Global Model Analysis by Parameter Space Partitioning

Mark A. Pitt, Daniel J. Navarro, Woojae Kim, and Jay I. Myung

Ohio State University

running head: Parameter Space Partitioning

Correspondence information:

Mark A. Pitt
Department of Psychology
1885 Neil Avenue
Ohio State University
Columbus, OH 43210-1222
office (614) 292- 4193
fax (614) 688-3984
pitt.2@osu.edu
http://lpl.psy.ohio-state.edu

Abstract

To model behavior, we need to know how models behave. This means learning what other behaviors a model can produce besides the one generated by humans in an experiment. This is a difficult problem because of the complexity of psychological models (e.g., their many parameters) and because the behavioral precision of models (e.g.,interval-scale performance) often mismatches their testable precision in experiments, where qualitative, ordinal predictions are the norm. *Parameter space partitioning* is a solution that evaluates model performance at a qualitative level. Given a definition of a qualitative data pattern, there exists a partition on the model's parameter space that divides it into regions that correspond to each data pattern. Markov-chain Monte Carlo methods are used to discover and define these regions. Three application examples, all using connectionist models, demonstrate its potential and versatility for studying the global behavior of psychological models. Among other things, one can easily assess how central and robust the empirical data pattern is to the model, as well as the range and characteristics of its other behaviors.

The experimental method of scientific inquiry has proven to work quite well in psychology. Its unique blend of methodological control and statistical inference are effective for testing qualitative (i.e., ordinal) predictions derived from theories of behavior. Data collection and dissemination have become very efficient, so much so that far more may be known about a behavioral phenomenon than is reflected in its corresponding theory.

That our knowledge extends beyond the reach of the theory may be a sign of productive science, and underscores the fact that theories are broad conceptualizations about behavior that cannot be expected to explain the minutia in data. Cognitive modeling is a research tool that can act as a counter-force to slow and fill this explanatory gap. It compensates for a theory's limitations of precision in data synthesis, description, and prediction. Whether the model is an implementation of an existing theory, or a neurally inspired environment is created in which to study processing (e.g., information integration, representational specificity, probabilistic learning), models are rich sources of ideas and information on how to think about perception and cognition. The pros and cons of various implementations can be evaluated. Inconsistencies and hidden assumptions can come to light during model creation and evaluation. In short, the modeler is forced to confront the complexity of what is being modeled, and in the process, can gain insight into the relationship between variables and the functionality of the model (see Shiffrin & Nobel, 1998 for a personal account of this process).

Of course, the virtues of modeling are accompanied by vices. One of the more serious, often leveled against connectionist models (Dawson & Shamanski, 1994; McCloskey, 1991) but by no means restricted to them, is that model behavior can be mysterious and difficult to understand, which can defeat the purpose of modeling. A model should not be as (or more) complex than the data being described. Rather, models should offer a simpler and tractable

description.

The preceding observations are not so much a comment about models or modeling per se, but a comment about the need for methods for analyzing model behavior. The computational power of cognitive models requires correspondingly sophisticated tools to study them. The wide variety of models in the discipline makes the need for universal tools all the more pressing and challenging. In this paper, we introduce such a general-purpose model analysis and comparison method in the context of connectionist models. We begin by situating it in relation to existing methods.

Methods of Model Analysis

Model analysis methods can be differentiated along two dimensions, whether they measure a model's local or global behavior, and whether model behavior is evaluated quantitatively or qualitatively. The position of many analysis methods along these, largely independent, dimensions is shown schematically in Figure 1. In the following discussion, quantitative methods are reviewed before qualitative ones.

Figure 1 here

Quantitative Model Analysis

Most quantitative evaluations of models are local. That is, they consider the behavior of the model only at its best-fitting parameter values. In contrast, global methods aim to elucidate a model's behavior across the full range of its parameter values. Note surprisingly, the two approaches are complementary in what they tell us about model behavior.

Local Quantitative Methods

The most common form of local model analysis is data fitting, in which a model is tested by measuring how closely it approximates (i.e., fits) human data. Because data are a reflection of

the psychological process under study, a good fit to the data is a necessary condition a model must satisfy to be taken seriously. A good fit determines how well a model passes the sufficiency test of mimicking human performance. It is especially useful in the early stages of model development as a quick and easy check on sufficiency. Quantitative measures of *goodness-of-fit* include percent variance accounted for, root mean square deviation, and maximum likelihood. Although a good fit makes a model a member of the class of possible contenders, this class will almost always be quite large. If two models that one is comparing fit the data similarly well, other analysis methods are needed to choose between them.

Another local method that is useful for probing model behavior more deeply than goodness-of-fit is a *sensitivity analysis,* in which a model's parameters are varied around its best-fitting values to learn how robust model behavior is to slight variations of those parameters. If a good fit reflects a fundamental property of the model, then this behavior should be stable across reasonable parameter variation. Another reason a model should satisfy this criterion is that human data are noisy. A model should not be so sensitive that its behavior changes noticeably when noise is encountered. *Cross validation*, in which a model is fit to the second of two data sets using the best fitting parameter values from fitting the first data set, is a fit-based approach to quantifying this sensitivity (Stone, 1974; Browne, 2000).

Global Quantitative Methods

A drawback of local analysis methods is the very fact that they are local. Each fit provides a view of the model's behavior at a particular point in its parameter space, but does not provide any information about how it behaves at other parameter settings. This can be particularly disconcerting if the behavior of the model is sensible only at a few settings.

Furthermore, relying on purely local methods leaves the one with a few "snapshots" of model performance that are difficult to piece together into a comprehensive understanding of the model. The task of comparing two models is even more arduous with local methods.

For these reasons, researches have begun developing global analysis techniques. They are intended to augment local methods, not replace them. Under a global view, the goal is to learn something about the full range of behaviors that a model exhibits. By doing so, we can gain a deeper understanding of the model and how it compares to competing models. Two of the most popular quantitative global methods are *Bayesian methods* (e.g., Myung & Pitt 1997; Kass & Raftery 1995) and *minimum description length* (MDL; e.g., Rissanen 1996, 2001; Grünwald 1998; Grünwald, Myung, & Pitt, in press). In both, the focus is on predictions made by the model at all of its parameter values. This global perspective yields a natural measure of a model's *a priori* data-fitting potential (i.e., the model's flexibility or complexity in fitting future, unseen data sets; Myung 2000). Although both methods are statistically rigorous, technical requirements currently limit their application to the diverse range of models in psychology.

More recently, Navarro et al (2004; Kim, Navarro, Pitt, & Myung, 2004; see also Wagenmakers, Ratcliff, Gomez, & Iverson, 2004) introduced a global quantitative method, called *landscaping,* that was developed with an eye toward increased versatility and informativeness about model discriminability. The essence of the technique involves determining how well two models fit each other's data. What is being measured is the extent to which two models mimic one another. The approach is attractive since landscapes are relatively easy to create, requiring only a comparison of fits to data sets. In addition, landscapes can be used to assess the informativeness of data in distinguishing pairs of models by inspecting where within these landscapes experimental data are located. In short, the landscape provides a global

perspective from which to understand the relationship between two models and their fits to data. However, like MDL and Bayesian methods, landscaping requires that the models make quantitative predictions.

<div align="center">Qualitative Methods</div>

Implicit in the use of quantitative methods of model analysis is that the goal of modeling is to approximate closely empirical data. Although there are good reasons for doing so, one must be careful to avoid modeling the quantitative minutiae of a data set while missing theoretically important qualitative properties or trends in the data. In the words of Box (1976, p. 792) "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". Thus, it is important not to lose sight of qualitative behavior. In this regard, model evaluation at the qualitative level can not only be informative, but sometimes more appropriate. It is certainly the most common in the psychological sciences.

Hypothesis Testing as A Local Qualitative Method

Hypothesis testing as used in the experimental method of psychological inquiry is in essence a local qualitative method of model analysis. Hypotheses generated in experimental settings are qualitative predictions about an ordinal pattern of data across conditions. One might, for instance, observe a preference reversal under some experimental conditions that only one of two models is able to reproduce. A correct prediction is generally held to be strong evidence in favor of the winning model, and not without reason (Platt, 1964). Only in rare circumstances will a model predict the exact (interval) quantitative differences among conditions. As such, most psychological models are often intended to be illustrative of some underlying process rather than a precise description of its inner workings. Accordingly, the failure of a model to reproduce the

"fine grain" of a data set is not necessarily fatal to the model. It may merely indicate that minor

changes are required. On the other hand, if a model cannot capture the gross qualitative pattern

of the data, something is seriously amiss. As with all local methods of model analysis,

hypothesis testing is most informative when the qualitative pattern in the data can be captured by

only one model.

<u>Global Qualitative Methods</u>

Hypothesis testing is the work-horse of model evaluation in much of psychology, but like

its quantitative counterpart, goodness-of-fit, its focus on local behavior is a limitation. We gain

only a glimpse of what the model can do. It be useful to know how many of the other qualitative

patterns in that same experiment the model could elicit. A model that can produce any logically-

possible pattern is no more impressive than one that fails to produce the empirical pattern

(Roberts & Pashler, 2000). In other words, there is the implicit problem pertaining to that we

might call "qualitative complexity or flexibility." The solution is the same as with quantitative

methods: Study the qualitative behavior of the model across a broad range of parameter values.

Dunn and James' (2003) *signed difference analysis* is an example of global qualitative

model analysis. One seeks to identify all of the signed difference vectors between two arbitrary

points in data space that a model allows by varying its parameter values. It is a simple means of

deriving testable ordinal predictions from models in which the functional relationship between

task performance measurements (e.g., dependent variables) and underlying constructs (i.e.,

model parameters) is assumed to be monotonic and otherwise unspecified. This method of global

analysis is probably most useful in the early stages of cognitive modeling where models are

defined primarily in terms of their qualitative predictions, and no or few assumptions are made

about other details of the underlying process. As models get more refined with elaborate

relationships between dependent variables and model parameters, a more sophisticated method is needed.

The purpose of the current paper is to introduce a general-purpose and highly informative method of global qualitative model analysis, *parameter space partitioning* (PSP). It involves doing exactly what the name implies: A model's parameter space is literally partitioned into regions that correspond to qualitatively different data patterns that the model could generate in an experiment. Study of these regions can reveal a great deal about the model and its behavior, as the three application examples below will show. In particular, one can easily learn how representative human performance is of the model, as well as the characteristics of other behaviors the model exhibits. The popularity and complexity of connectionist models made them an ideal class of models in which to demonstrate the potential of PSP for studying model behavior.

Parameter Space Partitioning: A Global Qualitative Method

A simple example illustrates the gist of PSP. Consider a visual word recognition experiment in which participants are asked to categorize stimuli as words or nonwords. The mean response time to words is then measured as the dependent variable across three experimental conditions, *A*, *B* and *C*. In this situation, it may be reasonable to claim that the important theoretical property is the ordinal relationship (fastest to slowest) across conditions. In this case, there are 13 possible orderings (including equalities) that can be observed across the three conditions (e.g., $A > B > C$, $A > B = C$, $B > C = A$, etc). Each of these orderings defines a *qualitative data pattern*. Suppose further that mean participant performance yielded the pattern $B > C > A$.

Now consider two hypothetical models, M_1 and M_2, of word recognition, each with two parameters. Using PSP, we can answer the following questions about the relationship between the models and the empirical data generated in the experiment: How many of the thirteen data patterns can each model produce? What part of the parameter space includes the empirical pattern? How much of the space is occupied by the empirical pattern? What data patterns are found in nearby regions as well as the rest of the parameter space?

Figure 2 shows the parameter space of each model partitioned into the data patterns it can generate. Model M_1 produces three, one of which is the empirical pattern. Note how it is central to model performance, occupying the largest portion of the parameter space. Even though the model generates two other patterns, they are smaller and differ only minimally from the empirical pattern. In contrast, M_2 produces nine of the thirteen patterns. Although one is the empirical pattern, M_2's performance is not impressive because it can mimic almost any pattern that could possibly have been observed in the experiment. Indeed, the fact that M_2 can mimic human performance seems almost incidental. Not only does the empirical pattern occupy a small region of the parameter space but larger regions are produced by patterns that do not human-like (e.g., $C>A>B$).

Figure 2 here

PSP can also be considered an extension of hypothesis testing. Instead of looking to see if the model captures an empirically observed pattern, we look at all of the patterns that a model could potentially capture to learn what else the model can do. The idea of looking for patterns across the entire parameter space is not new. For example, most recently Johansen and Palmeri (2002) used a grid search to look for predictions made by categorization models. What is new about PSP is the approach taken to solving the search problem, which overcomes the limitations

of grid search. With PSP, it possible to partition high-dimensional models with good accuracy in a reasonable amount of time, something that grid search cannot do.

<div align="center">Implementing PSP</div>

Implementation of PSP requires solving two non-trivial problems. One is how to define a data pattern, and the other is how to devise an efficient search algorithm to find the data patterns. Each set of a model's parameter values generates a model output, but not every model output is a distinct data pattern. How many qualitatively "different" outputs can a model produce? Answering this question, which is what PSP enables us to do, depends critically on how a data pattern is defined. This is something which will vary from experiment to experiment, and indeed may vary within an experiment depending upon what a researcher wants to learn. Although there is no general solution, the scientist usually knows what patterns should be found in order to support or falsify a model. Most of the time, ordinal predictions are being tested. In this case, a "natural" definition of a data pattern is the ordinal relationship of model outputs, as in the above example. Nevertheless, it is a good idea to try out a couple of different definitions and to perform sensitivity analyses to ascertain if and to what extent conclusions obtained under one definition hold across others. An example of this process is presented later in the paper.

The Search Algorithm

Once a data pattern is defined, the next challenge is to find all patterns a model can simulate. The set of all data-patterns forms a partition on the parameter space, in the sense that each point in the space can correspond to only one pattern, however improbable. Once we have a definition of a data pattern, we have in effect created an unknown partition on the parameter space. Accordingly, the problem to be solved is to search a multi-dimensional parameter space in such a way that we visit each part of the partition at least once. As the number of parameters

in a model increases, the space becomes higher dimensional, and the search problem can become very hard. The consequence of this is that brute force search methods to find all regions, such as grid search or a random search procedure like Simple Monte Carlo (SMC) will not work, or take far too long to succeed. Markov Chain Monte Carlo (MCMC; Gilk et al, 1996) is a much more sophisticated sampling method that we encorporated into an algorithm that efficiently finds all regions.

Application of the PSP algorithm begins with a starting set of parameter values at which the model can generate a valid data pattern (i.e., one that satisfies the definition). This initial set can be supplied by the modeler or from an exploratory run using SMC. Given the parameter set and the corresponding data pattern generated by the model, the algorithm samples nearby points in the parameter space to map the region that defines the data pattern. MCMC is used to approximate quickly and accurately the shape of this region. The process then begins anew by sampling a nearby point just outside of this region.

Figure 3 illustrates how the algorithm works in the space of a two-parameter model. The process begins with the initial parameter set serving as the current point in the parameter space (filled point in panel a). A *candidate* sample point (shaded point) is drawn from a small, predefined region, called a jumping distribution, centered at the current point. The model is then run with the candidate parameter values and its output is evaluated to determine if the data pattern is the same as that generated by the initial point. If so, the candidate point is accepted as the next point from which another candidate point is drawn. If the new candidate point does not yield the same data pattern as the initial one, it is rejected as belonging to the current region. Another jump from the initial point is attempted, accepting those points that yield the same data pattern. The sequence of all accepted points recorded across all trials is called the *Markov chain*

corresponding to the current data pattern. This sample of points is used to estimate the size of the

region occupied by the data pattern (panel b). The theory of MCMC guarantees that the sample

of accepted points will eventually be distributed uniformly over the region. This feature of

MCMC allows us to estimate the volume occupied by the region, regardless of its size. A

region's size is estimated by calculating the volume of a multi-dimensional ellipsoid that is

computed using the mean and the variance-covariance matrix of the sample.

Figure 3 here

Every rejected point (shaded point in panel b), which must be outside the current region,

is checked to see if it generates a new valid data pattern. If so, a new Markov chain

corresponding to the newly discovered pattern is started to define the new region. In effect,

accepted points are used to shift the jumping distribution around inside the current region to map

it completely, whereas rejected ones are used initiate new MCMC search processes to map new

regions. Over time, as many search processes as there are unique data patterns will be run.

Additional details about the algorithm are described in Appendix A.

Algorithm Evaluation

We tested the accuracy of the PSP algorithm by measuring its ability to find all of the

data patterns defined for a particular model. The difficulty of the search problem was varied by

manipulating the definition of a pattern (i.e., number of patterns) and the number of parameters

in the model. The extent of both (see Table 1) was deliberately made large to make the test

challenging. The efficiency of the algorithm was measured by comparing its performance to

SMC (random search).

The model was a hypercube whose dimensionality d (i.e., number of parameters) was 5,

10, or 15. To illustrate the evaluation method, a two-dimensional model ($d=2$) is depicted in

Figure 4 that contains twenty data regions (outlined in bold) that the algorithm had to find. Note that a large portion of the space does not produce any valid (i.e., the set of 20) data patterns. Also note that the sizes of the data regions vary a great deal. Some are elicited by a wide range of parameter values whereas others can be produced only by a small range of values. This contrast grows exponentially as the dimensionality of the model increases, and was purposefully introduced into the test to make the search difficult and approximate the complexities (i.e., nonlinearities) in cognitive models. Ten independent runs of each search method were carried out to assess the reliability of algorithm performance.

Figures 4 and 5 here

Figure 5 shows performance of the PSP algorithm for a search problem in which there were one hundred regions embedded in a 10-dimensional hypercube (d=10). The PSP algorithm found all regions and did so in nine minutes. SMC found only about 23 patterns in nine minutes, and given its sluggish performance, it seems doubtful that SMC would find all of them in anything close to a reasonable amount of time.

Table 1 summarizes results from the complete test. The mean proportion of patterns found is listed in each cell. Results are clear and consistent. The PSP algorithm almost always found all of the patterns whereas SMC failed to do so in every condition. Most noteworthy is the success of the PSP algorithm in the toughest situation, when there were 15 parameters and 500 data patterns. Its near perfect success suggests it is likely to perform admirably in other testing situations. In the remainder of this paper, we describe its applications to analyzing the behavior of a single model and to comparing design differences between models.

Table 1 here

Application 1: Evaluating the Qualitative Performance of ALCOVE

In standard model-fitting analyses, a model's ability to fit (or simulate) the data is taken as evidence that it approximates the underlying cognitive process. In a PSP analysis, the definition of "fit" is relaxed to be a qualitative, ordinal relation on the same scale as the experimental predictions themselves. Model performance is then evaluated by determining how many of the possible orderings of conditions it can produce and how central is the empirical pattern among them (Figure 2). In this section, we examined the behavior of ALCOVE (Kruschke 1992), a highly successful exemplar-based account of human category learning, in the context of the seminal Shepard, Hovland, and Jenkins (1961) experiment. While there are some category learning effects that it does not capture without extension or modification (e.g, Kruschke & Erikson 1995, Lee & Navarro 2002), ALCOVE remains a simple and powerful account of a broad range of phenomena.

<div align="center">Background to the Analysis</div>

The ALCOVE Model

In some category learning experiments, participants are shown a sequence of stimuli, each of which possesses some unknown category label. The task is to learn which labels go with which stimuli, using the feedback provided after responses are made. ALCOVE solves this problem in the following way (for a detailed description, see Kruschke 1992). When stimulus I is presented to ALCOVE, its similarity to each of the previously stored exemplars, s_{ij}, is calculated. Following Shepard (1987), similarity is assumed to decay exponentially (with a width parameter c) as a function of the attention-weighted city-block distance between the two stimuli in an appropriate psychological space. After estimating these similarities, ALCOVE forms response strengths for each of the possible categories. These are calculated using associative weights maintained between each of the stimuli and the categories. The probability

of choosing the *k*-th category follows the choice rule (Luce, 1963) with parameter ϕ.

Having produced probabilities for each of the various possible categorization responses, ALCOVE is provided with feedback from an external source. This takes the form of a "humble teacher" vector, in which learning is required only in cases where the wrong response was made. Two learning rules are then applied, both derived by seeking to minimize the sum-squared error between the response strengths and the teaching vector, using a simple gradient descent approach to optimization. Using these rules, ALCOVE updates the associative weights (with parameter η_w for the learning rate) and the attention weights (with a learning rate parameter η_a) prior to observing the next stimulus.

The Shepard, Hovland and Jenkins Task

In a classic experiment, Shepard et al. (1961) studied human performance in a category learning task involving eight stimuli divided evenly between two categories. The stimuli were generated by varying exhaustively three binary dimensions such as color (black vs. white), size (small vs. large) and shape (square vs. triangle). They observed that, if these dimensions are regarded as interchangeable, there are only six possible category structures across the stimulus set, illustrated in Figure 6a. This means, for example, that the category structure that divides all squares into one category, and all triangles into the other is regarded as equivalent to the category structure that divides small shapes from large ones, as shown in the lower right.

Empirically, Shepard et al. (1961) found robust differences in the way in which each of the six fundamental category structures was learned. In particular, by measuring the mean number of errors made by subjects in learning each type of category structure, they found that Type I was learned more easily than Type II, which in turn was learned more easily than Types III, IV and V (which all had similar error measures), and that Type VI was the most difficult to

learn. More recently, Nosofsky, Gluck, Palmeri, McKinley and Glauthier (1994), replicated

Shepard et al.'s (1961) task using many more subjects, and reported detailed information relating

to the learning curves. Figure 6b shows the mean proportion of errors for each category type.

Consistent with the conclusions originally drawn by Shepard et al. (1961), it is generally held

that the theoretically important qualitative trend in these data is the finding that there is a natural

ordering on these curves, namely that I<II<(III, IV, V)<VI. This kind of pattern is called a weak

order, since the possibility of ties is allowed.

Figure 6 here

The psychological importance of this weak order structure is substantial. Suppose we had

two models of the category learning process, M_1 and M_2, of roughly equal complexity. M_1

provides a reasonably good quantitative fit to the data, by assuming that all types are learned at

the same rate, and closely approximates the average of the six empirical curves. In contrast, M_2

reproduces the ordering I<II<(III, IV, V)<VI, but learns far too slowly and, as a result, fits the

data much worse than M_1. Since the models are of equivalent complexity, a classical model

selection analysis would prefer model M_1. However, while clearly both models have some flaws,

most psychologists would prefer M_2, because it captures the *theoretically relevant* property of

the data. This discrepancy arises because statistical criteria like MDL tend to assume that all

properties of the data are equally relevant. In many psychological applications, this is not the

case. In what follows, we assume that the weak order structure is the important theoretical

property of the empirical data, and use the PSP method to ask how effectively ALCOVE

captures this structure.

The PSP Analysis

As with any parameterized model, ALCOVE makes different predictions at different

parameter values. When applied to the Shepard et al. (1961) task, ALCOVE will sometimes produce curves that have the same qualitative ordering as the empirical data, but at other times they will look quite different. It would be nice to know something about the *other* orders that ALCOVE can produce, since it seems that we might learn something about the model itself. A PSP analysis can provide such information. Using the "weak order" definition of a pattern of curves, there are 4,683 different data patterns that a model could produce. One would hope that ALCOVE generates only a small proportion of these, and that the extra patterns it does produce are interpretable in terms of human performance.

Preliminaries

We now have a model, a data set, and an intuitive definition of a data pattern. To perform PSP analyses, a formal method of associating a set of learning curves with a particular pattern must be defined. The judgement that I<II<(III, IV, V)<VI is the appropriate empirical pattern has generally been based on visual inspection of the curves. It is possible to be more precise about this, allowing us to associate uniquely a set of learning curves with a qualitative ordering to yield a data pattern. The details of this procedure, which is essentially a clustering analysis, are provided in Appendix B.

For the PSP analysis of ALCOVE, we constrained the parameter vectors (c, ϕ, η_w, η_a) to lie between (0, 0, 0, 0) and (20, 6, 0.2, 0.2), and disallowed any parameter combination that did not produce monotonic curves. A technical complication is introduced by the fact that ALCOVE's predictions are slightly dependent on the order in which stimuli are observed. Each stimulus has a different effect on ALCOVE, so variations in order of presentation produce slight perturbations in the response curves. However, even these minor perturbations can violate the continuity assumptions that underlie the PSP algorithm. In order to deal with these order effects,

we chose 20 random stimulus orders, and ran the PSP algorithm 10 times for each stimulus

order, yielding a total of 200 runs. As it turns out, the important properties of ALCOVE appear

to be invariant under stimulus reordering, but some unimportant properties are not.

How Many Data Patterns can ALCOVE Produce?

After running the PSP algorithm 200 times, we observed that each run produced a

different number of patterns, ranging from a minimum of 32 to a maximum of 122. Although

this range is substantial, it reflects an inherent variability in ALCOVE more so than the PSP

algorithm itself. The mean number of patterns recovered for a particular stimulus order ranged

from 46.5 to 102.8, while the range in the number of patterns recovered within an order was

minimal: The smallest range was a mere 7 patterns, while the largest was 36. Moreover, there

was an important amount of redundancy across the 200 runs, with 17 patterns being found on

every occasion, which included the empirical pattern I<II<(III,IV,V)<VI. We will refer to these

17 patterns as "universal" patterns, and the other 183 patterns as "particular" patterns.

Compared to the set of all 4683 possible data patterns, even the largest count of

17+183=200 patterns encompassed by ALCOVE is quite a small number. In a sense, this is quite

a success for the model, because the empirical pattern suddenly looks far less unlikely if we

assume humans do something rather ALCOVE-like. Even in the scenario where we allow *all*

200 recovered patterns to be treated as a genuine ALCOVE prediction, the empirical pattern is

one pattern in 200, rising from the much less satisfying base rate of 1 in 4683. If the substantive

predictions are restricted to the set of universal patterns, the empirical pattern is now 1 in 17.

Either way, ALCOVE provides a reasonably good qualitative account of these data.

This initial analysis demonstrates that ALCOVE passes a basic sufficiency test, in that it

can account for the qualitative structure of the observed data without "going overboard" and

producing every possible pattern. Of course, as psychologists, we are interested in more than just

how many patterns ALCOVE produces, and this analysis is just the tip of the iceberg as far as

what can be learned from PSP. For example, it would be useful to know what kinds of category-

type orderings are generally preserved across all 200 data patterns. One crude method for

determining this is to find the average position (i.e., its rank among the six curves) of each

category type across all patterns. This is illustrated in Figure 7, which plots the mean rank for

each of the six types across all patterns for both universal and particular patterns. It is clear that

rank tends to increase as the index of the type increases. The main difference between the two

types of patterns is that the universals do not really distinguish between Types III and IV,

whereas the particulars do. Nevertheless, it seems to be the case that, on average, rank does not

decrease with index. This is encouraging, because both empirically and algebraically, the

difficulty of the task either increases or stays constant as the index increases (see Feldman,

2000). This analysis demonstrates that on average, the set of 200 patterns that make up

ALCOVE's entire set of qualitative predictions roughly preserve an important property of the

empirical data: A monotonic increase in learning difficulty across category type.

Figure 7 here

Which Patterns Matter?

The fact that some patterns are frequent (universals) and others rare (particulars) suggests

that some may be more representative of model behavior than others. However, the frequency

with which a pattern is found is an unsatisfying definition of its importance to a model's

behavior, since it confounds the model properties with the robustness of the search algorithm. It

was therefore necessary to make some assumptions that allow us to identify the major patterns

that are responsible for most of ALCOVE's behavior. If we accept the notion that ALCOVE's

parameters are interpretable and psychologically well-founded (see Kruschke 1993), then it makes sense to treat the parameter space itself as an important source of information about the model. Specifically, if a pattern can be produced only within a tiny region in the parameter space, then it is probably safe to dismiss it as irrelevant to the model. Using this information, we can estimate the proportion of the parameter space that is taken up by each pattern, and then use these quantities to identify the most prevalent and representative patterns.

Note that the outcome of such an analysis depends on the manner in which ALCOVE's parameters are formalized. In statistical terms, the conclusions are not invariant under an arbitrary nonlinear transformation, but will be invariant in any positively-scaled transformation. This is not necessarily a bad thing, so long as we have some principled reason for using the current parameterization. Arguably, the nature of the exemplar theory on which ALCOVE is based, and the manner in which the model captures Kruschke's (1993) "three principles", provide exactly this kind of justification.

As it turns out, there is a strong relationship between the size (i.e., volume) of the region occupied by a pattern and the frequency with which it was discovered across the 200 runs. In Figure 8, the log of the average volume for each of the 200 patterns discovered is plotted against the frequency with which they were discovered across the 200 runs of the algorithm. Patterns shown on the far left are the ultimate particulars, having been discovered only once, while patterns on the far right are the universals, having been discovered on every occasion. Noting that the scale is logarithmic, we observe that the universals are by far the largest patterns. The empirical pattern, indicated by the circle, is one of the larger patterns, and is a universal.

Figure 8 here

The data in Figure 8 let us refine the answer to the question, "To what extent does

ALCOVE *predict* the empirical pattern?" The fact that the empirical pattern is among the 17

universals is encouraging, as is the regularity suggested by Figure 7. However, by considering

the size of the various regions, we can take this analysis a step further. We could, for instance,

exclude all patterns that do not reach some minimum average size. This approach is illustrated

by the horizontal threshold shown in Figure 8: Only patterns that occupy more than 1% of the

parameter space on average lie above this line. This is a pretty stringent test, given that the

parameter space is four-dimensional. Indeed, the empirical pattern occupies only about 2% of the

space.

In total, only twelve patterns (all universals) occupy more than 1% of the space, as shown

in Table 2, and some general properties of ALCOVE's behavior emerge when examined

together. Looking across patterns, it is clear that Types III and IV are always (12 of 12) predicted

to be learned at about the same rate, and Type V is usually (10 of 12) also about the same. Type

VI, on the other hand, is mostly learned slower than III, IV and V (7 of 12). Type I is usually (9

of 12) faster than III-VI, as is Type II (8 of 12). So, not only is the empirically-observed pattern

I<II<(III,IV,V)<VI among the largest patterns (it is the eighth largest), but the other large

patterns generally preserve the pairwise relations found in the empirical data. They are, in short,

"close" to the empirical pattern. The exception to this claim regards the relationship between

Types I and II. Their ordering is ambiguous. It might be I<II (5 of 12), I=II (4 of 12), or even

II<I (3 of 12). In this case, ALCOVE does not make a strong qualitative prediction.

Table 2 here

<u>Summary</u>

The PSP analyses confirm a number of well-known properties of ALCOVE. Few

category learning researchers would be surprised to hear that ALCOVE captures the

I<II<(III,IV,V)<VI ordering in a robust manner, or that the various pairwise relations that the

ordering implies are almost always satisfied. What the current analyses add to this knowledge

base is how central the empirical pattern is to the model, that the dominant alternative patterns

are quite similar to the empirical one, and that "distant" patterns (i.e., violations of weak

orderings) are rarely if ever generated by ALCOVE. The added understanding provided by this

global analysis of ALCOVE increases one's confidence in claiming that the model "accounts"

for the data precisely because we know how the range of behaviors the model exhibits in this

experimental setting.

The PSP analysis also revealed some unexpected behaviors. It is somewhat surprising

that it is even possible for ALCOVE to predict II<I. It may be that this happens only at odd

choices of parameters, and is certainly something that would be interesting to look into in future

work.

Application 2: Comparing the Architectures of Merge and TRACE

In addition to learning about the behavior of a single model, PSP analyses can inform us

about the behavioral consequences of design differences between models. In this and the next

example, PSP is applied to two localist connectionist models of speech perception, TRACE

(McClelland & Elman, 1986) and Merge (Norris, McQueen, & Cutler, 2000). We compared

them in two experimental settings, one intended to bring out architectural differences, and the

other differences in weighting bottom-up (sensory) information.

Background to the Analysis

Architectural Differences of TRACE and Merge

The two models are illustrated schematically in Figure 9. They are similar in many ways.

Both have a phonemic input stage and a word stage. There are excitatory connections from the

phoneme input to the word stage, and inhibitory connections within the word stage. They differ

in how prior knowledge is combined with phonemic input to yield a phonemic decision

(percept). In TRACE, word (prior) knowledge can directly affect sensory processing of

phonemes. This is represented by direct excitatory connections from the word stage back down

to the phoneme stage. Also note that in TRACE the phoneme stages performs double duty, also

serving as a phoneme decision stage. In Merge, these two duties are purposefully separated into

two distinct stages to prevent word information from affecting sensory registration of phonemes.

Instead, lexical knowledge affects phoneme identification via excitatory connections from the

word to the phoneme-decision stage. In contrast to the direct interaction between phoneme and

word levels in TRACE, these two sources of information are integrated at a later decision stage

in Merge.

Figure 9 here

Although not visible in the diagrams in Figure 9, the models differ in another important

way. In keeping with the belief that bottom-up (sensory) information takes priority in guiding

perception early in processing, activation of a phoneme decision node in Merge must be initiated

by phoneme input *before* excitatory lexical activation can affect phoneme decision making.

TRACE contains no such constraint.

The goal of our investigations was to assess the impact of these two design differences on

global model behavior. Norris et al (2000) proposed Merge as an alternative to TRACE because

they felt the evidence from the experimental literature did not warrant interaction (i.e., direct

word-to-phoneme feedback). The adequacy of the Merge architecture was demonstrated in

simulation tests in which it performed just as well as, if not slightly better than, TRACE in

reproducing key experimental findings. PSP analyses of the models' behaviors were performed

in two of these experimental settings. In the first, the subcategorical mismatch study by Marslen-Wilson and Warren (1994), the consequences of splitting phoneme processing into separate input and decision stages was evaluated. In the second, the indirect inhibition experiment of Frauenfelder, Segui, and Dijkstra (1990), the contribution of the bottom-up priority rule to model performance was examined.

In order to compare the two models, it was necessary to equate them in every way possible to ensure that differences in performance were attributable only to design differences, not other factors, such as the size of the lexicon. In essence, we wanted to compare the fundamental structural and functional properties that define the models. We did this by first implementing the version of Merge described in Norris et al (2000), and then making the necessary changes to Merge to turn it into TRACE. In the end TRACE required four fewer parameters than MERGE. The names of the parameters, along with other model details, are in Appendix C.

The Subcategorical Mismatch Experiment

The first comparison of the two models was performed in the context of the subcategorical mismatch experiment of Marslen-Wilson and Warren (1994, Experiment 1; McQueen, Norris, & Cutler, 1999, Experiment 3). The setup is attractive because of the large number of conditions and response alternatives, which together permitted detailed analyses of model behavior at both the phonemic and lexical levels. In the experiment, listeners heard one-syllable utterances and then had to classify them as words or nonwords (lexical decision task) or categorize the final phoneme (phonemic decision task). The stimuli were made by appending a phoneme (e.g., /b/ or /z/) that was excised from the end of a word (e.g., *job*) or nonword (e.g., *joz*) to three preceding contexts, to yield six experimental conditions (listed in Table 3). The first

context was a new token of those same items but with the final consonant removed (e.g., *jo*), to create cross-spliced versions of *job* and *joz*. The second consisted of equivalent stretches of speech from a word that differed only in the final consonant (e.g., *jo* from *jog* in both cases). The third was the same as the second except that the initial parts were excised from two nonwords (e.g., *jo* from *jod* and *jo* from *jov*).

Table 3 here

Because cues to phoneme identity overlap in time (due to coarticulation in speech), a consequence of cross-splicing is that cues to the identity of the final consonant will conflict when the first word ends in a consonant different from the second. For example, *jo* from *jog* contains cues to /g/ at the end of the vowel, which will be present in the resulting stimulus when combined with the /b/ from *job* (Condition 2 in Table 3).

Marslen-Wilson and Warren (1994) were interested in how such stimuli affect phoneme and lexical processing. As the results in Table 3 show (taken from McQueen et al, 1999), in the lexical decision task, reaction times slowed when listeners heard cross-spliced stimuli, but responding was not affected by the source of the conflicting cues (i.e., equivalent RTs in Conditions 2 and 3). Phoneme categorization, in contrast, was sensitive to the subtle variation in phonetic detail, but only when the stimulus itself formed a nonword (e.g., *joz*; Conditions 5 and 6). Importantly, the use of cross-spliced stimuli nullifies the bottom-up priority rule in Merge, which otherwise might have contributed to any differences (see Norris et al, 2000, for details). To the extent that differences are found across models, they are probably a result of their different architectures.

The PSP Analysis

The subcategorical mismatch experiment contains two dependent measures of

performance, classification decisions and the speed with which those decision were made

(response time). We treated classification performance as a qualitative pattern in the PSP

analysis. We then performed a separate investigation of the RT predictions.

<u>Preliminaries</u>

Classification in these two models is usually defined in terms of the activation state of the

network when specific decision criteria are met, such as a phoneme node exceeding an activation

threshold. Because there are multiple conditions in the subcategorical mismatch experiment, a

data pattern is really a profile of classifications across these conditions. In this experiment there

are six conditions, with a phoneme and lexical response in each, for a total of 12 categorization

responses that together yield a single data pattern. With four possible phoneme responses and

three possible lexical responses, there were a total of 2,985,984 (4^6 x 3^6) patterns. Of interest is

how many and which of these patterns TRACE and Merge produce.

We applied two classification decision rules. The effect of these rules is that they

establish the necessary mapping between the continuous space of network states to the discrete

space of data patterns, making it possible to associate each data pattern with a region in

parameter space. Because any one rule could yield a distorted view of model performance, the

use of two rules enabled us to assess the generality of the results. In addition, we found that

some model properties that are not evident with one criterion emerged when performance was

compared across rules. The first rule, labeled *weak threshold*, was a fixed activation threshold,

with values of 0.4 for phoneme nodes and 0.2 for lexical nodes. It is the same rule used by Norris

et al (2000), and was adopted to maintain continuity across studies.

The second rule, called the *stringent threshold*, required classification to be more

decisive, by requiring the activation level of competing nodes to be significantly lower than the

winning node. Two thresholds were used, the higher of which was the lower bound for the

chosen node and the lower of which was the upper bound for the nearest competitor. These

values were 0.45 and 0.25 for phoneme classification, and 0.25 and 0.15 for lexical decision.

Two other constraints were also enforced as part of the stringent threshold. There had to be a

minimum difference in activation between the winning node and its closest competitor of 0.3 for

phoneme classification and 0.15 for lexical decision. Finally, for nonword responses in lexical

decision, the difference in activation between the two lexical items, *jog* and *job*, could not be

more than 0.1.

The models were designed as depicted in Figure 9. The only lexical nodes were *job* and

jog. All necessary phoneme nodes were included, with /b/, /z/, /g/, /v/ being of primary interest.

The PSP algorithm was run for both models and both decision rules. To obtain a data pattern, all

six stimuli were fed into the model and both phoneme and lexical classification responses

assessed. The algorithm and models were run in Matlab on a Pentium IV computer. The time

required to find all patterns varied greatly, taking as little as 22 minutes (TRACE, stringent

threshold) and as long as 24 hours (Merge, stringent threshold). The consistency of results was

ascertained using five multiple runs for each model and threshold.

How Many Patterns do the Models Produce?

The analysis began by comparing classification (asymptotic) performance of the two

models. Table 4 contains a number of measures that define their relationship under the weak and

stringent thresholds. The second column contains venn diagrams that depict the similarity

relation between the models when measured in terms of the number of data patterns each can

generate. Looking first at the weak threshold data, both models generate 22 common patterns,

with TRACE producing only a few unique patterns compared to Merge (3 vs 29). The filled dot

represents the empirical pattern, which both models produced.

The nature of the overlap in the diagram indicates that TRACE is virtually nested within Merge, with 22 of its 25 patterns also being ones that Merge produces. However, Merge can produce an extra 29 patterns, suggesting that in the subcategorical mismatch design, Merge is more flexible (i.e., complex) than TRACE[1]. That said, the difference between them is it not all that great, although it does not disappear under further scrutiny (see below). Perhaps most impressively, both models are highly constrained in their performance, generating fewer than 60 of the some 3 million patterns that are possible in the design.

When the stringent threshold is imposed, both models generate fewer patterns, with TRACE producing 7 and Merge 32, as indicated in the lower half of Table 4. Nevertheless, the relationship between the models remains unchanged: TRACE is almost nested within Merge. However, the one unique TRACE pattern turns out to be the empirical data, which Merge no longer generates.

Table 4 here

Interestingly, the change in threshold primarily caused a drop in the number of shared patterns, indicating that the models are more distinct under the stringent threshold. To understand why, turn your attention to columns 3 and 4, which contain estimates of the proportion of the parameter space occupied by the valid data patterns. Far less of TRACE's parameter space is used than Merge's. Predictably, the regions shrink when the stringent threshold is applied, although the shrinkage is much more dramatic for Merge than TRACE. If one then examines how much of each valid volume is occupied by common and unique patterns, the reason for the increased distinctiveness presents itself. For TRACE (column 5), this region is occupied almost entirely by the 22 common patterns (99%). Under the stringent threshold, this

value drops slightly to 84%, but because there is only one unique pattern in this case, its value must be 16%, the size of the region occupied by the empirical pattern. The remaining 15 common patterns occupied such tiny regions in TRACE's parameter space under the weak threshold that application of the stringent threshold eliminated them. A similar situation occurred with Merge (column 6), with 12 of the 16 common patterns (of which the empirical pattern was one) disappearing due to a change in threshold. Four became unique to Merge. A few patterns unique to each model also failed to satisfy the stringent threshold (3 for TRACE and 7 for Merge).

Although a change in threshold brings out differences in the models, analysis of their common patterns under both thresholds reveals an impressive degree of similarity. For example, under the weak threshold, the regions in parameter space of all 22 patterns are comparable in size across models. To measure this, we correlated the rank orderings (from smallest to largest) of the volumes in the two models. Use of the actual volume estimates themselves is inappropriate because of differences in model structure and parameterization. With $\rho = .79$, the correlation is high. For the stringent threshold it is even higher, $\rho = 1.0$, but keep in mind that there were only six data points in this analysis. Such strong associations indicates that the overlap between models is not just nominal in terms of shared data patterns, but those regions are similar in relative size with all other common regions, making them functionally highly similar.

Is the Variation Across Patterns Sensible?

To the extent that a model produces data patterns other than the empirical one, a mark of a good model is for its performance to degrade gracefully. In the first application of PSP, ALCOVE's performance degraded gracefully because all of its patterns looked similar to the

empirical one. In the current analysis, we measured deviation by counting the number of

mismatches (different decisions) between a particular pattern and the empirical one. Since a

pattern consists of 12 decisions, there are a maximum of 12 possible mismatches (6 phonemic

and 6 lexical). Histograms of the mismatch counts were created for both models, and are shown

in Figure 10 alongside a histogram showing the base rate at each mismatch distance (e.g., there

are thousands of possible patterns with six mismatches, but only 12 patterns with one mismatch).

Figure 10 here

Both models show a remarkable proclivity to produce human-like data. The distributions

are positioned near the zero endpoint (empirical pattern) with peaks between two and four

mismatches. The probability is virtually zero that a random model (i.e., a random sample of

patterns) would display such a low mismatch frequency. The weak threshold distributions are

slightly to the right of stringent threshold distributions, and Merge distributions are slightly to

the right of TRACE's distributions. In both cases, this is probably a base rate effect; allowing

more patterns increases the likelihood of more mismatches[2].

An equally desirable property of a model is that patterns that mismatch across many

conditions occupy a much smaller region in the parameter space than those that mismatch by

only one or two conditions. That is, larger regions should correspond to patterns that are more

similar to the empirical pattern. This is generally true for both models, and to a similar extent.

When region volume is correlated with mismatch distance, there is a modest relationship

between the measures (r= -0.35 for TRACE and r=-0.34 for Merge).

The mismatch distributions beg the question, What types of classification errors do the

models make? To answer this, the proportion of mismatches in each of the 12 conditions was

computed for each model and are plotted in Figure 11. The profile of mismatches across

conditions reveals more similarities than differences between the models. For the most part the models performed similarly. The major difference is that Merge produced more phoneme misclassifications (conditions 3 and 6) and TRACE produced a greater proportion of lexical misclassifications. In the four phoneme conditions in which errors were made, the final phoneme was created by cross-splicing two different phonemes, creating input that specified one weakly and the other strongly. The errors are a result of misclassifying the phoneme as the more weakly specified (lowercase) alternative (e.g., *g* instead of *B*). By design, Merge's bottom-up only architecture heightens its sensitivity to sensory information, which in the present simulation made it a bit more likely than TRACE to misclassify cross-spliced phonemes.

Lexical misclassification errors in TRACE are due primarily to a bias to respond "nonword." However, a small percentage of these lexical errors are due to classifying the stimulus as *jog* (condition 5). This also occurred in condition 2, but much less often, with misclassifications as *jog* constituting 8% of the mismatches for TRACE and 3% for Merge.

Figure 11 here

Which Patterns Matter?

Although some misclassifications can be legitimized on many grounds (e.g., humans make such errors, perception of ambiguous stimuli will not be constant), it is important to determine whether they are characteristic behaviors of the model or idiosyncratic patterns, rather like the "particulars" defined in the ALCOVE analysis. That is, it is useful to distinguish between unrepresentative and representative behaviors. To do so, we measured the volumes of all regions identified by the PSP algorithm. As in the ALCOVE analysis, a threshold of 1% of the valid volume was adopted to define a meaningful pattern. When this is done, many patterns turn out to be noise and the set of representative patterns is reduced to a handful. For TRACE, 21

of its patterns (3 unique and 18 common) do not meet this criterion. Four patterns, all common, dominate in volume, together accounting for 99% of the volume (range 3.8 % - 45.2%). For Merge, the set of dominant patterns is larger, and is split equally between common and unique patterns. Thirty six patterns (21 unique and 15 common) fail to reach the 1% criterion. Seven common (range 1.3% - 21.1%) and eight unique (1.4% - 15.7%) patterns do so and make up 96% of the valid volume. Even with a threshold that eliminated 75% of all patterns, the asymmetry in pattern generation between the models is still present (TRACE=4; Merge=15).

The volumes of the representative common patterns are graphed in Figure 12. The numerals in the legend refer to the mismatch distance of each common pattern from the empirical pattern. Most obvious is the fact that the empirical pattern is much larger in TRACE than in Merge (33.1% and 6.8%) and that one mismatching pattern (filled black) dominates in both models (45.2% and 21.1%). This pattern turns out to be one in which there is a bias to classify all stimuli as nonwords. As a group, the eight unique Merge patterns mismatch the empirical pattern more than the common patterns. The largest pattern occupies a region of 15.7% (six mismatches), nearly twice the next largest region (8.9%). In this pattern, not only did Merge exhibit the same nonword response bias, but it also categorized cross-spliced phonemes as the competing (remnant) phoneme.

Figure 12 here

Response Time Analyses

Up to this point in the analysis we have compared only the classification behavior of the models, but the time course of processing is equally important, as experimental predictions often hinge on differences in RT between conditions. Connectionist models are generally evaluated on their ability to classify stimuli at a rate (e.g., number of cycles) that maintains the same ordinal

relations across conditions found with RTs. To assess the robustness of simulation performance, parameters can be varied slightly and additional simulations then run to ensure that the model does not violate this ordering (e.g., by producing a reversal of the RT pattern). In their comparison of Merge and TRACE, Norris et al (2000) found neither produced an RT reversal. Our test was a more exhaustive version of theirs.

The PSP analysis defines for us the region in the parameter space of each model that corresponds to the empirical pattern. When assessing RT performance, what we want to know is whether there are any points in this region (i.e., parameter sets) that yield invalid RT patterns, as defined by the ordinal relations among conditions in the experiment (i.e., the RT pattern in the six phonemic and three lexical conditions in Table 3). To perform this analysis, 10,000 points were sampled over the uniform distribution of the empirical region. Simulations were then run with each sample and the cycle time at which classification occurred was measured and compared across all conditions. Violations were defined as reversals in cycle times between adjacent conditions (e.g., condition 1 vs. condition 2) in phoneme classification and lexical decision. Just as Norris et al (2000) reported, we did not find a single reversal between conditions for either model.

<div align="center">Summary</div>

The PSP analyses reveal that the consequence of splitting the phoneme level in two is to produce a slightly more flexible model. This was found when all patterns were considered and when those most representative were considered, so the nature of this additional flexibility is of some interest. By splitting the phoneme level in two, Merge was not transformed into a different model. Rather, the model's behavior was expanded beyond that of TRACE, as the nested relationships in Table 4 show. Merge retained many of TRACE's behaviors (producing most of

its patterns) plus acquired new ones. A consequence of this expansion is that the representativeness of these behaviors is considerably different across models, as shown in Figure 12. It is because these differences are quantitative more so than qualitative that the models are so similar on other dimensions (e.g., frequency and types of mismatches, relative region size, response time relations between conditions).

Application 3: The Bottom-Up Priority Rule in Merge and TRACE

Background to the Analysis

The Bottom-Up Priority Rule

This second comparison of TRACE and Merge was performed to determine how the addition of a bottom-up priority (BUP) rule in Merge distinguishes it from TRACE. Recall that in Merge, for a phoneme decision node to become activated (Figure 9), excitation must be initiated from the phoneme input stage (i.e., evidence for the phoneme must be in the acoustic signal). In TRACE, this is not required. Initial activation via top-down connections from the word stage is possible. For example, having been presented with *jo* in *job*, excitation from the *job* node will feedback and excite the *b* node. Inhibitory connections between phoneme nodes makes it possible to observe indirect word-to-phoneme inhibition, because the activated *b* node will in turn inhibit competing phoneme nodes (e.g., *g)*. Thus, word nodes have the ability to excite directly and inhibit indirectly phoneme nodes.

Of course, one can easily incorporate a BUP rule into TRACE, simply by not allowing top-down feedback to influence a phoneme node until after that node has received some bottom-up input. Similarly, the BUP rule can be removed from Merge, by allowing phoneme decision units to be activated by word layer units without first having to be activated by phonemic input units.

This observation suggests an elegant way to assess the rule's contribution to model behavior: Run PSP analyses on both models with and without the BUP rule. This amounts to the 2x2 factorial design shown in Table 5. The lower left and upper right cells represent the models as originally formulated. This comparison serves as a reference from which to understand the contribution of the priority rule and the models' structures in affecting behavior. Comparisons of results between columns (i.e., models) neutralizes the effects of the priority rule. If the rule is primarily responsible for differences in model behavior, then the results for both models should be quite similar when the rule is and is not operational. If differences still remain, then structural differences are also contributing to their diverse behaviors. In short, these analyses will tell us whether Merge, without its priority rule, behaves like TRACE, and whether TRACE, with the priority rule, mimics MERGE.

Table 5 here

The Frauenfelder et al. Data

If indirect word-to-phoneme activation is possible, then one would expect anomalous word endings to slow down human performance. Frauenfelder et al (1990, Experiment 3) tested this prediction by having listener monitor for phonemes that occurred late in multisyllabic words and nonwords. Three conditions are of interest in the present analysis. A word condition (e.g., *habit*, with /t/ as the target phoneme) served as a lower bound on responding because lexical and phonemic information should combine to yield fast RTs. A control nonword condition (e.g., *mabil*, with /l/ as the target phoneme) served as a reference against which to measure inhibition. Because *mabil* is not a word, there should be no top-down lexical facilitation or inhibition when responding to /l/; responses should be based on sensory input alone. In the third, inhibitory

condition, listeners heard nonwords like *habil*, with *l* as the target phoneme. A slowdown in RT relative to the control nonword condition is expected if there is in fact inhibition. This is because the first part of the stimulus, *habi*, will excite *habit*, whose activation should then feed back down and excite /t/, which will then inhibit /l/.

The RT slowdown in the inhibition condition was small and not reliable, a null result which has been interpreted as arguing against word-to-phoneme excitation in TRACE. However, in simulations of indirect inhibition, TRACE's behavior is not cut and dry, with inhibition being more likely with longer than shorter words (Norris et al, 2000). In contrast, the bottom-up priority rule in Merge guarantees that it produces consistent performance that never yields inhibition. TRACE's variable behavior is a sign that it can generate more data patterns than Merge. If this is the case, is the absence of the priority rule the main cause, or is it also due to structural differences between the models?

<div align="center">The PSP Analysis</div>

Preliminaries

The design of the indirect inhibition experiment is much simpler than the subcategorical mismatch experiment. There are only three conditions and only a single response decision (phonemic). To simulate the experiment, the combination of so few conditions and a simple model design (one lexical and two critical phoneme nodes) would yield so few potential data patterns that the analysis might not provide satisfying answers to our question. We therefore added a lexical decision response (word or nonword) to the design on the grounds that listeners would, if asked, accurately categorize each stimulus as a word or nonword. The models should perform similarly. This additional response permitted a more fine-grained analysis and comparison of the models.

Input to the models consisted of three utterances: *habit*, *mabil*, *habil*. They were selected to be

moderately long (five phonemes) so that indirect inhibition would have a chance to emerge. Both

models were modified from the previous test to consist of only one lexical node (*habit*) and the

appropriate phoneme input/decision nodes, with /t/ and /l/ of most interest because their

activation functions were used to test for inhibition. With two classification response, each with

two alternatives, and three stimulus conditions, there were a total of 64 possible data patterns (2^3

x 2^3).

To examine the effect of the BUP rule on model performance, phoneme activation

parameters were adjusted accordingly in each model prior to running the PSP algorithm. The

same two decision rules were again used to assess the generality of results. With two decision

rules and two priority rules, the algorithm was run four times on each model. The consistency of

the results for each analysis was ascertained by rerunning the algorithm five times. The averaged

data are presented below. No more than seven minutes were required to find all patterns in any

run.

Classification Analyses

The classification data are shown in column 3 of Table 6, with the first row containing

the comparison of the models as originally designed (TRACE without BUP and Merge with it).

The Venn diagram shows that the relationship between the models is opposite of that in the

subcategorical mismatch experiment (Table 4), with Merge now embedded in TRACE, and

TRACE producing three times as many data patterns as Merge (12 vs. 4). Both models are again

highly constrained in their behavior, producing nowhere near the 64 possible data patterns in the

experimental design.

Table 6 here

TRACE's extra flexibility in this experimental setup agrees with Norris et al's (2000) observation of TRACE's variable behavior in producing indirect inhibition. That this flexibility is due to the absence of a BUP rule can be seen by examining the venn diagrams in the remaining rows. When the priority rules are swapped between models (row 2), their relationship reverses. TRACE shrinks from twelve to four patties while Merge grows from four to twelve, embedding TRACE in Merge.

This reversal is not merely in terms of the number of data patterns, but extends to the actual data patterns themselves. That is, even though the models trade places in terms of their relationship, the common and unique patterns remain the same. To see that this is the case, look at the venn diagrams in rows 3 and 4, where the priority rule is either on or off at the same time in the two models. In both situations TRACE and Merge overlap completely, producing only common patterns, four when the priority rule is on and 12 when it is off, which matches exactly what was found in rows 1 and 2. The preceding observations can be neatly summarized by noting that when the priority rule was on, the models always generated four patterns. When it was off, they generated 12.

The isomorphism of the models with and without the priority rule is surprising. Within this experimental design, their qualitative behaviors are identical in terms of the patterns they can generate, and point to the rule itself as a primary determiner of behavioral differences. This interchangeability shows up in the RT analyses as well.

Columns 4-7 in Table 6 contain volume measurements whose relationship between models is similar to that found in the subcategorical mismatch analyses. TRACE's valid region of the parameter space is much smaller than Merge's. This relationship changes little as a function of the priority rule. When common data patterns occupy only a portion of the valid

volume (rows 1 and 2), the region is larger for TRACE (75%) than Merge (56%), indicating that

the unique patterns occupy a smaller region in TRACE's parameter space. This result

foreshadows what happens under the stringent threshold.

A quick glance down the last column of Table 6 (stringent threshold) shows the

isomorphism between the models no longer holds. This can be seen most clearly in the top cell,

where the embedding is the reverse of that found with the weak threshold. TRACE produced

only three of the 12 patterns, two of which are shared by Merge. In contrast, Merge produced the

same four patterns as found under the weak threshold. In rows 2-4, TRACE remains embedded

in Merge, always producing the same few patterns (never the empirical one). It appears that

under the stringent threshold, TRACE generated so few patterns to begin with that there was

little opportunity for the priority rule to alter model behavior. In contrast, Merge behaved just as

it did under the weak threshold, generating more patterns without the rule (rows 2 and 4) than

with it (rows 1 and 3). Notice that what changes most down this column is the number of unique

patterns generated by Merge[3].

The reason for this asymmetric effect of the stringent threshold on the two models can be

understood by examining the volume measurements under the weak threshold (Figure 13).

Region sizes are shown for both models with and without the priority rule. Focus on the two No

BUP bars. Each slice of a vertical bar represents a different data pattern, except those denoted

with gray shading, which represents the combined area of eight regions, most of which are

visible for Merge but not for TRACE. In fact, these eight bars together total less than 1% of

TRACE's valid volume. These regions plus the human region are so small that they all disappear

when the stringent threshold is applied, which is the cause of the sharp drop in the number of

data patterns. These same regions are much larger in Merge (only one is less than 1% of the

volume), and although they shrink in size, they do not disappear when the stringent threshold is applied.

Figure 13 here

Further insight into the effects of the priority rule on model behavior can be obtained by comparing the types of misclassification errors made by the models when the rule was on and off. As in Figure 12, each slice of a vertical bar in Figure 13 is filled with a graphic pattern that denotes the mismatch distance of the pattern from the empirical data. In TRACE, the empirical pattern (bottom row) itself occupies a small region in the parameter space that changes imperceptibly in size when bottom-up information is given priority. In Merge, however, not only is this region much larger, but it increases in size when the priority rule is invoked, occupying 29% of the valid volume. Although TRACE can produce the empirical pattern, it is clearly a more central pattern in Merge. What is more, the priority rule enhances this centrality.

Most of the mismatching patterns, especially those that occupied the largest regions, rarely veered far from the empirical pattern, differing by one or two responses out of a possible six. Comparison of the bars between models shows that the same biases exhibited by the models in the subcategorical mismatch experiment are present here. A few patterns dominate the parameter space in TRACE whereas the space is split more equitably among patterns in Merge. This is most evident in the No BUP conditions, where a pattern with two mismatches occupies half (50.6%) of TRACE's volume. The errors in this instance are due to lexical misclassifications in which the two nonwords (*mabit* and *habil*) were categorized as words. Merge exhibits a much weaker tendency to do this (6.8%). Instead, the pattern occupying the largest volume in Merge does just the opposite: *habit* is classified as a nonword. TRACE produces this error as well. Application of the priority rule (bars 2 and 4) increases these

tendencies in both models (i.e., the regions increase in size).

Reaction Time Analyses

RT analyses were performed only on the weak threshold data because TRACE failed to generate the human pattern under the stringent threshold. The region occupied by the empirical pattern was probed for parameter sets that violated the ordering of mean participant response times across the three conditions (*habit* < *mabil* = *habil*). Model behavior with and without the priority rule was also examined.

The procedure was the same as that used in the subcategorical mismatch analysis. Samples (10,000) were drawn from the uniform distribution over the region of the empirical pattern and the model was run with each set of parameter values. The cycle at which phoneme classification occurred was recorded across conditions. Because violations amount to a reversal in cycle classification time between adjacent conditions, we calculated the difference in cycle time of neighboring conditions, *habit* vs. *mabil*, and *mabil* vs. *habil*. Cycle times to *mabil* were subtracted from those in the other two conditions. Distributions of these difference scores (over all samples) are plotted in Figure 14. The top graphs contain the comparison of the models as originally formulated (TRACE without BUP, Merge with BUP). In the bottom pair of graphs, the priority rules were swapped across models. The left-hand graphs contain the *habit-mabil* comparison and the right-hand graphs the more theoretically important, *mabil* vs *habil* comparison.

Figure 14 here

In the upper left graph, both models produce the correct ordering, with phoneme classification always being faster in a word (*habit*) than nonword (*mabil*) context. What differs between the models are the shapes of the distributions, with TRACE's being sharply peaked and

Merge's more diffuse. Inn the right graph, the effects of the priority rule are visible, with Merge always generating the correct prediction of identical classification times in the two conditions (zero RT difference, a single point) and TRACE showing indirect inhibition, with recognition times slightly longer to /l/ in *habil* than in *mabil*, hence the negative difference score. Although this effect is almost always small, it is clear from the shape of the distribution that choice of parameter settings could lead to different conclusions, where on rare occasion the effect would be large.

That equivalent classification times in the *mabil* and *habil* conditions are due to the priority rule can be seen in the lower right graph, where TRACE was run with the BUP rule and Merge without it. Just as in the classification data in Table 6, their performance reverses: TRACE now produces no difference across conditions whereas Merge displays indirect inhibition.

Finally, scanning across all four graphs, it is clear that the interactive architecture of TRACE constrains classification time more so than the integrative architecture of Merge. Regardless of priority rule, both models show a tendency to produce small rather than large differences scores, but TRACE more so than Merge given the height and location of the peaks of the distributions.

<div align="center">Summary</div>

This second comparison of TRACE and Merge was undertaken to learn how the bottom-up priority rule can differentiate the two models. The PSP analyses under the weak and stringent threshold showed that the rule reduces model flexibility and confirmed its necessity for simulating the correct pattern of response times. By turning the rule on and off across models, we discovered that they can behave identically, producing qualitatively indistinguishable data

patterns. Only when the volumes of these patterns were inspected did performance differences due to model architecture emerge. Merge's integrative architecture is somewhat better suited for mimicking human behavior in this experimental design. Not only is the empirical pattern more central in Merge, but this was true regardless of whether the priority rule was in place. These findings and conclusions are possible because of the global perspective on model behavior that a PSP analysis provides.

General Discussion

Models are becoming increasingly useful tools for psychologists. To use them most productively, it is important to have a thorough understanding of their behavior. Although a host of methods have been developed with this purpose in mind (Figure 1), their applicability is limited, in large part because of the diversity of models in psychology. To the extent that new methods are introduced, it would be most useful to understand model behavior at a level of granularity that matches the qualitative predictions that are made when testing models experimentally.

PSP was developed to meet these goals. A model's ability to mimic human behavior is evaluated in the context of the other data patterns that it can generate in an experimental design, thereby providing a richer understanding what it means for a model to account for an experimental result. We specifically chose connectionist models to demonstrate PSP's versatility, as there are few methods available for analyzing these widely-used and complex models. Other types of models can be compared. Indeed, one reason PSP is so versatile is that the model is really just a module in the algorithm, making it possible to insert models of almost any type, be they algebraic, algorithmic, connectionist, etc. Comparison of models across types is relatively straight forward.

The three example applications demonstrate some of what can be learned with PSP. In the case of ALCOVE, PSP was used to perform a basic task in model analysis - evaluate the model's soundness in capturing an empirical result. Not only did the analysis inform us about whether the model can mimic human data (its local behavior), but much more importantly, by studying the partitioned parameter space, we learned how representative this behavior is of the model and how many other data patterns the model can produce. Some of these were plausible alternative response patterns. Others occupied such a small region in the parameter space that they should be considered spurious and uncharacteristic of the model.

In the second and third applications, PSP was used to study the behavioral consequences of slight design differences between two localist connectionist models. To do this, data patterns and their corresponding volumes in parameter space were compared across models in two experimental settings that were intended to bring out design differences. Many more similarities than differences were found in classification behavior. RT analyses of the region occupied by the empirical data led to the same conclusion. Qualitatively, TRACE and Merge were indistinguishable. Only in the volume analyses did differences emerge. These took the form of biases in emphasizing some data patterns over others.

Although one might not be surprised by the models' similarities, the PSP analysis provides the evidence to solidify such a claim. Furthermore, the analyses suggest how difficult it could be to find a situation in which they make different qualitative predictions. That said, PSP might be just the tool to assist in such an endeavor, by enabling the researcher to essentially pretest an experiment to learn whether it could discriminate between the models. Once an experiment is designed, the models can be run through the PSP algorithm to determine which

data patterns are shared and unique. To the extent that participants are likely to produce those in the latter category, the experiment has the potential to discriminate between the models.

Of course, the meaningfulness of any such analysis depends on how a data pattern is defined. Conclusions may be specific to a definition, which is why it can be worthwhile to use more than one. Because ordinal predictions dominate in most experimental work, the task of defining a data pattern can be relatively straight forward. This is another reason why PSP is widely applicable. It does not depend on quantitative fits.

That said, for other psychological models, there may be no obvious "natural" definition of a data pattern. A case in point is the power model of retention (forgetting). The response probability y, representing proportion recall, is a continuous function of retention interval t in the form of $y = at^{-b}$ where a and b (a, b > 0) are two parameters. Application of PSP requires one to decide whether the retention curve associated with parameter values $(a, b) = (1, 0.5)$ should be considered indistinguishable, in some defined sense, from the curve associated with parameter values $(a, b) = (1.1, 0.4)$. In the absence of a justifiable definition of a data pattern in a situation like this, it would be useful if one could devise a "generic", suitably interpretable, definition. Work is underway on this problem.

PSP can also be of considerable use in model development by making the modeler aware of the broader consequences of the choices made in model design. For example, there may be four key experimental results that any model of memory must capture to be taken seriously. By running the algorithm in each experimental setting, a comprehensive analysis of the model's design can be undertaken. Desirable and undesirable behaviors can be readily identified. The model can then be refined on the basis of this knowledge and the entire process repeated. From successive iterations, one can develop a clear understanding of the model's inner workings (e.g.,

how parameters contribute to model behavior, whether there are redundancies across parameters, the appropriateness of their ranges, and even optimal parameter values).

Figure 15 provides a glimpse of what can be learned from such detailed analyses. Shown are distributions of parameter values for three of TRACE's parameters that were obtained within the empirical region of the parameter space in the indirect inhibition experiment (no BUP). Each distribution contains 10,000 values and gives a sense of how the parameter contributes in capturing the experimental result. For the phoneme excitation parameter, the distribution of possible values is fairly broad, although there is a clear optimum. The phoneme-to-word excitation parameter is much more tightly constrained, so much so that one might wonder whether its range should be cut in half, from 0 to 0.5. The distribution for the phoneme-to-phoneme inhibition parameter is essentially flat, suggesting a high degree of context sensitivity, which goes along with Norris et al's (2000) observations about the stability of indirect inhibition. Analyses like these expose the modeler to tendencies and behaviors that might not be anticipated.

Figure 15 here

Throughout the paper, the global perspective on model behavior that PSP provides has been examined from the standpoint of the model to determine its flexibility or scope (Cutting, 2000). It can be equally productive to view the situation from the vantage point of the experiment, by asking how much the experiment is likely to challenge the model. To the extent that it will, the experiment has the potential to be a good, meaningful test of the model. In this regard, one lesson that comes out of our applications of PSP is that the fewer the number of conditions and the fewer the possible relationships between them, the less informative the test may be.

Because PSP is a search problem, it is subject to many of the same difficulties found in nonlinear function optimization and integration, which work under a set of regularity conditions about the target function (i.e., model), such as continuity, smoothness, stationarity, existence of a solution within a finite limit and so on. For this reason, the algorithm is currently limited in scope. Another requirement is that the range of parameters must be finite. If some unconstrained parameters are unavoidable (i.e., plausible data patterns could still be generated from their extreme values), one could reparameterize the model utilizing log, inverse logistic, or other transformations.

As currently implemented, volume estimation of a region is performed using a multi-dimensional ellipsoid. This is satisfactory for models that generate regions that are not oddly shaped (sharply bowed) or discontinuous. Exploratory work ensured that these conditions were satisfied for the current simulations. However, the algorithm would be more powerful, precise, and useful were it more general and applicable to such situations. Work is underway to extend its capabilities.

In conclusion, analysis of the global behavior of a model can provide a wealth of information that is useful for understanding a model's performance in a particular testing situation. PSP is a flexible and powerful method of globally model analysis, as the examples presented here attest, suggesting it has considerable potential to assist researchers in their quest to model human behavior.

References

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

Browne, M. W. (2000). Cross-validation methods. Journal of Mathematical Psychology, 44, 108-132.

Cutting, J. (2000). Accuracy, scope, and flexibility of models. Journal of Mathematical Psychology, 44, 3-19.

Dunn, J. C. & James, R. N. (2003). Signed difference analysis; Theory and application. Journal of Mathematical Psychology, 47, 389-416.

Dawson, R.W., & Shamanski, K.S., (1994). Connectionism, confusion, and cognitive science. The Journal of Intelligent Systems, 4, 215-262.

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. Nature, 407, 630-633.

Frauenfelder, U.H., Segui, J., & Dijkstra, T. (1990). Lexical effects in phonemic processing: Facilitatory or Inhibitory? Journal of Experimental Psychology: Human Perception and Performance, 16, 77-91.

Grunwald. P. (1998). The Minimum Description Length Principle and Reasoning Under Uncertainty. PhD Thesis. ILLG Dissertation Series D5 1998-03. CWI. The Netherlands.

Grunwald, P., Myung, I.J., & Pitt, M.A. (in press). Advances in Minimum Description Length: Theory and Application. Cambridge, MA: MIT Press.

Johansen, M. A. & Palmeri, T. J. (2002). Are there representational shifts during category learning? Cognitive Psychology, 45, 482-553.

Kass, R.E., & Raftery, A.E. (1995). Bayes factors. Journal of the American Statistical

Association, 90, 773-795.

Kim, W., Navarro, D.J., Pitt, M. A., Myung, I.J. (2004). An MCMC-based method of comparing connectionist models in cognitive science. Advances in Neural Information Processing Systems, 16, (937-944). MIT Press

Kontkanen, P., Myllymäki, P., Buntine, W., Rissanen, J. & Tirri, H. (in press). An MDL framework for data clustering. To appear in P. Grünwald, I. J. Myung & M. A. Pitt (Eds.) Advances in Minimum Description Length: Theory and Applications. Cambridge, MA: MIT Press.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. Psychological Review, 99, 22-44.

Kruschke, J. K. (1993). Three principles for models of category learning. The Psychology of Learning and Motivation 29, 57–90.

Kruschke, J. K. & Erikson, M. A. (1995). Six principles for models of category learning. Talk presented at the 36th Annual Meeting of the Psychonomic Society, 10 November 1995, Los Angeles, CA.

Lee, M. D., & Navarro, D. J. (2002). Extending the ALCOVE model of category learning to featural stimulus domains. Psychonomic Bulletin and Review, 9, 43-58.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), Handbook of Mathematical Psychology (Vol. 1, p. 103-190). New York, NY: Wiley.

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

McCloskey, M. (1991). Networks and Theories: The Place of Connectionism in Cognitive Science. Psychological Science,2, 387-395.

McQueen, J., Norris, D., Cutler, A. (1999). Lexical influence in phonetic decision making: Evidence from subcategorical mismatches. <u>Journal of Experimental Psychology: Human Perception & Performance, 25</u>, 1363-1389.

Marslen-Wilson, W., & Warren, P. (1994). Levels of perceptual representation and process in lexical access: Words, phonemes, and features. <u>Psychological Review</u>, <u>101</u>, 653-675.

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

Navarro, D.J., Pitt, M.A., & Myung, I.J. (2004). Assessing the distinguishability of models and the informativeness of data. <u>Cognitive Psychology</u>, <u>49</u>, 47-84.

Norris, D., McQueen, J.M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. <u>Behavioral & Brain Sciences 23</u>, 299-370.

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C. & Glauthier, P. (1994). Comparing models of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins. <u>Memory & Cognition 22</u>, 352–369.

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

Platt, J.R. (1964). Strong Inference, <u>Science, 146</u>, 347-353.

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

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

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

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

Shepard, R. N., Hovland, C. L. & Jenkins, H. M. (1961). Learning and memorization of classification. Psychological Monographs 75 (13), Whole No. 517.

Shiffrin, R., & Nobel, P. (1998). The art of model development and testing. Behavior Research Methods, Instruments & Computers, 29, 6-14.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). Journal of Royal Statistical Society, Series B, 36, 111-147.

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. Journal of Mathematical Psychology, 48, 28-50.

Appendix A: Additional Details about How the PSP Algorithm Works

Application of the PSP algorithm is based on the following five premises: (1) regions are contiguous with one another in the sense that a path between any two regions exists *within* the partition; (2) the size of the regions changes smoothly in the parameter space, so small regions tend to cluster together; (3) model behavior is stationary in the sense that a given parameter set always generates a single, fixed data pattern. This means that the boundaries of the region are fixed, not varying every time the model generates a data pattern; (4) the range of the parameter space to be searched is finite. (5) the data space can be discretized in such a way that the total number of data patterns to be discovered is finite to the extent that they can be found within a reasonable amount of computing time. With these premises, the PSP algorithm can be summarized as the following three steps:

Given θ_1 and Pattern 1, set $m = I = 1$.

Step 1.   Establish $q_m(\cdot|\cdot)$ by adapting the size of its hyper-spherical domain. Go to Step 2.

Step 2.   Set $I = \mathrm{mod}(I, m) + 1$. Go to Step 3.

Step 3.   Sample θ_y from $q_i(\cdot|\theta_i)$. If θ_y generates a new valid pattern, set $m = m + 1$, $\theta_m = \theta_y$, and record the new pattern as Pattern m, and then go to Step 1. If θ_y generates Pattern I, set $\theta_i = \theta_y$ and go to Step 2. Otherwise, go to step 2.

In the above, $q_i(\cdot \mid \theta_i)$ denotes the jumping distribution of the region corresponding to pattern I centered at θ_i, and the subscript I $(1 \leq i \leq m)$ indexes the region from which we are currently sampling, and *m* represents the number of regions (or number of data patterns) found so far. Below are three additional observations on the algorithm.

Firstly, the size of the jumping distribution (i.e., the radius of the hyper-sphere) in each region must adapt to its size and shape. If it is too small, almost all candidate points will be accepted, but every jump will be so small that it will take too many jumps for an exhaustive search of a region. Also, rejected points will rarely be generated. In contrast, if the size of the jumping distribution is too large, candidate points will be rejected too often, and the granularity of the jumps will not be small enough to define the edges of a region, which requires a properly sized jumping distribution to succeed. Unless one is dealing with a normal distribution, no theory exists that defines the optimal jumping distribution. With the PSP algorithm, we have found it best to use an adaptive jumping distribution, which on average accepts 20% of sample points.

Secondly, the search process terminates if the following two conditions are met. First, a certain preset size of MCMC samples is obtained for each of the discovered regions. To compensate for the fact that regions discovered early in the search process are likely to be sampled more than regions discovered later, the algorithm concentrates more on the newly discovered regions in such a way that the total number of trials in the search history will eventually be the same for all

discovered regions. Second, if a new pattern is not discovered after a set number of trials (or time), the algorithm terminates.

Finally, it is worth noting that the PSP algorithm has the desirable property of focusing on each region in equal proportion, irrespective of its size. An equal number of search trials is performed in each region. As a consequence, closer attention is paid to the small regions. This means that the resulting sampling distribution over the whole parameter space is essentially a mixture distribution that gives higher density to points known to lie near many regions.

Appendix B

The weak order that defines a data pattern in the ALCOVE example can be decomposed into a set of equivalence relations, and a strong order on the equivalence classes. The strong order part is easy. We simply rank them in terms of mean learning rate. The equivalence relations are more difficult to derive, requiring that we partition the six curves using a suitable clustering procedure.

The clustering procedure we employed was a minor variant on the clustering technique introduced by Kontkanen, Myllymäki, Buntine, Rissanen, and Tirri (in press). The essence of the technique is to view a clustering solution as a probabilistic model for the data. In the current application, the likelihood function for the data takes the form of a mixture of binomials, with a single multivariate binomial for each cluster. The clustering procedure now reduces to a statistical inference problem, which is solved by choosing the set of clusters that optimizes a Minimum Description Length statistic. The six learning curves reported by Nosofsky et al. (1994) are averaged across 40 subjects over the first 16 blocks, consisting of 16 stimulus presentations each. Each data point is thus pooled across 40 x 16 = 640 trials. Using this technique it is possible to infer that I<II<(III, IV, V)<VI is indeed the natural structure for these data.

Lastly, we need to be able to associate a set of *predicted* learning curves with a data pattern, which is not the same thing as associating a set of *observed* learning curves with a data pattern. Nevertheless, it is not difficult to do. A set of response probabilities is first discretized to the same resolution as the empirical data. This is straightforward, by finding the expected values for the data, given by np, where n is the sample size and p is the average response probability

predicted for some category type across any given block of trials, and then rounding to the

nearest integer. While the rounding error is a nuisance, it is negligible for $n=640$. The discretized

curves are then mapped onto a qualitative data pattern by using the same clustering technique

used to classify the empirical data.

Appendix C: TRACE and Merge Parameters

Parameter	TRACE	Merge	Range
Phoneme excitation	*V*	*V*	(0,1)
Phoneme to word excitation	*V*	*V*	(0,1)
Phoneme to word inhibition	-	*V*	(0,1)
Phoneme to target excitation	-	*V*	(0,1)
Phoneme decay	*V*	*V*	(0,1)
Word to target excitation	-	*V*	(0,1)
Word to phoneme excitation	*V*	-	(0,1)
Word to word inhibition	*V*	*V*	(0,1)
Word decay	*V*	*V*	(0,1)
Target to target inhibition	-	*V*	(0,1)
Target decay	-	*V*	(0,1)
Target momentum	-	*V*	(0,1)
Phoneme to phoneme inhibition	*V*	-	(0,1)
Cycles per input slice	*V*	*V*	Fixed

Note: The word-to-word inhibition parameter was used in the indirect inhibition example.

Author Notes

Mark A. Pitt, Woojae Kim, Jay I. Myung, Department of Psychology, Ohio State University. Daniel J. Navarro, Department of Psychology, University of Adelaide.

Correspondence concerning this article should be addressed to Mark Pitt, Department of Psychology, Ohio State University, 1885 Neil Avenue, Columbus, OH, 43210. Electronic mail should be sent to Pitt.2@osu.edu

Footnotes

[1] Indeed, equating the number of patterns with model complexity is a natural and well-justified definition of complexity (Myung, Balasubramanian, & Pitt, 2000).

[2] Subsequent analyses in this section are restricted to the weak-threshold data, in part because of the paucity of common patterns. General conclusions do not change.

[3] The four corresponding columns of percentages for the stringent threshold were omitted from Table 6 because the values changed in the same ways as those in the analysis under the weak threshold.

Table 1

Search efficiency of the PSP algorithm and SMC as a function of the number of parameters and data patterns. Shown in each cell is the mean proportion of patterns found based on ten independent runs.

| Number of parameters | Number of patterns to find | | | | | |
| | 20 | | 100 | | 500 | |
	PSP	SMC	PSP	SMC	PSP	SMC
5	1.00	0.85	1.00	0.78	1.00	0.87
10	1.00	0.32	1.00	0.23	0.99	0.24
15	1.00	0.09	1.00	0.02	0.99	0.03

Table 2

 The twelve major data patterns predicted by ALCOVE, shown in rank order (smallest to largest) of region size. A major pattern is defined as one that occupies more than 1% of the parameter space on average. All twelve are universal patterns.

	Ranking					
Pattern Number	**1**	**2**	**3**	**4**	**5**	**6**
1	$1 = 2 = 3 = 4 = 5 = 6$					
2	$1 = 2 < 3 = 4 = 5 = 6$					
3	$2 < 1 < 3 = 4 = 5 = 6$					
4	$2 < 1 = 3 = 4 = 5 = 6$					
5	$1 = 2 < 3 = 4 = 5 < 6$					
6	$1 = 2 = 3 = 4 = 5 < 6$					
7	$1 < 2 < 3 = 4 < 5 < 6$					
8	$1 < 2 < 3 = 4 = 5 < 6$					
9	$1 < 2 = 3 = 4 = 5 < 6$					
10	$2 < 1 < 3 = 4 = 5 < 6$					
11	$1 < 2 = 3 = 4 < 5 < 6$					
12	$1 < 2 < 3 = 4 = 5 = 6$					

Table 3

Design of the subcategorical mismatch simulation in Norris et al (2000). Underlined segments denote the sections of the utterances that were excised and then combined. The condition names describe whether the cues to the final consonant in the first word matched those of the second word. Reaction time data are from McQueen et al (1999). Response sets in each task in the simulations are shown on the right. Asterisks denote listeners' dominant response.

	Condition	Example	Phonemic categorization	Lexical Decision	Phonemic categorization				Lexical Decision		
					/b/	/g/	/z/	/v/	"job"	"jog"	nonword
	Word				/b/	/g/	/z/	/v/	"job"	"jog"	nonword
1	matching cues	jo<u>b</u> + jo<u>b</u>	668	340	*				*		
2	mismatching cues from word	jo<u>g</u> + jo<u>b</u>	804	478	*				*		
3	mismatching cues from nonword	jo<u>v</u> + jo<u>b</u>	802	470	*				*		
	Nonword				/b/	/g/	/z/	/v/	"job"	"jog"	nonword
4	matching cues	jo<u>z</u> + jo<u>z</u>	706	na			*				*
5	mismatching cues from word	jo<u>g</u> + jo<u>z</u>	821	na			*				*
6	mismatching cues from nonword	jo<u>v</u> + jo<u>z</u>	794	na			*				*

Table 4

Classification results from the subcategorical mismatch test.
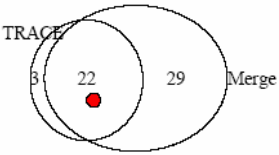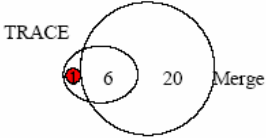
Threshold	Pattern overlap	Percentage of total volume occupied by valid patterns		Percentage of valid volume occupied by common patterns	
		TRACE	Merge	TRACE	Merge
Weak		10%	69%	99%	55%
Stringent		3%	18%	84%	21%

Table 5

Factorial combination of the PSP comparisons that were performed to determine whether the difference bottom-up priority in the two models is what distinguishes their behaviors in the test of indirect inhibition.

	Model	
Bottom-up Priority Rule	TRACE	Merge
Yes		
No		

Table 6

Results from the PSP analyses of Merge and TRACE in the indirect inhibition test.
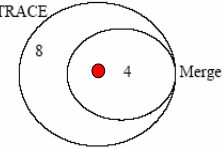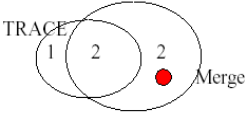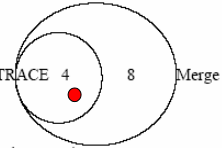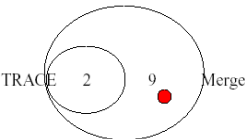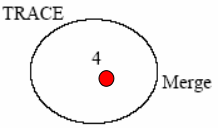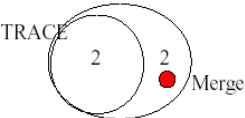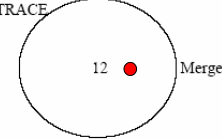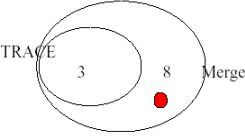
Bottom-up Priority Rule		Pattern overlap Weak Threshold	Percentage of total volume occupied by valid patterns		Percentage of valid volume occupied by common patterns		Pattern overlap Stringent Threshold
TRACE	Merge		TRACE	Merge	TRACE	Merge	
no	yes		20%	67%	75%	100%	
yes	no		15%	60%	100%	56%	
yes	yes		15%	67%	100%	100%	
no	no		20%	60%	100%	100%	

Figure Captions

Figure 1. The locations of methods of model analysis and comparison in a two-dimensional space, defined by the degree to which the method evaluates quantitative vs. qualitative model performance (horizontal axis) and whether the method focuses on local or global model behavior (vertical axis).

Figure 2. The partitioned parameter space of two hypothetical two-parameter models in an experiment with three conditions (A, B, C). M1 generates only three of the possible 12 patterns. M2 generates nine, indicating it is more complex. See text for further details.

Figure 3. Illustrations of two initial steps in the PSP algorithm. Panel a shows the selection of a point (black circle) in the parameter space and its accompanying jumping distribution (dashed circle). The shaded circle represents a sample from within the distribution. Panel b shows a region in parameter space that the algorithm has mapped out as generating the same data pattern (enclosed area filled with black dots). The point outside the region (shaded circle), along with a jumping distribution, denotes the start of a new search process to map out another region in the parameter space.

Figure 4. Illustration of the two-parameter model used to test the PSP algorithm. The algorithm had to find the 20 areas outlined in bold (valid patterns). Note that there are (invalid) regions outside of this area. They should be viewed as distracting portions of the parameter space and have the effect of increasing the difficulty of the search problem.

Figure 5. Search efficiency of the PSP algorithm compared to Simple Monte Carlo (random search). 100 patterns had to be found in a ten parameter model.

Figure 6. The six fundamental category types for the Shepard et al. (1961) task (panel a), and the learning curves for those types found by Nosofsky et al. (1994) in their replication (panel b). For visual clarity, the curves in panel b are labeled using arabic numerals rather than the more conventional roman numerals.

Figure 7. Means and standard errors for the rank of each category type in the data patterns recovered by PSP. The rankings for universals do not differ strongly from the rankings of the particulars.

Figure 8. Scatterplot of the log average-volume for each data pattern against the frequency with which the pattern was recovered. Note that since volume is shown on a logarithmic scale, a few universals ("major patterns") occupy the vast majority of the parameter space. The empirical pattern is indicated by the hollow circle in the upper right corner, the dashed horizontal line designates the 1% volume threshold, and the dashed vertical line designates the 95% recovery threshold.

Figure 9. Schematic diagrams of TRACE and Merge.

Figure 10. Similarity of all Merge and TRACE data patterns to the empirical (human) pattern measured in terms of the number of mismatches to the empirical pattern. Data from the

application of the weak and stringent thresholds are shown separately. The dashed line represents the distribution of mismatches for all patterns in the experimental design.

Figure 11. Proportion of mismatches by Merge and TRACE across the 12 conditions. Condition number corresponds to that in Table 3. The two-letter sequences are shorthand descriptions of the cross-spliced stimuli used in the conditions. The lowercase letter denotes the final phoneme of the first word and the uppercase letter denotes the final phoneme of the second word.

Figure 12. Comparison of Merge and TRACE volumes in the subcategorical mismatch experiment. Regions in the valid parameter space that correspond to data patterns that occupy more than 1% of the volume. Each slice represents a different data pattern. Numerals in the legend specify the mismatch distance of the pattern from the empirical (human) data.

Figure 13. Comparison of Merge and TRACE volumes in the indirect inhibition experiment. Regions that correspond to data patterns of each model without (No BUP) and with (BUP) the bottom-up priority rule. Each bar represents a different data pattern, whose shading specifies its mismatch distance from the empirical pattern (see numerals in legend).

Figure 14. A comparison of model classification response times in the indirect inhibition experiment. Differences in classification times (cycles) between adjacent experimental conditions are plotted for each model, with the *mabil* vs *habit* comparison on the left, and the *mabil* vs *habil* comparison on the right. In the top graphs, the models were run as originally formulated (TRACE without BUP, Merge with BUP). In the bottom graphs, TRACE was run

with the priority rule and Merge without it.

Figure 15. Distributions of values of three TRACE parameters from 10,000 samples of the region in parameter space corresponding to the empirical data pattern (without the priority rule). Parameter names are listed inside each graph.
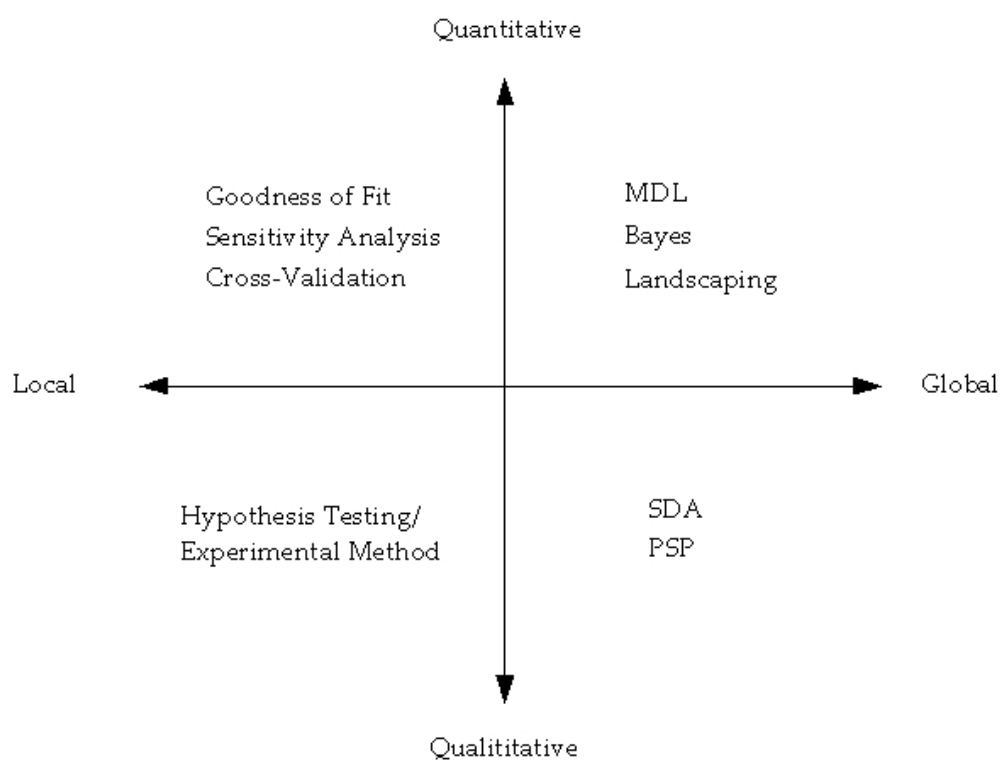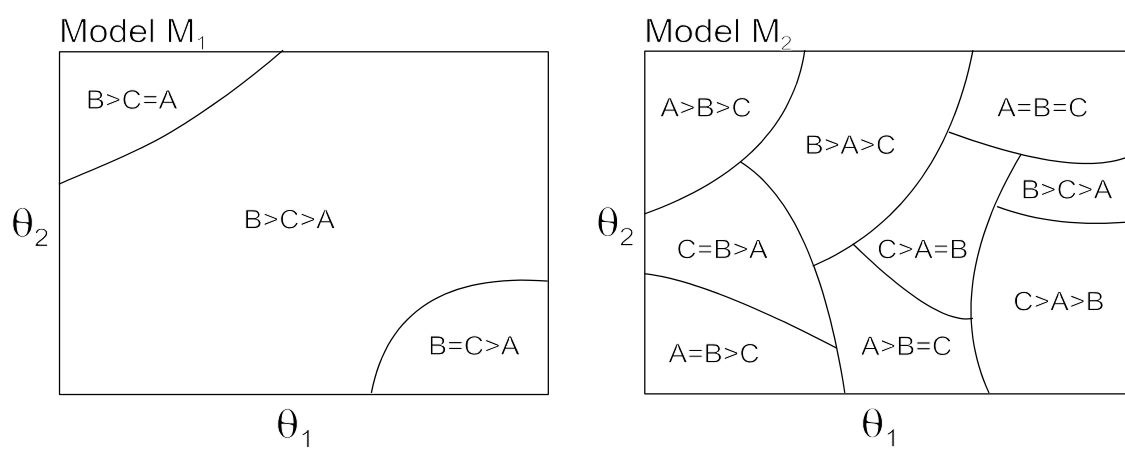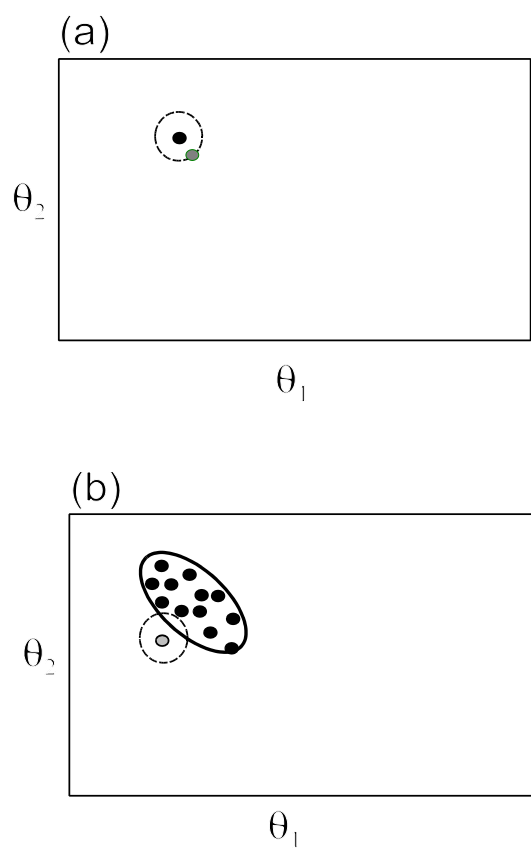
Figure 1

Quantitative

Goodness of Fit
Sensitivity Analysis
Cross-Validation

MDL
Bayes
Landscaping

Local

Global

Hypothesis Testing/
Experimental Method

SDA
PSP

Qualititative

Figure 2

Model M₁ — Regions labeled: B>C=A, B>C>A, B=C>A, with axes θ₂ (vertical) and θ₁ (horizontal)

Model M₂ — Regions labeled: A>B>C, B>A>C, A=B=C, B>C>A, C=B>A, C>A=B, C>A>B, A=B>C, A>B=C, with axes θ₂ (vertical) and θ₁ (horizontal)

Figure 3

(a)

θ_2

θ_1

(b)

θ_2

θ_1

Figure 4

Figure 5

Figure 6

(a)

(b)

Type I   Type II   Type III

Type IV   Type V   Type VI

dim 3
dim 2
dim 1
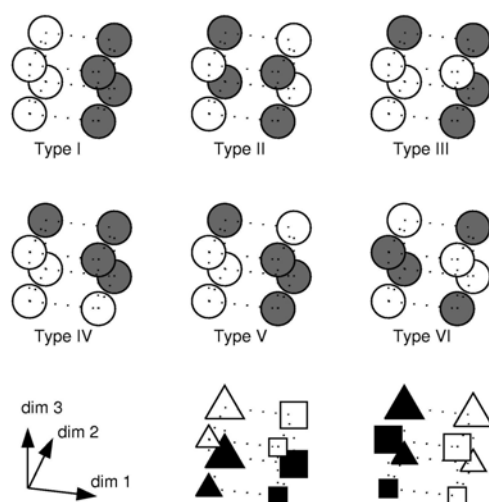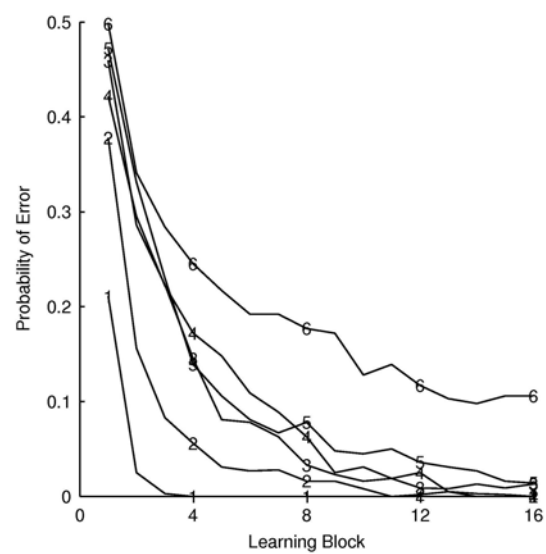
Figure 7
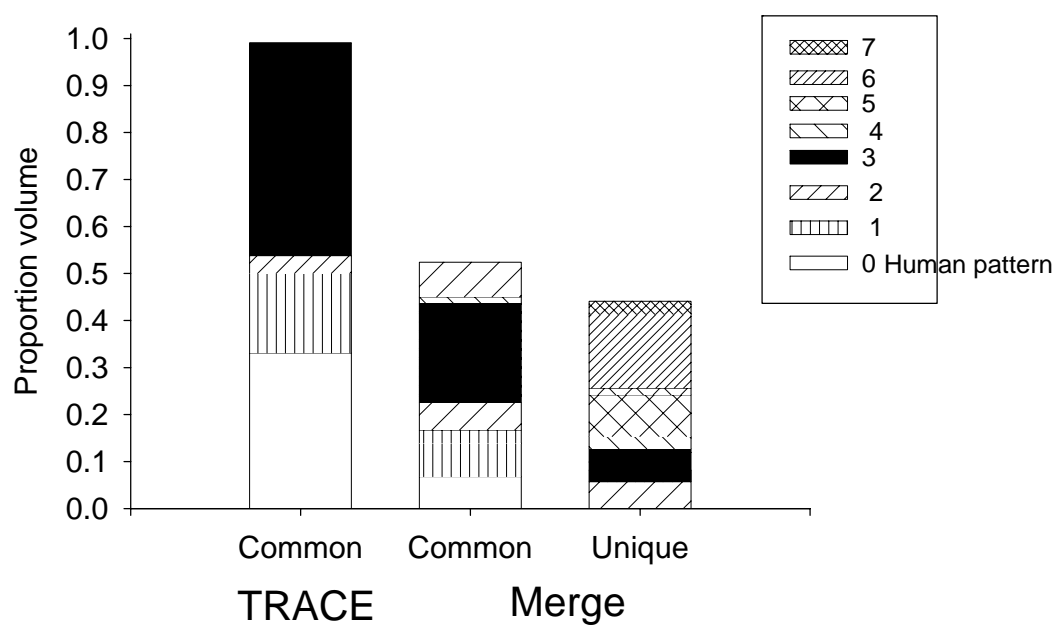
Figure 8

Figure 9

TRACE

Merge
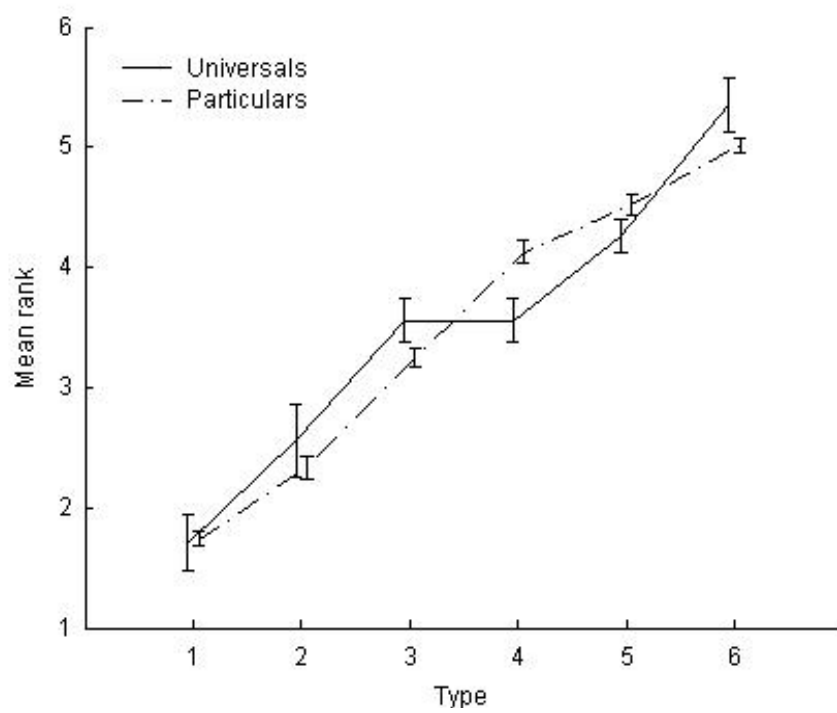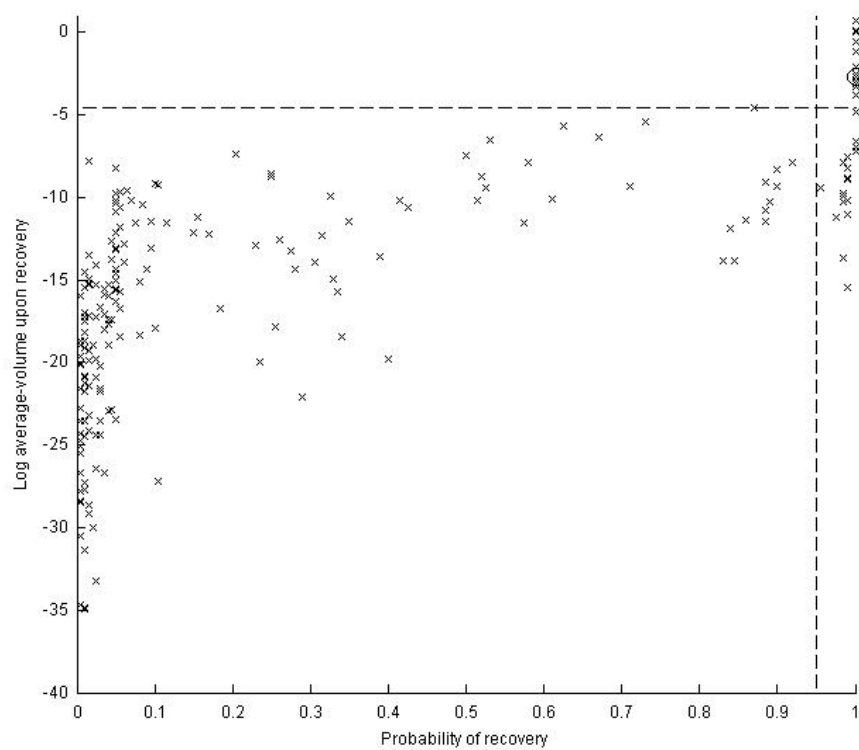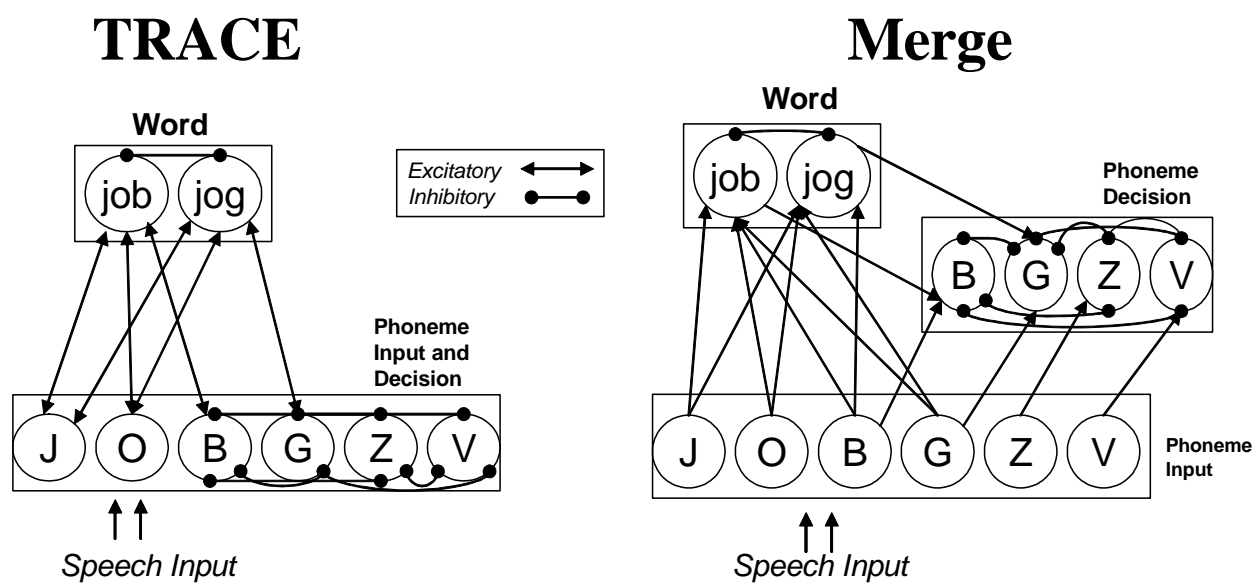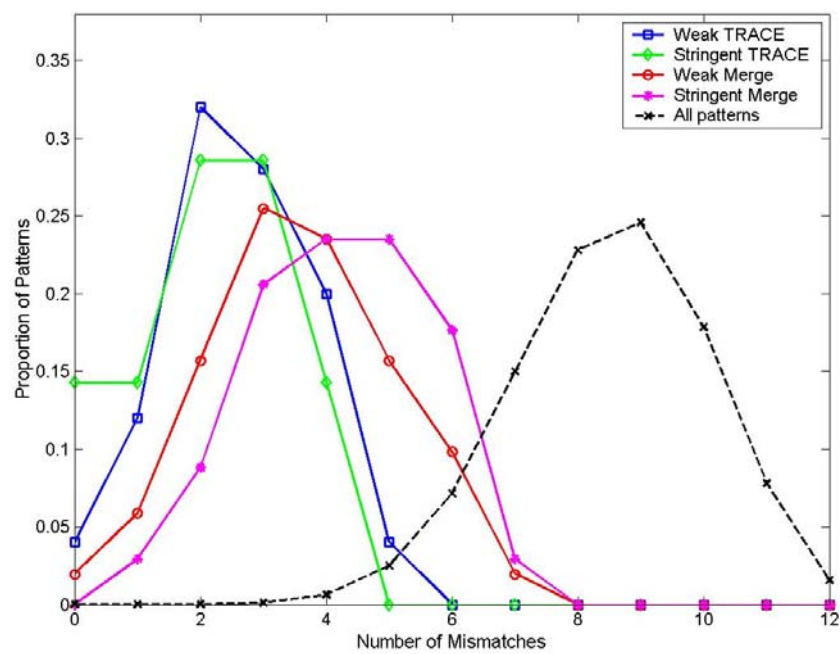
Figure 10

Figure 11
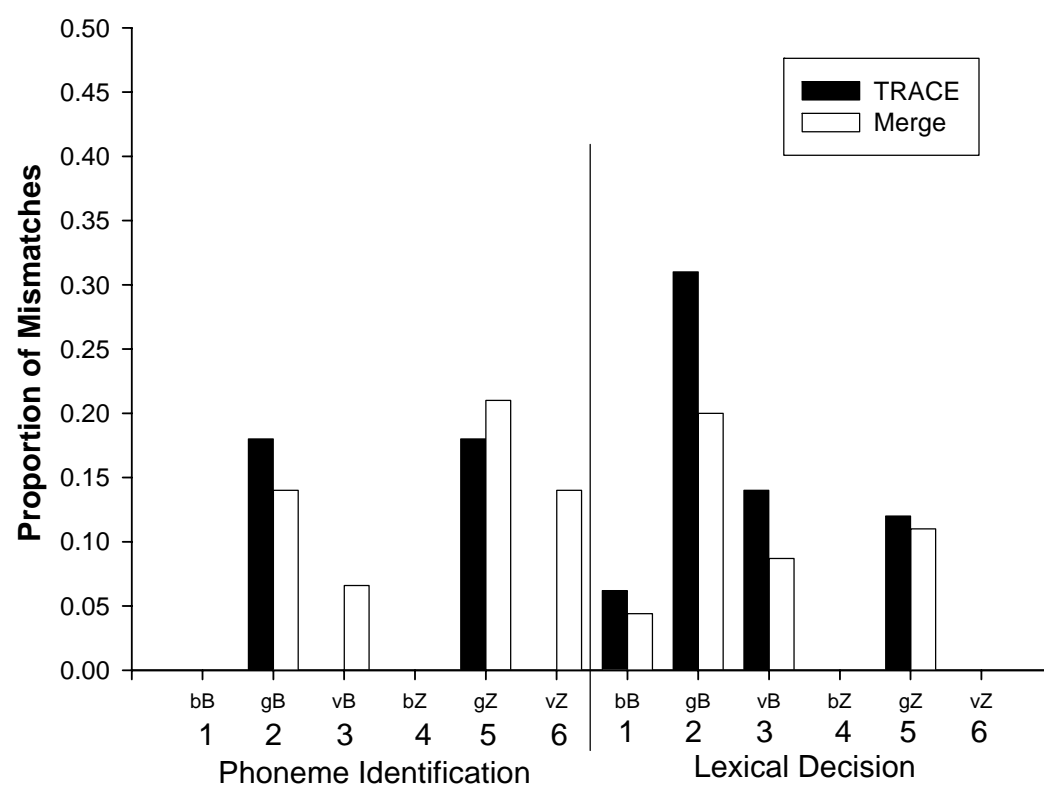
Figure 12

Figure 13

Figure 14

Figure 15