the data were generated by EX, the mean log-odds is always near zero when compared to PE (grey circles) or SE (black circles), just as one would expect for nested models. Interestingly, while the EX/SE comparison (white squares) generally displays one of the highest mean log-odds, the EX/PE comparison (white diamonds) rarely fares so well. Together, these findings indicate that, while EX is a submodel of both PE and SE, the "extra" PE patterns tend to look a lot more like EX patterns than the extra SE patterns. Furthermore, since the black diamonds (SE/PE) are generally below the black squares (PE/SE), it appears that SE mimics PE

better than the reverse. Nevertheless, both comparisons tend to suggest poor distinguishability across all designs.

Turning to the HY model, it is apparent that neither the HY/SE (grey square) nor SE/HY (black triangle) comparisons figure highly in the mean log-odds plots, with the implication that, although these models are clearly distinct, they mimic each other fairly
well. Nevertheless, SE mimics HY better than vice versa. In contrast, the HY/EX (white circles) and EX/HY (white triangles) comparisons are much higher up in the plots, suggesting greater distinguishability. Interestingly, the white circles (HY/EX) are generally higher than the grey squares (HY/SE), implying that HY can mimic SE much better than it can mimic EX. Since EX is nested in SE, this is initially confusing, but turns out to have a very elegant explanation: The "extra" (non-EX) data patterns generated by SE tend to look more like HY patterns than do the EX data patterns. Therefore, when sampling data from SE, there is a tendency to get a larger number of HY-like patterns than one would if sampling from the EX model.

Analyses such as these, which can be carried out prior to data collection, provide guidance on the combination of N and $|t|$ needed to distinguish pairs of models. For example, an N greater than 1000 will be required to distinguish some pairs of models (e.g., PE vs SE; HY vs SE), whereas other comparisons (e.g., EX vs HY) can get by with a smaller design ($N = 250$ appears to be sufficient). By stepping back from a particular data set and examining the overall data-fitting relationship between models given an experimental design, landscaping enhances model testing by revealing the potential distinguishability of the models under study.

3.4 Applying the Landscapes to Empirical Data

How do landscapes help with the interpretation of empirical data? The first step in answering this question is to plot the experimental data in the landscape. Recall that each point in the landscape represents a data pattern whose location in the graph is determined by the relative fits of both models. Experimental data can be overlaid onto the graph by fitting both models and

then following the same procedure. For a given landscape, the further this "experimental data" point is from the criterion, the larger the difference in model fits. Empirical data that fall close to the criterion are fit equally well by both models.

Figure 5 displays the HY/SE landscapes for Wixted and Ebbeson's (1991) design on the left, and Peterson and Peterson's (1959) design on the right, with the empirical data (indicated by circles and triangles) overplotted. Despite the similarity between the landscapes, the empirical data behave rather differently to one another. In the landscape on the right, the data fall close to the tip of the landscape, virtually as far away from the criterion line as possible while still remaining in the SE landscape. It is as unlike HY as possible while still remaining fairly SE-like. In the landscape on the left (same recall task, but different stimuli) the data sets fall within the landscape, but adjacent to the criterion line. The SE model fits the data slightly better, but with log-odds values of 0.72 and 0.42, it is hardly enough to conclude with confidence that they belong to SE.

Insert Figure 5 about here.

Empirical data will move around in a landscape as a function of the models being compared, since each model fits a given data set differently. The graphs in Figure 6 clearly illustrate this point. They are the same as those in the bottom row of Figure 2, except that the data from the two experiments of Burtt and Dobell (1925) are plotted as well. The circles represents the location of the data from the recognition experiment and the triangles are from the recall experiment. Although neither data set abuts the criterion in the left-most graph, the recall data are most useful for discriminating between EX and SE. When the EX model is replaced by the HY model, just the reverse is the case. When the PE and SE models are compared, neither data set is useful for distinguishing them, as both lie next to the criterion line. The landscape itself is especially informative in the right-most graph because even though the data do not distinguish between the models, it shows the models are distinguishable:

Those data patterns that the SE model can fit much better than the PE model are further down the landscape.

Insert Figure 6 about here.

4. Representativeness Analysis

Representativeness analysis is a natural extension of landscaping. The aim is to attempt to *quantify* the relationship between the landscape and a data set. That is, the landscape can be used to estimate how typical a pair of ML values are of the models under consideration. After introducing the measure, we apply it to the 77 historical data sets.

4.1 Defining Representativeness

The issue of where data fall in a landscape highlights the fact that the most useful data are those that are much more representative of one model than another. To be representative of a model, data must fall in its landscape. On the basis of the landscapes in Figure 6, it would be very strange to recommend the use of one model over another if the empirical data did not fall within *any* of the landscapes, irrespective of what the actual ln(ML) values were. If SE tends to generate data in one region of the plot, and the data fall outside of it, then SE is probably not a very good account of those data, even if its fit is better than any of the competitors'.

Landscapes can be used to indicate how representative the empirical data are of a particular model (or, more precisely, the representativeness of the *fits* to that data). By sampling data from one model and fitting them by it and its competitor, we obtain some information about the probability with which the data-fits will end up in a particular region of the plot. Therefore, a way to quantify the representativeness of a region is to estimate the probability that the relative fit will fall within that region. One effective method is to use Gaussian kernels (e.g., Hastie, Tibshirani & Friedman, 2001), in which an unknown probability distribution is approximated by a mixture of a large

number of normal distributions, one for each data set in the landscape (see Appendix B). The result is a distribution of fit *representativeness*, with each point in the landscape having an associated probability[4].

When a data set falls inside a representativeness distribution, we learn how commonplace the relative fit is in the context of the competing model. Such information can be useful to understand the relative fits. For example, empirical data that are located in the region of the landscape near the distribution's peak have relative fit that are quite typical. However, when the relative-fit data set falls outside of the representativeness distribution, we can take a further step in interpretation and safely conclude that the empirical data pattern itself is very unlikely to have been generated by that model.

Because the fits to an empirical data set can be representative of one, both, or neither of a pair of models, it is most informative to combine the data from their two landscapes (e.g., HY/SE and SE/HY) to assess representativeness. This is accomplished simply by plotting both sets of landscape data on the same axes, as illustrated by the bottom graph of Figure 7. In this graph, the black dots denote data generated from HY using the design of Burtt and Dobell (1925), and the grey dots denote data from SE using the same design. Clearly, both models tend to fit their own data better (since the two sets of dots are very distinct), but there is some region of overlap in the top right corner. The upper plot shows the estimated representativeness distributions (SE in white, HY in grey) for the same data. By adding the third dimension, we are better able to understand the relationship between the models and data, because we are now shown the probability with which data will fall in any given location.

Insert Figure 7 about here.

There are two important points to make about our method. Firstly, notice that the representativeness distributions cover a much broader region of the graph than the landscapes themselves. This is a deliberate choice. As Box (1976, p. 792) observes: "Since all models are wrong the scientist must be alert to what is importantly wrong. It is inappropriate to be concerned about mice

when there are tigers abroad". If a model's description of the data is only a little bit wrong, then we have found a mouse, not a tiger. With this in mind, we adopted a very liberal policy and gave all models the benefit of the doubt when assessing representativeness, thereby minimizing the risk of turning mice into tigers. The second point is a complementary one and pertains to how we dealt with tigers. A representativeness value of zero is assigned to all regions that fall "too far" from the landscape. Formally speaking, we excluded the most extreme one thousandth of a percent of the distribution. The technique is just like setting a rejection region in null hypothesis testing, with an extraordinarily low $\alpha = 0.00001$. The reason for this is that if one model assigns a representativeness of 10^{-50} to the data and another assigns 10^{-45}, the correct conclusion is that the data are unrepresentative of both models. Thus we have found a tiger. If we simply took the log-odds of these two probabilities, we would arrive at the conclusion that the data are 10,000 times more representative of the second model than the first. While technically accurate, it is hardly the appropriate lesson to learn from such improbable data.

4.2 Representativeness Analyses of Retention Data

The representativeness of all 77 data sets was measured in the context of all model pairs. The results are easiest to interpret when the data sets are divided into three categories based on whether they were representative of both, one, or neither model. In the first category, representativeness is non-zero for both models (i.e., the data fall inside both of the distributions - see Figure 7), allowing quantitative comparisons to be made between them. The points will be located in the regions of the landscape where the models' distributions overlap. Most of the data sets fall into this category (70%-81%, depending on the pair of models under consideration). Figure 8 plots (the logarithm of) the representativeness probability of all of these data sets with respect to each pair of models. They are laid out identically to the landscape plots of distinguishability, only what is plotted is the representativeness (log probability) assigned to each empirical data set. Rather than identifying each

data set individually with a separate number, they are identified by the number corresponding to the design of the experiment in which it was generated. Numbers below the diagonal indicate that the observed ML values are more typical of the model on the x axis. Just the opposite is the case for data sets above the diagonal.

Insert Figure 8 about here.

The small range of log probabilities on both axes indicates that these data sets are the mice of our database: The evidence provided by one of these data sets is weak to moderate. All are highly and similarly representative of the models. Although the sheer number of these minimally informative data sets might seem startling, it should not come as a surprise when one considers that all four models fit many of the data sets well (see Tables 1 and 2). Inspection of these plots reveals that, in the top row, data tend to fall above the line of equal representativeness, suggesting that HY, PE and SE all consistently outperform EX by a small margin. The other three plots are less clear. While there is some evidence that might suggest that HY and SE outperform PE, it is not particularly convincing. There is very little indication that either of HY or SE is superior: HY wins more often, but SE's wins are more decisive. At best, such data tentatively suggest (HY, SE) > PE > EX. Perhaps the most important contribution of this first and largest category of data sets is to indicate that all models are "near the mark", in the sense that they satisfy the necessary condition of capturing well most of the data sets reported in the empirical literature.

The second category includes those data sets whose ML values are representative of one model, to the extent that they are assigned a non-zero representativeness probability, but are totally unrepresentative of the other model. These are the data sets that fall inside one of the model distributions (as in Figure 7), but not the other. In these cases, it makes little sense to compare representativeness values (or fit values, for that matter). Rather, we should simply acknowledge that one model fits the data sufficiently well and the other does not. These 13 data sets are listed in Table 2. The model

specified in each cell provided the superior fit of the pair listed at the top of each column. Those marked in bold are instances in which the superior fit was particularly dramatic. In those cases, the data fell within the representativeness distribution for one model, but fell so far away from the other distribution that – had we not excluded the tails of the distribution – the representativeness probability would have been less than 10^{-40}. Once again, a cursory look across columns shows that there is clear evidence that EX is inferior to the other three models; note the absence of EX in the cells in the first three columns. The results in the last three column suggest that the ranking of the other models should be SE > HY > PE. However, this ordering is not decisive since there are data sets in which PE and HY are favored over SE.

Insert Table 2 about here.

The third and smallest category contains the true tigers. Ten data sets, all from designs 1, 14 and 15, fell so far outside the landscapes of all four models that they were assigned a representativeness probability of zero. There are at least two interpretations of such universally unrepresentative data: (a) The data are unrepresentative of all models because the data are just too noisy; (b) the data display regularities that are not captured by any model. In this latter case, the data do not tell us which model is to be preferred so much as indicate that they are all wrong. The retention functions for all ten data sets (3 studies) are displayed in Figure 9.

Insert Figure 9 about here.

The four data sets from Bahrick et al. (1975; left panel) display systematic departures from monotonicity, suggesting that the data are fairly noisy. Combined with the fact that this was a field study examining long term retention of high school acquaintances, it is almost certain that these data were influenced by a great many immeasurable factors. In this situation, we might safely conclude that this noise is the source of the unrepresentativeness. The cause of the outliers in Runquist (1983; middle panel) is less clear-cut, since the

study was a little more controlled, and the data do not violate monotonicity quite so extensively. However, the violations are large enough to conclude that noise was probably a contributing factor to the unrepresentativeness of these data. In contrast, the three outliers from the Rubin et al. (1999; right panel) study are smooth, monotonic, and highly similar to one another. Since the three nearly-identical curves resulted from three nearly-identical experiments, it would appear that noisy data are not the cause of the unrepresentativeness. One explanation would be that the data may represent a mixture of retrieval from short-term and long-term memory[5]. Since none of the models discussed here is designed to deal with this situation, it is perhaps unsurprising that none provide a good account of the data.

The representativeness analyses provide another example of the usefulness of landscaping. By mapping the relationship between models and data, we gain new insights about both. When empirical data fall inside both representativeness distributions, we learn that both models can express some of the regularities present in the data. Data sets that are representative of only one model take on a great deal of significance in discriminating between them. In this way, landscaping adds a new dimension to model testing by differentiating data sets in terms of their contribution to distinguishing models. Finally, data that are universally unrepresentative indicate a large mismatch between the models and the data (which may be due to problems in one or both). Without a landscape, this crucial circumstance can be extremely difficult to identify. Although poor fits by both models are a good indication that something is amiss, it is not easy to tell when a fit is bad enough for the data to be considered unrepresentative of the model. Indeed, Rubin et al. (1999) concluded that the SE model provided a very good account of their data (though they were ambivalent regarding its overall status), simply because r^2 was always greater than 0.9. In contrast, the landscaping analysis reveals that the SE model (bounded version) accounts only for the recognition data, and that none of the models captures the cued recall data.

5. Landscaping as a Design Tool

Up to this point we have discussed landscaping as a method of evaluating past research. It can also be used to guide future research. In this section, we briefly describe how the method can be used by experimentalists to learn more about the models under consideration and how to distinguish them.

If the goal is to test which of two models is superior, then the landscape and representativeness plots can help ascertain how they can be distinguished and the degree to which this is possible. To begin with, the locations and sizes of the landscapes will reveal the relationship between the models and thus how they can be distinguished. For example, if, as in Figure 7, there is an asymmetric relationship between them, data capable of discriminating between the models can be generated by one model only, in this case SE. If data are obtained in the center of SE's landscape, they demonstrate the superiority of SE. On the other had, a test of HY would be much more difficult to perform because most of the data it generates are fit well by SE.

As we have seen, the informativeness of empirical data varies a great deal. From a model selection standpoint, ideal data would fall within the landscape of only one model (i.e., category two in the previous discussion). The ease with which such data can be produced in an experiment will depend on a host of factors. Recall that the landscape is created by varying the data-generating model's parameters over what is thought to be a reasonable range. The family of data patterns will likely be larger than the subset of patterns that humans can produce, so experimenters must rely on their knowledge of the field to identify regions in the landscape that yield plausible human-like patterns. Once these regions have been identified, the experimenter can work backwards from this subset of data patterns and design experiments that will yield similarly-shaped patterns of data in one of these regions. Inevitably this will require some trial and error on the part of the experimenter, who will have to fine-tune the experiment (e.g., by altering stimuli, increasing task difficulty, etc.) so that participants produce the desired pattern[6].

This method of experimentation might seem disagreeable because it appears to be devoid of theoretical guidance, but most of the time it is likely to

be nothing more than fine-tuning variables, only it is carried out with knowledge of what the data must look like to yield maximally divergent quantitative differences between models. Comparison of mathematical models requires a level of precision rarely found in experiments, where predictions are most often cast in the form of qualitative, ordinal differences. Landscaping is a technique to increase the precision of experimentation so it approaches that at which the models themselves operate (i.e., the same unit of measurement). Seen in this light, landscaping is a tool intended to aid experimentation by bridging the gap between the coarseness of experimentation and exactness of models.

One way to think about how experimental design influences the landscapes is to distinguish between those variables that affect the shape of the landscape and those that affect where data fall in the landscape. For these four retention models, only N and t affect its shape. Larger samples and more retention intervals increase distinguishability (see Figure 3), providing better conditions under which to collect informative data. These conditions may be necessary to distinguish models, but since the relative representativeness of the empirical data – defined as the log odds obtained from their representativeness probabilities – correlates very weakly with both N and $|t|$ (at 0.05 and 0.07 respectively), it is clear that they alone are insufficient.

Other design variables (e.g., task, stimuli, participants) can affect where data fall in the landscape, so it is these that must be manipulated to obtain maximally informative data. For example, if changes in task or stimuli produce different data patterns, then they will fall in different regions of the landscape. On the other hand, if a variable has no effect on performance (e.g., participant age), the data patterns will be identical and yield points in the landscape that lie on top of one another.

It is tempting to analyze the informativeness data to look for systematic effects of task and stimuli. Are data from cued recall experiments overall more informative than data from recognition experiments? Meta-analyses like this are complicated by the fact that other variables were not held constant across experiments. For example, data from cued recall experiments consistently yielded the most informative data relative to other tasks (recall and

recognition), but these experiments were also the ones with the largest N and $|t|$. Furthermore, even if these other variables were fixed, the relative representativeness measure depends not only on experimental design, but the set of models being compared, as is illustrated in Figure 6. For these reasons, general statements about how other design variables affect model distinguishability and selection are difficult to make unless conclusions are restricted to a small set of models.

In the following two subsections, we illustrate further how landscaping can be used as a design tool. Other types of models are used in these analyses to demonstrate the generality of the method. The ease with which model complexity can be evaluated in a landscape is also discussed.

5.1 Information Integration Models

Consider the task of distinguishing between Oden and Massaro's (1978) Fuzzy Logic Model of Perception (FLMP) and Anderson's (1981) Linear Integration Model (LIM), which are primarily concerned with questions of stimulus identification. A classic example is the perception of speech when participants see and hear a talker speak a syllable. How is the auditory and visual information combined into a single percept (e.g. was it a /ba/ or a /da/)?

Suppose that we decide to try a two-choice categorization task (i.e. choose A or B) with a two by eight design, and 24 participants. This design involves two different levels of one information source (e.g., visual) and eight different levels of the other (e.g., auditory). Thus there are a total of 16 stimuli that may be produced by combining the two evidence sources. 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 $p_{ij} = \theta_i\lambda_j / (\theta_i\lambda_j + (1- \theta_i)(1- \lambda_j))$, whereas LIM predicts that $p_{ij} = (\theta_i + \lambda_j)/2$. The top panel of Figure 10 displays the results of a landscaping analysis for this experimental design, in the form of representativeness plots for FLMP and LIM. The two distributions overlap partially, indicating that there is a fundamental indiscriminability of the models using this design. Worse, ML

is a poor criterion because 30% of the LIM distribution falls on the wrong (FLMP) side of the decision line.

Insert Figure 10 about here.

A minor alteration remedies this situation. The preceding design does not ask

how participants would respond when only one source of evidence is provided, even though the models make different predictions in these circumstances. LIM predicts $p_i = \theta_i$ whereas FLMP predicts that $p_i = \theta_i / (1- \theta_i)$. By adding the 10 extra "unimodal" stimuli (two visual alone and eight auditory alone) to the design and then repeating the representativeness analysis, we obtain the plots in the lower panel of Figure 10. Clearly, the new design is far better able to discriminate between FLMP and LIM. The distributions barely overlap, and the ML decision line separates them quite effectively. The FLMP data are all on the correct side of the line, and only 1% of the LIM data are on the wrong side. This design is far more likely to distinguish the models and has the attractive property of being able to collect data that clearly favor one model or the other because both representativeness regions are distinct (see Navarro et al., 2003, for details).

Notice that the decision threshold depends on the measure used to choose between models. ML, like all measures that are based solely on goodness-of-fit, is biased in favor of the more complex model: That is, the model which is more adept or flexible in fitting data. This is very evident in the first design, where 30% of the LIM data sets are on the FLMP side of the threshold. This problem can be remedied by calculating the complexities of the models and then shifting the decision criterion upward or downward to correct for the difference (note that complexity is a property of the model *and* the experimental design). After calculating the "geometric complexity" measure discussed by Pitt et al. (2002), we found that the criterion should be shifted towards the FLMP distribution[7] by 1.88 (the adjusted criterion is indicated in the plots by the broken line). Although the correction is small, the

improvement is dramatic because the LIM distribution is so peaked: The LIM error rate falls to 3.5% whereas the FLMP error rate rises to only 0.3%.

Contrastingly, in the lower panel of Figure 10, ML makes no errors on FLMP data, and only 1% errors on LIM data. Calculating geometric complexity for this design suggests that the criterion line should be shifted by 4.86 towards the FLMP distribution. However, the improvement that this produces is fairly small, with 0.2% errors on FLMP data and 0.01% errors on LIM data. This is because the models are highly discriminable in this design, so a large complexity difference has little effect. Since it can be difficult to calculate geometric complexity (but see Su, Myung & Pitt, in press), it is nice to note that landscaping can be used to find out when it is really needed: The potential impact of complexity to be gauged simply by moving the criterion upward and downward[8].

5.2 Categorization Models

As a final example we consider two slightly more complex models, interesting properties of which come to light when landscaped. These are Nosofsky's (1986) Generalized Context Model (GCM), and an extension of this model, GCM-γ (Minda & J. Smith 2002; Shin & Nosofsky, 1992). In the GCM, the probability that stimulus i is judged to belong to category K is proportional to its similarity to the exemplars that are known to belong to that category. That is,

$$p(K \mid i) = \sum_{x \in K} s_{ix} \Big/ \sum_{J} \sum_{y \in J} s_{iy}$$

In the GCM-γ model, the probability of category membership is assumed to be proportional to some power γ of the similarity-to-exemplars:

$$p(K \mid i) = \left(\sum_{x \in K} s_{ix} \right)^{\gamma} \Big/ \sum_{J} \left(\sum_{y \in J} s_{iy} \right)^{\gamma}$$

Obviously, the GCM is a special case of the GCM-γ when $\gamma = 1$. In both models, similarity is assumed to decline exponentially with distance in a psychological space (Shepard 1987). In this example, we used the six-dimensional spatial representation employed in Shin and Nosofsky's (1992) Experiment 1. As with previous examples, we sampled data sets from Jeffreys'

distribution and used these to construct landscapes for both models (Fisher information results are presented by Su et al., in press).

The representativeness distributions for these models are shown in Figure 11. As is immediately apparent, the models are remarkably different from each other. This is true despite the fact that GCM is nested within GCM-γ, arising because the γ parameter adds a large set of new data patterns that GCM-γ can produce and GCM cannot. This set is so large that GCM-like patterns are very atypical of GCM-γ.

Insert Figure 11 about here.

Comparison of the solid decision threshold (ML) to the broken one (complexity adjusted) reveals that the latter is far superior. Since the models are nested, ML classifies all patterns as GCM-γ. To compensate for complexity differences between the models, the criterion line should be shifted by 5.2 units, resulting in a drop in error rate, 0% for GCM-γ data, but still 67% for GCM data. While this is clearly a substantial improvement, it leaves a great deal to be desired.

Why was the complexity adjustment not better? Inspection of the representativeness landscapes in Figure 11 reveals that complexity only partly accounts for the differences between the models. Complexity measures consider the relationship between a model and data, but do not consider the interrelationship between models. The result is that in this kind of model discrimination task, a complexity measure can suggest only a constant correction to the ML decision threshold. In this case, however, GCM and GCM-γ have a complicated relationship with each other as well as the data. As is clear from the top-down view shown in the inset of Figure 11 (looking down on the distributions from above), the tails of the GCM-γ distribution "wrap around" the GCM distribution. Because the GCM distribution is so sharply defined, almost any pattern inside that region (which is basically a semi-circular area) is more representative of GCM: Anything outside of this area is more representative of GCM-γ. Therefore, the best way to discriminate between these models would be to define a nonlinear decision threshold along

the borders of this semi-circular region. Measures of model complexity cannot achieve this.

6. General Discussion

Advancement in psychology requires good models and good data. It also requires good methods of integrating the two. Landscaping connects the them by answering questions such as, "What is the relationship between models and data?" and "How representative are data of a model?" The nature of these questions should make it obvious that landscaping focuses on the global behavior of a model, not on a model's data-fitting performance in a single setting.

Perversely, these questions become more pressing and harder to answer the better we do our jobs as scientists. The retention literature is a good example of this. With over a century's worth of data collected, there is little doubt that retention follows a smooth, convex, monotonically decreasing function, and its long-run behavior should be slower-than-exponential (Jost's law; see Alin 1997). Beyond this, it is difficult to discriminate between models that satisfy these constraints if a model's fit to data is used as the sole criterion on which to choose a model. Data-fitting all by itself, as we illustrated at the beginning of the paper (Table 1) is simply is not a good tool for discriminating closely competing models. It is inappropriate given the demands of the job because to advance the science, we need to know more about the models and data than can be learned from fit alone. Landscaping was designed with this goal in mind.

Analyses of the 77 data sets in the context of four leading retention models not only showed how (in)distinguishable the models are, but also demonstrated how experimental design variables, such as N and t, affect distinguishability. For discriminating retention models, N is more effective than t. Representativeness analyses, in which prior data were merged with the landscapes, enabled us to determine how informative data are in distinguishing pairs of models. The fits to most data sets (70% or more depending on the pair of models being compared) are highly typical of those observed of data

generated by all models. Although these data are not very useful for distinguishing between models, they do show that all models can capture some aspects of the process underlying retention. A much smaller number of data sets (17%) proved much more informative, providing clear evidence in favor of one model over another. The results (Table 2) are fairly orderly and suggest a ranking of SE > HY > PE > EX. However, the analyses also revealed that the impressively reliable data of Rubin et al. (1999) are unrepresentative of all models: None of these models could plausibly have generated these data. This indicates that even the best models are incomplete in their description of the form of the retention function.

The knowledge that landscaping contributes about the relationship between models and the representativeness of data fits can be used to choose the next course of action in modeling retention. If the goal is to propose a new model, then it is not enough for the model to fit the data better than it competitors (SE can do that), it is also necessary to show that the data fits are representative of this model (i.e., it could plausibly have generated them). If the goal is to collect new data to decide between the top two models, for example, one should identify an experimental design that yields landscapes that do not completely overlap, inspect data sets in these non-overlapping regions, and then design an experiment to yield such data.

The preceding discussion brings out the point that landscaping can be used to improve postdiction and prediction in experimentation, serving as a bridge between model and data. Postdiction is a common form of model evaluation in psychology. Landscaping makes such tests more stringent because it imposes the additional condition that the data are representative of the new model, not just that the model fitted the data better than its competitors. Prediction is an even more stringent and convincing test of model adequacy because the data have not yet been collected. Landscaping provides the means to identify optimal tests by determining the experimental conditions that will be most favorable to generating data that are more representative of one model than another.

Even with an intimate knowledge of a field, it can be difficult to predict the impact of specific variables on performance. Landscaping allows one to

learn how some design changes will affect model distinguishability. Without this knowledge, the effect of specific manipulations are unknown and can even be misunderstood, as we demonstrated with sample size. An increase in N is not guaranteed to improve model distinguishability. It reduces error variance, which should assist in distinguishing models, but unless the data are representative of one and only one model, the outcome of the experiment may be disappointing. This situation could be prevented by first viewing the relevant landscapes, which collectively would suggest at least one course of action, even if it means trying something else.

The brief examples of landscaping information integration models and categorization models demonstrate the impact model complexity can have on the landscape and the tool's wider applicability. As is shown in Figures 10 and 11, a landscape plot is tailor-made for not only displaying but also evaluating the effects of model complexity on model distinguishability. The criterion shifts toward the more complex model by the amount it exceeds its competitor in complexity. The effect of this adjustment will be substantial if the distribution of at least one model is concentrated on or near the criterion. Otherwise they probably will be negligible. When used for this purpose, landscaping provide an easy means with which to evaluate the impact complexity might have on model selection without having to calculate complexity, which can be challenging.

The method of landscaping described in this paper is applicable to statistical models (i.e., those for which there exists a likelihood function). As a tool, landscaping is in principle applicable to any type of model (e.g., connectionist, qualitative). The necessary ingredients are a way to express the performance relationship between the models themselves and also with experimental data. We are currently developing a method to landscape localist connectionist models. Details of this work can be found in W. Kim, Navarro, Pitt and Myung (in press).

References

Alin, L. (1997). The memory laws of Jost. Technical Report: *Götebotg Psychological Reports, 27*, no. 1.

Anderson, J. R. & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science, 2*, 396-408.

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

Ashby, F. G., Maddox, W. T. & Lee, W. W. (1994). On the dangers of averaging across subjects when using multidimensional scaling or the similarity-choice model. *Psychological Science 5,* 144-151.

Bahrick, H. P., Bahrick, P. O. & Wittlinger, R. P. (1975). Fifty years of memory for names and faces: A cross-sectional approach. *Journal of Experimental Psychology: General, 104*, 54-75.

Balasubramanian, V. (1997). Statistical inference, Occam's razor, and statistical mechanics on the space of probability distributions. *Neural Computation, 9*, 347-368.

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association, 71*, 791-799.

Bregman, A. S. (1968). Forgetting curves with semantic, phonetic, graphic, and contiguity cues. *Journal of Experimental Psychology, 78*, 539-546.

Brown, S. & Heathcote, A. (2003). Averaging learning curves across and within participants. *Behavior Research Methods, Instruments and Computers, 35,* 11-21.

Burtt, H. E. & Dobell, E. M. (1925). The curve of forgetting for advertising material. *Journal of Applied Psychology, 9*, 5-21.

Conway, M. A., Cohen, G. & Stanhope, N. (1991). On the very long term retention of knowledge acquired through formal education: Twelve years of cognitive psychology. *Journal of Experimental Psychology: General, 120*, 395-409.

de Bruijn, N. G. (1958). *Asymptotic Methods in Analysis.* Amsterdam: North-Holland.

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Review 53,* 134-140.

Gehring, R. E., Toglia, M. P. & Kimble, G. A. (1976). Recognition memory for words and pictures at short and long retention intervals. *Memory & Cognition, 4*, 256-260.

Geweke, J. (1999a). Using simulation methods for Bayesian econometric models: Inference, development and communication. *Econometric Review, 18*, 1-126.

Geweke, J. (1999b). Simulation methods for model criticism and robustness analysis. In J. O. Berger, J. M. Bernardo, A. P. Dawid, and A. F. M. Smith (eds) *Bayesian Statistics 6* (pp. 275-299). Oxford: Oxford University Press.

Gilks, W. R. , Richardson, S., & Spiegelhalter, D. J. (1995). *Markov Chain Monte Carlo in Practice.* London: Chapman and Hall.

Gill, J. (2002). *Bayesian Methods: A Social and Behavioral Sciences Approach.* Boca Raton: Chapman & Hall.

Grünwald, P. (2000). Model selection based on minimum description length. *Journal of Mathematical Psychology, 44*, 133-151.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The Elements of Statistical Learning*. New York: Springer-Verlag.

Jeffreys, H. (1961). *Theory of Probability* (3rd ed.). London: Oxford University Press.

Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*, 773-795.

Kass, R. E. & Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association, 91*, 1343-1370.

Kim, W., Navarro, D. J., Pitt, M. A. & Myung, I. J. (in press). An MCMC-based method of comparing connectionist models in cognitive science. *Advances in Neural Information Processing Systems, 16*.

Krueger, W. C. F. (1929). The effect of overlearning on retention. *Journal of Experimental Psychology, 12*, 71-78.

Lee, M. D. & Pope, K. J. (2003). Avoiding the dangers of averaging across subjects when using multidimensional scaling. *Journal of Mathematical Psychology,47*, 32-46.

Longmore, B. E. & Knight, R. G. (1988). The effect of intellectual deterioration on retention deficits in amnesic alcoholics. *Journal of Abnormal Psychology 97*, 448-454.

Luh, C. W. (1922). The conditions of retention. *Psychological Monographs, 31,* whole no. 142.

McClelland, J. L. & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18*, 1-86.

MacLeod, C. M. (1988). Forgotten but not gone: Savings for pictures and words in long term memory. *Journal of Experimental Psychology: Learning, Memory and Cognition, 14*, 195-212.

Minda, J. P. & Smith, J. D. (2002). Comparing prototype-based and exemplar-based accounts of category learning and attentional allocation. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 28,* 275-292.

Murdock, B. B. Jr (1961). The retention of individual items. *Journal of Experimental Psychology 62*, 618-625.

Murray, M. K. & Rice, J. W. (1993). *Differential Geometry and Statistics.* London: Chapman & Hall.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology* , 47, 90-100.

Myung, I. J., Balasubramanian, V., & Pitt, M. A. (2000). Counting probability distributions: Differential geometry and model selection. *Proceedings of the National Academy of Sciences USA, 97,* 11170-11175.

Myung, I. J., Kim, C., & Pitt, M. A. (2000). Toward an explanation of the power-law artifact: Insights from response surface analysis. *Memory & Cognition, 28* , 832-840.

Myung, I. J. & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review, 4* , 79-95.

Navarro, D. J. (submitted). Misbehavior of the Fisher information apprioximation to minimum description length. Submitted to *Neural Computation*.

Navarro, D. J., Myung, I. J., Pitt, M. A., & Kim, W. (2003). Global model analysis by landscaping. *Proceedings of the 25th Annual Conference of the Cognitive Science Society*.

Nocedal, J. & Wright, S. J. (1999). *Numerical Optimization*. New York: Springer-Verlag.

Norris, D., McQueen, J. M. & Cutler, A. (2000). Merging phonetic and lexical information in phonetic decision-making. *Behavioral & Brain Sciences, 23*, 299-325.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship, *Journal of Experimental Psychology: General, 115,* 39-57.

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

Peterson, L. R. & Peterson, M. J. (1959). Short term retention of individual verbal items. *Journal of Experimental Psychology 58*, 193-198.

Pitt, M. A., Kim, W., & Myung, I. J. (2003). Flexibility vs generalizability in model selection. *Psychonomic Bulletin & Review*, 10, 29-44.

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.

Raftery, A. E. (1995). Bayesian model selection in social research. In P. V. Marsden (Ed.), *Sociological Methodology*, (p. 111-196). Oxford: Blackwell.

Ratcliff, R. & Smith, P. (in press). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review.*

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

Rissanen, J. (2001). Strong optimality of the normalized ML models as universal codes and information in data. *IEEE Transactions on Information Theory 47,* 1712-1717.

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

Rubin, D. C., Hinton, S. & Wenzel, A. (1999). The precise time course of retention. *Journal of Experimental Psychology: Learning, Memory & Cognition 25*, 1161-1176.

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

Runquist, W. (1983). Some effects of remembering on forgetting. *Memory & Cognition, 11*, 641-650.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science, 237*, 1317-1323.

Shin, H. J. & Nosofsky, R. M. (1992). Similarity-scaling studies of dot-pattern classification and recognition. *Journal of Experimental Psychology: Learning, Memory and Cognition, 121,* 278-304.

Sikström, S. (2002). Forgetting curves: Implications for connectionist models. *Cognitive Psychology 45*, 95-152.

Sloman, S. A., Hayman, C. A. G., Ohta, N., Law, J. & Tulving, E. (1988). Forgetting in primed fragment completion. *Journal of Experimental Psychology: Learning, Memory & Cognition, 11*, 812-816.

Squire, L. R. (1989). On the course of forgetting in very long term memory. *Journal of Experimental Psychology: Learning, Memory & Cognition, 15*, 241-245.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E. J. & Blum, B (2003). Inferring causal networks from observations and interventions. *Cognitive Science, 27*, 453-489.

Strong, E. K. (1913). The effect of time interval upon recognition memory. *Psychological Review, 30*, 339-32.

Su, Y., Myung, I. J. & Pitt, M. A. (in press). Minimum description length and cognitive modeling. In P. Grünwald, I. J. Myung, I. J., & M. A. Pitt (eds.) *Advances in Minimum Description Length: Theory and Applications*. MIT Press.

Thompson, C. P. (1982). Memory for unique personal events: The roommate study. *Memory & Cognition, 10*, 324-332.

Wagenmakers, E. J., Ratcliff, R., Gomez, P. & Iverson, G. J. (in press). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology.*

Wickelgren, W. A. (1968). Sparing of short-term memory in an amnesiac patient: Implications of strength theory of memory. *Neuropsychologica, 6,* 235-244.

Wickelgren, W. A. (1972). Trace resistance and decay of long-term memory. *Journal of Mathematical Psychology, 9*, 418-455.

Wixted, J. T. & Ebbesen, E. B. (1991). On the form of forgetting. *Psychological Science, 2*, 409-415.

Wixted, J. T. & Ebbesen, E. B. (1997). Genuine power curves in forgetting: A quantitative analysis of individual subject forgetting functions. *Memory & Cognition, 25,* 731-739.

Author Note

The authors were supported by NIH grant R01-MH57472 awarded to IJM and MAP. DJN was also supported by a grant from the Office of Research at OSU. We would like to thank Nancy Briggs, Woojae Kim, Yong Su, Eric-Jan Wagenmakers, John Wixted, and two anonymous reviewers for helpful comments and insights that substantially improved the paper. DJN would also like to thank Marcus Butavicius for his thoughts on experimental design.

Appendix A: Summary of Database

Design	Source	Mean N	$\|t\|$	ID	Participants	Stimuli	Task	Notes
1	Bahrick et al. (1975)	436	9	a	elderly	pictures	recognition	yearbook pictures
1	Bahrick et al. (1975)	436	9	b	elderly	word strings	recognition	names of people from yearbook
1	Bahrick et al. (1975)	436	9	c	elderly	words + pictures	matching	names and yearbook pictures: face->name
1	Bahrick et al. (1975)	436	9	d	elderly	words + pictures	matching	names and yearbook pictures: name->face
1	Bahrick et al. (1975)	436	9	e	elderly	words + pictures	free recall	
2	Bregman (1968)	92	8	a	undergrads	words	cued recall	semantic association
2	Bregman (1968)	92	8	b	undergrads	words	cued recall	stem completion
2	Bregman (1968)	92	8	c	undergrads	words	cued recall	phonetic (rhyming) association
2	Bregman (1968)	92	8	d	undergrads	words	cued recall	associate (contiguity)
2	Bregman (1968)	92	8	e	undergrads	words	cued recall	paired association varied across blocks
2	Bregman (1968)	92	8	f	undergrads	words	cued recall	paired association varied across blocks
2	Bregman (1968)	92	8	g	undergrads	words	cued recall	paired association varied across blocks
2	Bregman (1968)	92	8	h	undergrads	words	cued recall	paired association varied across blocks
3	Burtt & Dobell (1925)	1160	5	a	undergrads	words	recognition	paired associates learned previously
3	Burtt & Dobell (1925)	1160	5	b	undergrads	words	cued recall	paired associates learned previously
4	Conway et al. (1991)	187	12	a	adults	words	recognition	names
4	Conway et al. (1991)	187	12	b	adults	words	recognition	concepts
4	Conway et al. (1991)	187	12	c	adults	words	recall	names
4	Conway et al. (1991)	187	12	d	adults	words	recall	concepts
4	Conway et al. (1991)	187	12	e	adults	words	verification	
4	Conway et al. (1991)	187	12	f	adults	words	verification	
5	Gehring et al. (1976)	1447	6	a	undergrads	pictures	recognition	
5	Gehring et al. (1976)	1447	6	b	grads	words	recognition	nouns denoting the pictures
6	Krueger (1929)	280	6	a	grads	words	recall	100% overlearn: monosyllables
6	Krueger (1929)	280	6	b	grads	words	recall	150% overlearn: monosyllables
6	Krueger (1929)	280	6	c	grads	words	recall	200% overlearn: monosyllables

7	Longmore & Knight (1988)			240	5	a	adults	words	recall	normals: monosyllables
7	Longmore & Knight (1988)			240	5	b	adults	words	recall	Korsakoffs; monosyllables
8	Longmore & Knight (1988)			120	5		adults	words	recall	alcoholics with no signs of Korsakoffs; monosyllables

Design	**Source**	**Mean *N***	$\|t\|$	**ID**	**Participants**	**Stimuli**	**Task**	**Notes**
9	Luh (1922)	240	5	a	grads	trigrams	antic. recall	no learning
9	Luh (1922)	240	5	b	grads	trigrams	free recall	no learning
9	Luh (1922)	240	5	c	grads	trigrams	recognition	no learning
9	Luh (1922)	240	5	d	grads	trigrams	ordering	no learning
9	Luh (1922)	240	5	e	grads	trigrams	free recall	% of previous learning: 100
9	Luh (1922)	240	5	f	grads	trigrams	recognition	% of previous learning: 100
9	Luh (1922)	240	5	g	grads	trigrams	ordering	% of previous learning: 100
9	Luh (1922)	240	5	h	grads	trigrams	free recall	% of previous learning: 67
9	Luh (1922)	240	5	i	grads	trigrams	recognition	% of previous learning: 67
9	Luh (1922)	240	5	j	grads	trigrams	ordering	% of previous learning: 67
9	Luh (1922)	240	5	k	grads	trigrams	free recall	% of previous learning: 33
9	Luh (1922)	240	5	l	grads	trigrams	recognition	% of previous learning: 33
9	Luh (1922)	240	5	m	grads	trigrams	ordering	% of previous learning: 33
10	Luh (1922)	240	7	a	grads	trigrams	free recall	% of previous learning: 150
10	Luh (1922)	240	7	b	grads	trigrams	recognition	% of previous learning: 150
11	MacLeod (1988)	320	5		undergrads	word + pictures	recognition	paired associates: words were numbers
12	Murdock (1961)	192	6	a	undergrads	trigrams	recall	backward counting interference
12	Murdock (1961)	192	6	b	undergrads	words	recall	backward counting interference
12	Murdock (1961)	192	6	c	undergrads	word triads	recall	backward counting interference
12	Murdock (1961)	192	6	d	undergrads	words	recall	varied # prior words in list 0
12	Murdock (1961)	192	6	e	undergrads	words	recall	varied # prior words in list 3
12	Murdock (1961)	192	6	f	undergrads	words	recall	varied # prior words in list 6

12	Murdock (1961)	192	6	g	undergrads	words	recall	varied # prior words in list 9
12	Murdock (1961)	192	6	h	undergrads	words	recall	varied # prior words in list 12
13	Peterson & Peterson (1959)	192	6		undergrads	trigrams	recall	backward counting interference
14	Rubin et al. (1999)	900	10	a	undergrads	words	cued recall	paired associates in same color
14	Rubin et al. (1999)	900	10	b	undergrads	words	cued recall	all words white font
14	Rubin et al. (1999)	900	10	c	undergrads	words	cued recall	words in random colors
14	Rubin et al. (1999)	900	10	d	undergrads	words	recognition	old-new recognition
14	Rubin et al. (1999)	900	10	e	undergrads	words	recognition	remember-know recognition

Design	**Source**	**Mean *N***	**$\|t\|$**	**ID**	**Participants**	**Stimuli**	**Task**	**Notes**
15	Runquist (1983)	1728	6	a	undergrads	words	cued recall	paired associates seen 3x previously and tested
15	Runquist (1983)	1728	6	b	undergrads	words	cued recall	paired associates seen 1x previously and tested
15	Runquist (1983)	1728	6	c	undergrads	words	cued recall	paired associates seen 3x previously and not tested
15	Runquist (1983)	1728	6	d	undergrads	words	cued recall	paired associates seen 1x previously and not tested
16	Sloman et al. (1988)	672	14		grads	words	completion	fragments presented in reverse order
17	Squire (1989)	1078	15		UG, elderly	word strings	recognition	names of TV shows (mean age 41)
18	Strong (1913)	100	13	a	adults	words	recognition	recognize from a list twice as long as the study list
18	Strong (1913)	100	13	b	adults	words	recognition	recognize from a list twice as long as the study list
19	Strong (1913)	40	13	a	adults	words	recognition	recognize from a list twice as long as the study list
19	Strong (1913)	40	13	b	adults	words	recognition	recognize from a list twice as long as the study list
20	Strong (1913)	20	13		adults	words	recognition	recognize from a list twice as long as the study list
21	Thompson (1982)	128	9		undergrads	dates	recognition	memorable events generated by college roommate
22	Wixted & Ebbeson (1991)	432	5	a	undergrads	words	recall	learned list well before test
22	Wixted & Ebbeson (1991)	432	5	b	undergrads	words	recall	learned list poorly before test
23	Wickelgren (1968)	46	8		H.M.	spoken digits	recognition	determine whether 1 digit was in the list
24	Wickelgren (1968)	40	5		H.M.	spoken digits	recognition	determine whether 3 digit number was in a list of 5
25	Wickelgren (1968)	50	7		H.M.	spoken digits	recognition	determine whether 3 digit number was in a list of 7

Appendix B: Statistical Comments

B.1 Jeffreys' Prior and Worst-Case Inference

In this paper we have sampled parameter sets from Jeffreys' (1961) distribution,

$$p(\theta \mid M) = \frac{\sqrt{\det(\mathbf{I}(\theta))}}{\int_{\Omega} \sqrt{\det(\mathbf{I}(\theta))} d\theta}$$

where Ω denotes the parameter space and $\mathbf{I}(\theta) = [i_{ab}(\theta)]$ denotes the Fisher information matrix,

$$i_{ab}(\theta) = -E_D\left[\frac{\partial^2 \ln p(D \mid \theta)}{\partial\theta_a \partial\theta_b}\right].$$

Using Jeffreys' distribution as a prior over the parameters is common in Bayesian statistics, partly because it is reparameterization-invariant, but also because it is a uniform distribution on the space of probability distributions (Balasubramanian, 1997). It is in this sense that Jeffreys' distribution is a "worst case" (or noninformative) distribution. Under Jeffreys' prior, the amount of prior uncertainty about the probability distribution $p(\ \cdot\ \mid M, \theta)$ is maximized. This differs from the uniform distribution, which maximizes uncertainty about θ. Since it is generally acknowledged that parameters of a model are rather arbitrary "indexing tools", it is better to work with the probability distributions themselves when justifying a prior. We have chosen Jeffreys' prior here because it treats each (distinguishable) probability distribution equally.

It is worth noting that this issue falls within the larger question of how to set reasonable priors for Bayesian inference (see Kass & Wasserman for a thorough discussion). While this matter is beyond the scope of this paper, we note in passing that it is often worth considering a number of different priors, in order to check that the results are not unduly influenced by the prior. In general, however, we feel that "worst-case" priors are a reasonable first choice.

Unfortunately, these desirable properties can come at a cost: In some cases it is difficult to sample from Jeffreys' distribution, and may require

Markov Chain Monte Carlo methods (MCMC; see Gilks, Richardson & Spiegelhalter, 1995). Fortunately, MCMC is now a standard technique in statistics, available through widely-available software such as BUGS. In other cases, such as when the error distribution is normal, closed forms are available, making the sampling simpler.

B.2 Landscaping and Bayesian Marginals

We briefly discuss the connections between landscaping and Bayesian statistics (e.g., Gill 2002, Robert 2001). Although the basic statistical foundations of the technique are established here, it would be nice to explore this further, pursuing the connections with Bayesian methods and Minimum Description Length (e.g. Grünwald 2000; Rissanen 1996). From a Bayesian perspective, one defines the predictive distribution for the data as,

$$p(D \mid M) = \int p(D \mid M, \theta) p(\theta \mid M) d\theta .$$

This states that the data D are sampled from M according to its marginal distribution. It is well-known that, because they consider the behavior of the model across its entire parameter space (i.e., by adopting a global model analysis), Bayesian approaches are able to identify and compensate for model complexity (Myung & Pitt, 1997).

Notice, however, that the landscaping approach also specifies a prior distribution $p(\theta \mid M)$, from which the parameters are sampled, and a likelihood function $p(D \mid M, \theta)$ from which the data are generated. In short, landscape data are sampled from the Bayesian marginal distribution $p(D \mid M)$. In this sense, there is a direct connection between landscaping and Bayesian model selection.

As a final note, this observation allows us to make the notion of "distinguishability" a little more explicit. Under one definition, two models would be indistinguishable if they both contained the data-generating distribution. We prefer not to use this definition, since it would allow one to "cheat", by proposing very elaborate models that incorporate an enormous number of distributions. Rather, we adopt the "universal distribution" approach (e.g. Rissanen 2001), in which a family of distributions can be

"summarized" by a *single* distribution. In this case, the marginal distribution $p(D \mid M)$ is a universal distribution. Under this approach, two models are considered to be indistinguishable if their universal distributions are highly similar to one another.

B.3 Representativeness and Partial Information Bayes.

Suppose, however, that we were not interested the likelihood of the data itself $p(D \mid M)$. Rather, we were interested in the likelihood of the fits to that data for two models, x and y. That is, $x = \ln p(D \mid M_X, \boldsymbol{\theta}^*)$ and $y = \ln p(D \mid M_Y, \boldsymbol{\theta}^*)$, the maximum log-likelihood for the data obtained under the models. The quantity that we are interested in is the probability of observing both x and y if the data were truly generated from M_X. This is denoted $p(x, y \mid M_X)$, and corresponds to the quantity that we have called the *representativeness* of the fits. It is important to note that x and y depend directly on the data set D, but only indirectly on $p(D \mid M)$, through the data itself. A Bayesian approach yields

$$p(x, y \mid M_X) = \int p(x, y \mid M_X, \theta)\, p(\theta \mid M_X)\, d\theta$$

The $p(x, y \mid M_X, \boldsymbol{\theta})$ quantity is the probability of generating a data set from model X at some parameters $\boldsymbol{\theta}$ that yields the same fits x and y as the original data set D. Note that x and y are statistics of the data set D and the model set (X and Y), and do not carry as much information as the data itself (and is therefore called the "partial information" Bayesian marginal by Geweke 1999a, 1999b). Indeed the relationship between $p(x, y \mid M)$ and $p(D \mid M)$ may be non-trivial. However, given that the fits x and y are commonly used to draw inferences about models, it is useful to consider the likelihood of observing them given M_X or M_Y. Accordingly, the representativeness of these fits is an important consideration when evaluating models.

Ideally, the representativeness could be found by solving this integral analytically. In general, however, the integral is intractable and must be approximated. It is worth noting that the commonly used Laplace approximation (de Bruijn 1958; Kass and Raftery 1995) is inappropriate, as the posterior is not well-approximated by a multivariate Gaussian. With this in

mind, we estimate it numerically, using the N sets of landscape data, denoted $L = (L_1,..., L_N)$. Formally, we estimated the representativeness of D for model X using the Nadaraya-Watson kernel-weighted average,

$$\hat{p}(x_D, y_D \mid M_X) = \frac{\sum_{i=1}^{N} K\left[x_D - x_{L_i}, y_D - y_{L_i}\right] \times S(i)}{\sum_{i=1}^{N} K\left[x_D - x_{L_i}, y_D - y_{L_i}\right]}$$

where $K(x, y)$ denotes the kernel, a bivariate Gaussian distribution with mean at (x, y), and $S(i)$ is an indicator function that equals 1 if the i-th data set is from M_X, and 0 otherwise. Subscripts are used to indicate which data set the fits x and y refer to. This is the representativeness measure reported in this paper. Since we wished to be very generous to the models, these distributions had covariance matrix equal to 10x**I**, where **I** denotes the identity matrix. The choice of 10 is somewhat arbitrary, and users wishing to do a precise analysis should certainly consider some formal procedure for choosing an optimal kernel variance. In this application, however, our interest is in setting an overly-large kernel width, because of our "mice and tigers" approach. Accordingly, the kernel width of 10 is *excessively* large for these data. In any case, the statistical properties of the Nadaraya-Watson estimator can be found in Hastie et al. (2001).

Table 1. Fits of five retention models using ln(ML) to five data sets of Rubin et al. (1999; top), the average fit to the remaining 72 data sets (bottom).

Data set	EX	HY	PE	SE
Recall: matched color	-561	-228	-296	-162
Recall: white	-565	-213	-288	-143
Recall: random color	-534	-194	-267	-133
Recognition: remember + know	-231	-58	-83	-39
Recognition: remember	-196	-49	-64	-44
Average from remaining 72 set	-60	-38	-36	-31

Table 2. Data that are qualitatively captured by one model, and not by another. Boldfaced items indicate decisions in which the evidence in favor of making this decision is particularly compelling.

Study	ID	EX v HY	EX v PE	EX v SE	HY v PE	HY v SE	PE v SE
Bregman (1968)	e	HY	PE	SE			
Bregman (1968)	f			SE		SE	SE
Bregman (1968)	g	HY		SE	HY		SE
Bregman (1968)	h			SE		SE	SE
Burtt & Dobell (1925)	b	HY	PE	SE			
Conway et al. (1991)	c			SE	HY	HY	SE
Gehring et al. (1976)	a	**HY**	**PE**				PE
Gehring et al. (1976)	b	**HY**	**PE**	**SE**			
Rubin et al. (1999)	d			**SE**		SE	SE
Rubin et al. (1999)	e	**HY**		**SE**	HY		SE
Runquist (1983)	c			**SE**		SE	SE
Strong (1913)	a	HY	PE	SE			
Strong (1913)	b	HY	PE	SE			

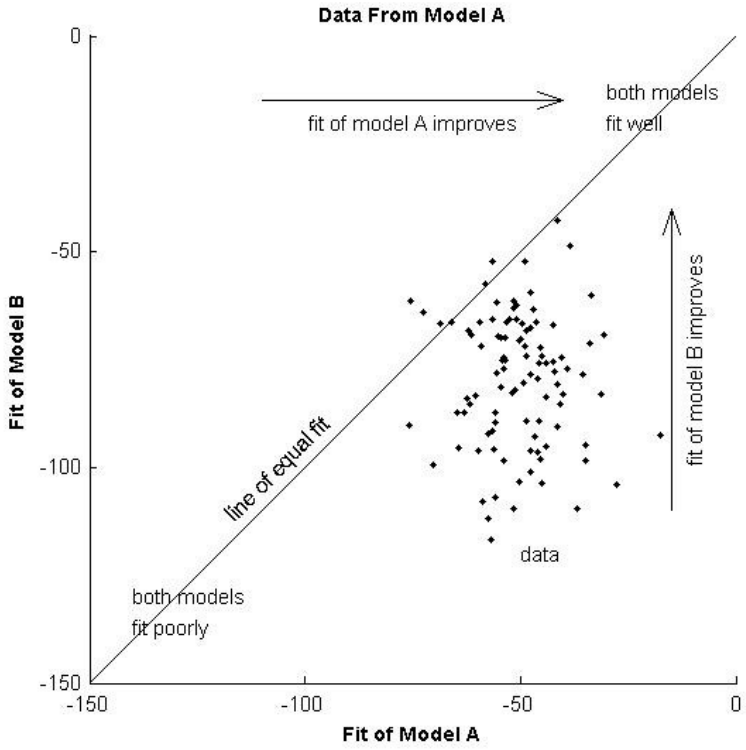


Figure 1. Stylized depiction of a landscape graph in which the fits of model A to a large number of data sets are plotted against the fits of model B. Since the data sets are sampled from model A, the plot provides an indication of how effectively model B can mimic model A.

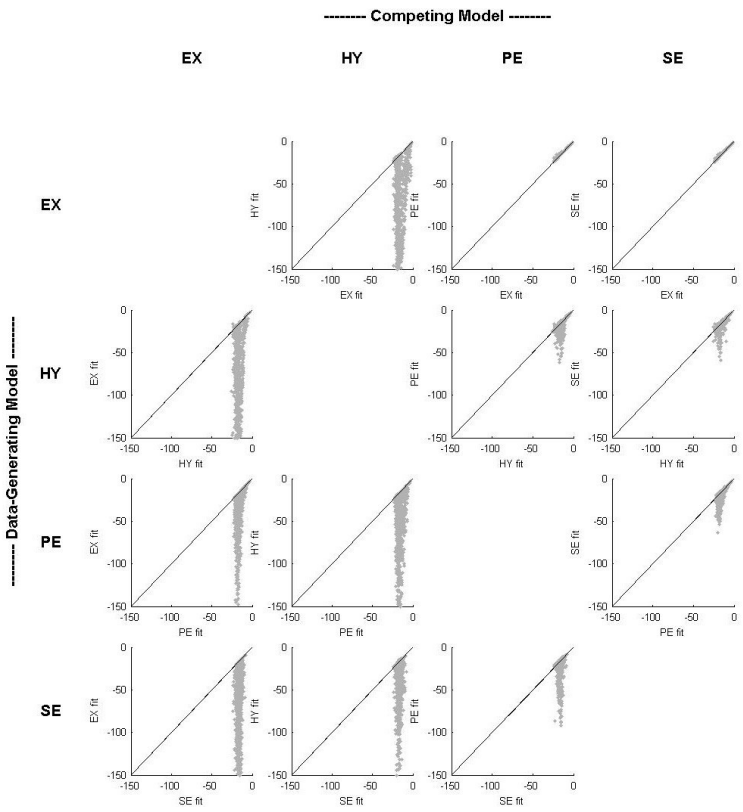


Figure 2: Landscapes corresponding to the Burtt and Dobell (1925) experimental design. For each subplot, the data-generating model appears on the x axis, and the competing model appears on the y axis. The ordering of models from left to right (and top to bottom) is EX, HY, PE, SE.

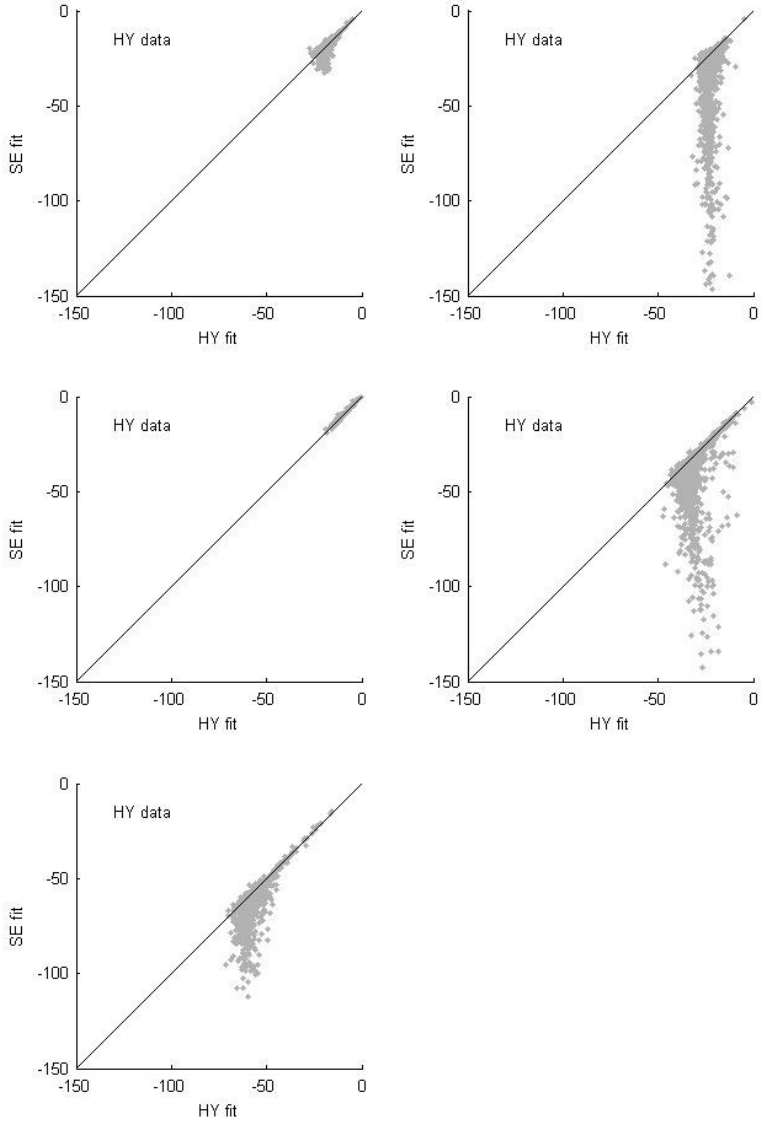


Figure 3: Illustrations of the effects of N and $|t|$ on the landscapes. Landscapes in the top row are from Krueger (1929; left; N=280, $|t|$=6), and Runquist (1983; right; N=1728, $|t|$=6). Those in the middle row are from Wickelgren (1968; left; N=40, $|t|$=5) and Strong (1913; right; N=40, $|t|$=13). The bottom panel is from Squire (1989; N=1078, $|t|$=15).

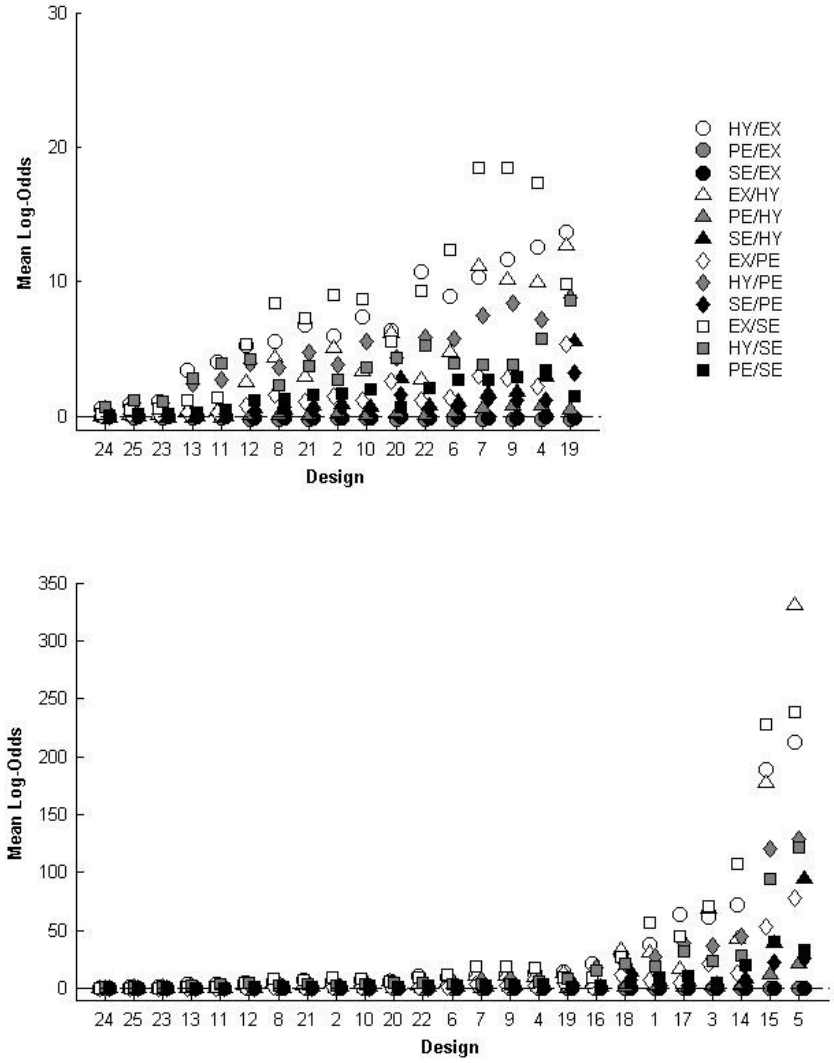


Figure 4: Mean log-odds for all 25 designs and all 12 pairwise comparisons, in order of increasing overall distinguishability. The upper panel is a close-up view of the 17 least discriminating designs.

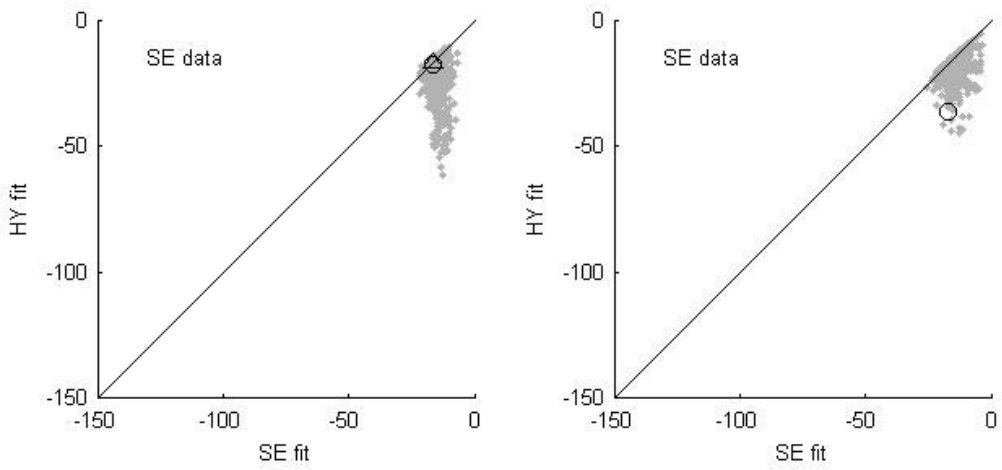


Figure 5: Informative and uninformative data. The data on the left are from Wixted and Ebbeson (1991). The data on the right are from Peterson and Peterson (1959).

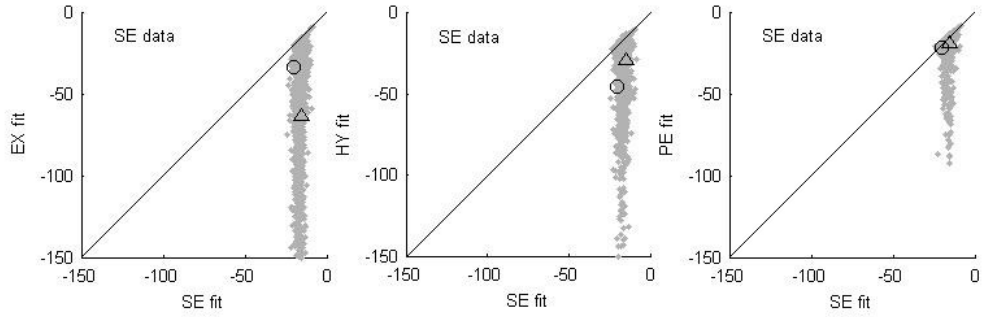


Figure 6: Data sets move as a function of the competing model. The circles correspond to a single empirical data set (from Burtt and Dobell 1925), as do the triangles. The landscape data are from the SE model.

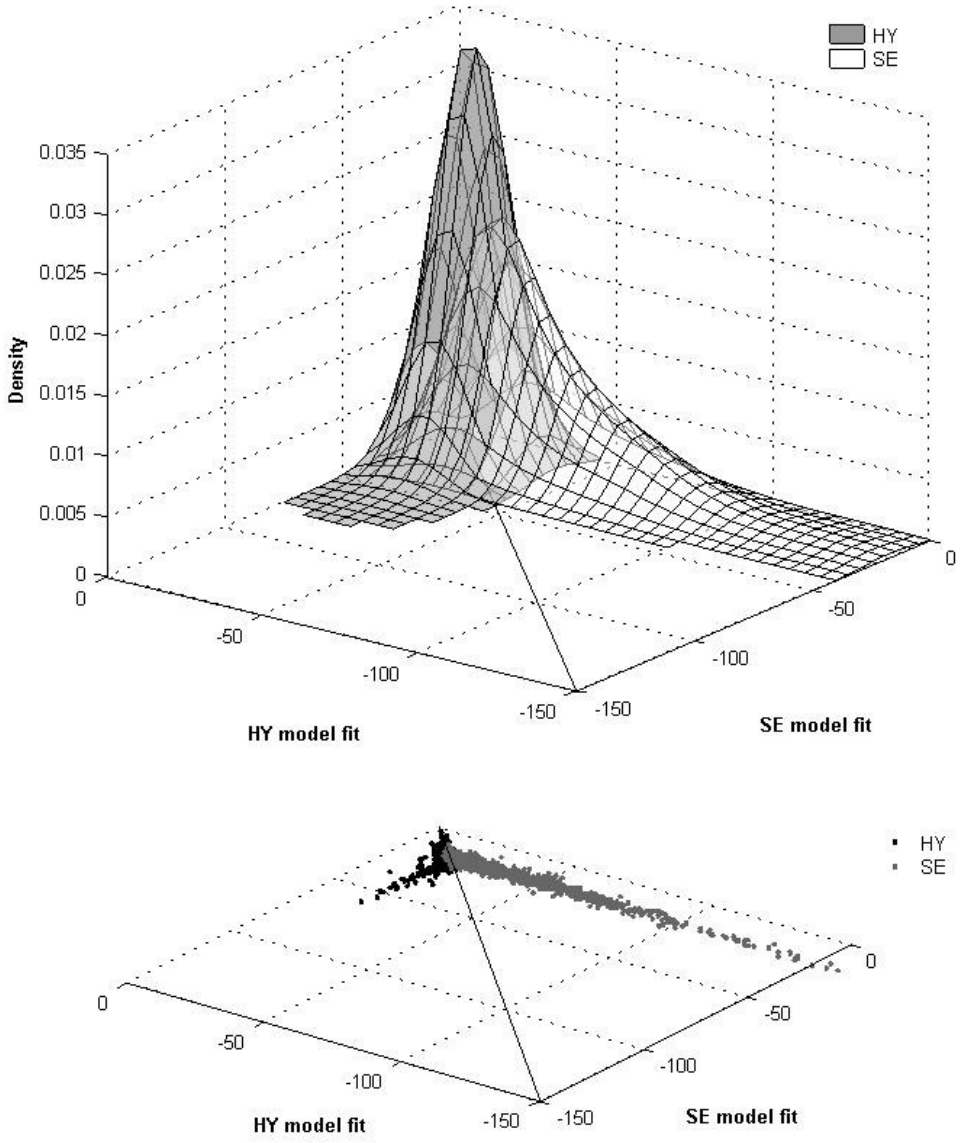


Figure 7: Representativeness distributions for the HY and SE models. The lower panel shows the landscapes for these models (HY in black, SE in grey), and the upper panel shows the (very generous) distributions that correspond to the two models.

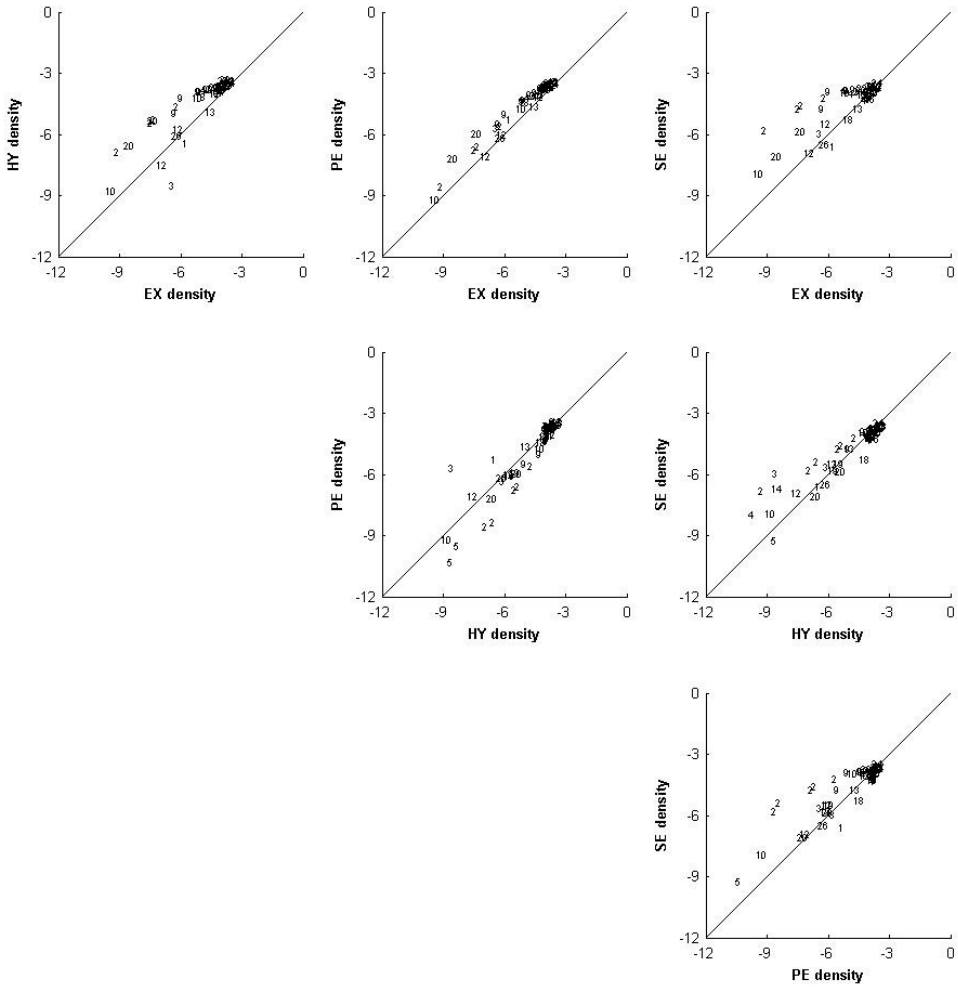


Figure 8: Pairwise comparisons of representativeness probability (densities) when both models perform adequately. Numbers denote the design to which the corresponding empirical data set belongs.

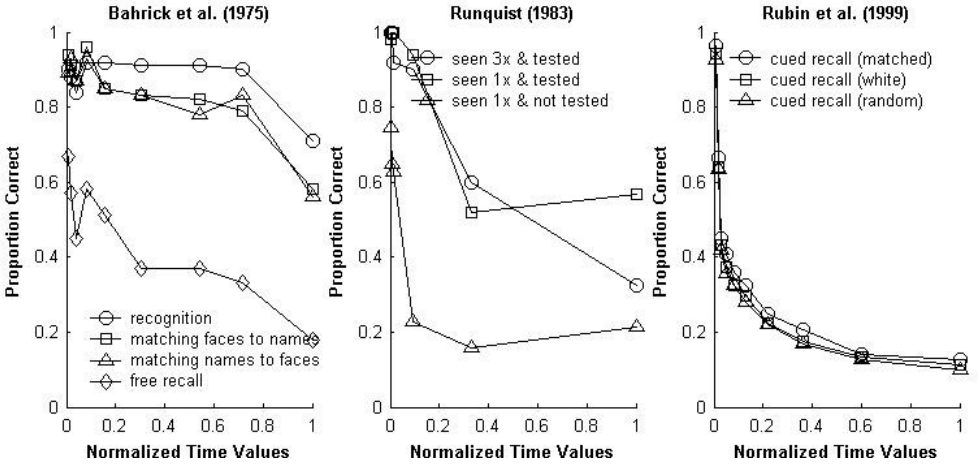
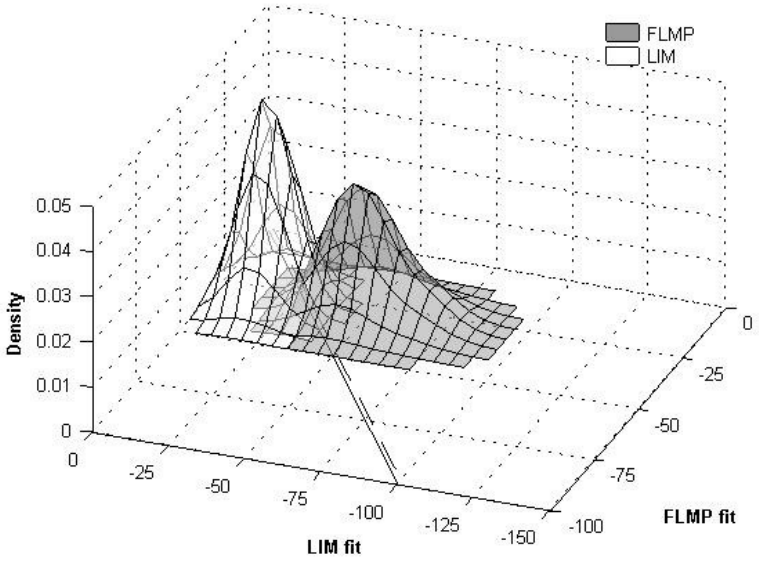


Figure 9: Retention functions from Bahrick et al. (1975) are in the left graph. Those from Rubin et al. (1999) are on the right. Data in the middle panel are from Runquist (1983).



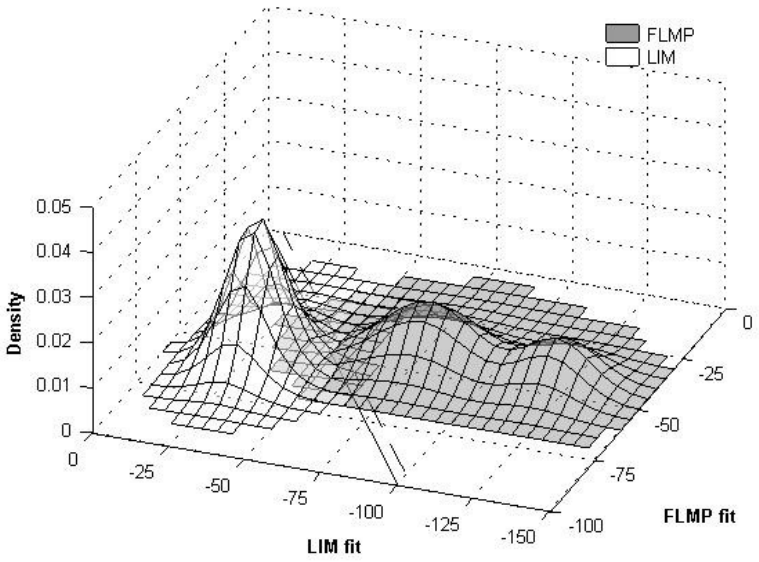


Figure 10: Representativeness plots for data sampled from FLMP (in grey) and LIM (in white) for two different experimental designs. Solid lines denote the

ML decision thresholds, and broken lines denote the complexity-adjusted thresholds.

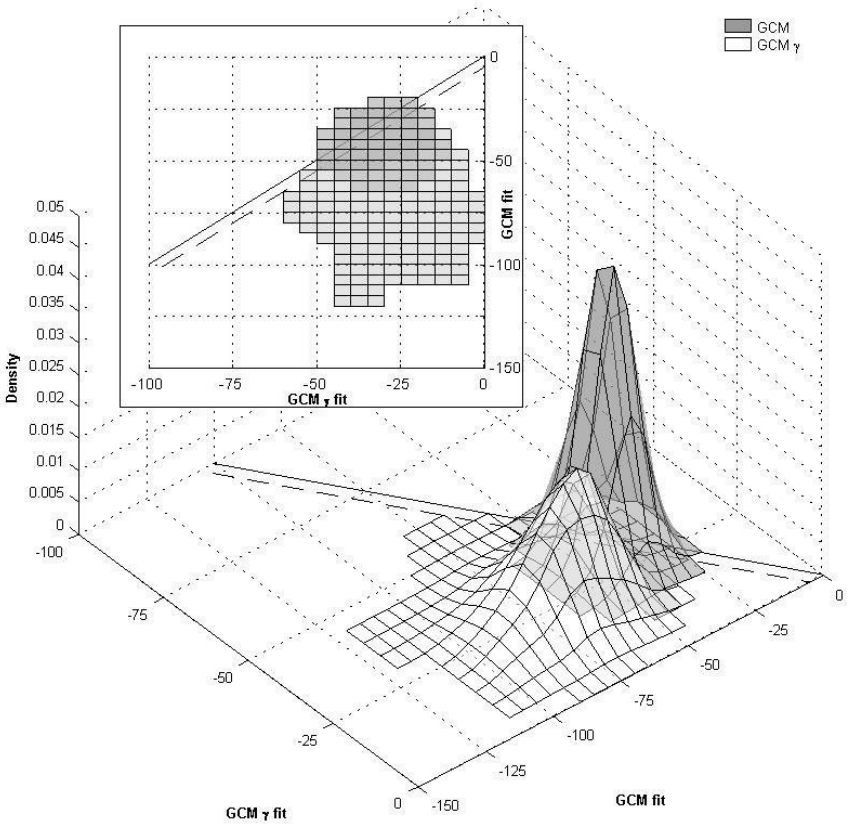


Figure 11: Representativeness plots for data sampled from GCM (in grey) and GCM-γ (in white). The solid line denotes the ML decision threshold, and the broken line denotes the complexity-adjusted threshold. The inset panel shows a top-down view of the same plots.

[1] As an aside, we note that in many cases r^2 or *d'* will be appropriate, and there is no reason in principle why the methods introduced in this paper could not accommodate them.

[2] The other two recognition curves from that paper were not used here because they introduce an extra complication to the analysis. The two recognition data sets that we have used are simple retention data and can be treated as independent near-replications. In contrast, the "remember only" and "know only" data are not independent of the "remember + know" data.

[3] Our interpretations in this section are based on the scale suggested by Jeffreys' (1961; see also Raftery 1995), referring to standards of evidence in science. The relevant qualification is that the scale originally corresponded to marginal likelihoods, not maximum likelihoods, which is the source of our "difference in numbers of parameters" disclaimer.

[4] Strictly, what is obtained is a probability density, not a probability. A probability density is the continuous version of a probability. Thus, whereas probabilities sum to 1, probability densities integrate to 1. As a result, it is possible for a probability density function to exceed 1, so long as the definite integral across any interval is less than or equal to 1.

[5] We thank John Wixted for suggesting this interpretation.

[6] Curiously, this concept is not unrelated to the idea of "active learning" in human inference (Steyvers, Tenenbaum, Wagenmakers & Blum, 2003).

[7] This number differs substantially from that reported by Pitt et al. (2002). The reason for this is that we have incorporated order constraints on the parameters, whereas Pitt et al. did not.

[8] Readers may wonder why a complexity analysis for the retention functions was not presented. The reason lies in the technical assumptions that underlie the geometric complexity measure. The measure derived by Myung et al. (2000) is a ratio of two Riemannian volumes: The volume occupied by the model in the space of probability distributions, and a small ellipsoid near the maximum likelihood parameters. The derivation implicitly requires that the extension of the model volume be much larger than the small ellipsoid in every dimension (which is always true as N becomes arbitrarily large). However, some unusual behavior of the geometric complexity for the retention functions led us to suspect that this assumption is often violated in the set of designs that we considered (see Navarro, submitted). Given this, we did not feel it was appropriate to compare complexity differences for these models. Nevertheless, for the designs with large Ns and $|t|$ (e.g. design numbers 1, 5, 14, 15, and 17), in which this problem is unlikely to be too severe, the ordering of complexity was always SE > PE > EX > HY.