The Design and Analysis of State-Trace Experiments

Melissa Prince
Scott Brown
&
Andrew Heathcote

The School of Psychology,
The University of Newcastle, Australia

Address for Correspondence

Prof. Andrew Heathcote
The School of Psychology,
Psychology Building, The University of Newcastle,
University Avenue, Callaghan, 2308, Australia
Email: andew.heathcote@newcastle.edu.au
Phone: 61-2-49216778
Fax: 61-2-49216906

Suggested running head: State-Trace Analysis

Abstract

State-trace analysis (Bamber, 1979) addresses a question of interest in many areas of psychological research: does one or more than one latent (i.e., not directly observed) variable mediate an interaction between two experimental manipulations? Despite its increasing use, there is little guidance available on how to design an experiment suited to state-trace analysis, and existing statistical methods for state-trace analysis are problematic. We provide a framework for designing and refining a state-trace experiment, and statistical procedures for the analysis of accuracy data using Klugkist, Kato and Hoijtink's (2005) method of estimating Bayes Factors. The statistical procedures provide estimates of the evidence favoring one vs. more than one latent variable, as well as evidence that can be used to refine experimental methodology.

Keywords: state-trace analysis, dimensional analysis, Bayes factors, inequality constraints, model selection

Psychologists and neuroscientists are often interested in whether a behavior involves one or more than one cognitive module, representation, process or brain region, and whether this involvement differs between participant populations, stimuli, tasks, experimental manipulations or dependent measures. Although these questions seem straightforward, standard methods used to address them (e.g., Shallice, 1988) have repeatedly been shown to be problematic (e.g., Bogartz, 1976; Busemeyer & Jones, 1983; Dunn & Kirsner, 1988; Henson, 2006; Poldrack, 2006; Wixted, 1990). In this paper we examine an alternative method, called *state-trace analysis* (Bamber, 1979), which avoids many of these problems. Loftus, Oberg and Dillon (2004) characterized state-trace analysis as specifically addressing the question of *dimensionality*, that is, whether one or more than one latent (i.e., not directly observed) variable mediates behavior.

In the first major section of this paper (*State-Trace Analysis: What is it and why use it?*) we provide an informal introduction to state-trace analysis. We explain why state-trace analysis can support more certain inference about dimensionality, particularly for bounded dependent measures such as choice accuracy, where inference can be confounded by floor and ceiling effects (Loftus, 1978). In the next major section (*State-Trace Experiment Design*) we provide guidance on the design of state-trace experiments. We also demonstrate previously undocumented limitations of state-trace analysis, and suggest how experimental designs can be refined to avoid them.

In the final major section (*Statistical Analysis of State-Trace Data*) we develop a new application of Klugkist, Kato and Hoijtink's (2005) Bayesian encompassing prior method to the state-trace analysis of accuracy data. In this section we focus on the rationale for the proposed new analysis, and how it can be applied to support both

inferences about dimensionality and the refinement of experimental designs. Mathematical and computational details are provided in an Appendix. However, it is not necessary for users to implement the computations, as we provide freely available R language (R Development Core Team, 2010) open-source software (see Prince, Hawkins, Love & Heathcote, submitted).

State-Trace Analysis: What is it and why use it?

Loftus (2002) provides a historical context for state-trace analysis, describing it as one of several *equivalence* based techniques for determining the rules by which different combinations of independent variables lead to equivalent states in one or more latent variables. He notes that such techniques can be applied to experimental data from "virtually any area of psychology" (p.382) and cites the color-matching experiments that led to the tri-chromatic (i.e., three-dimensional) theory of color vision as an exemplary application of equivalence methods to behavioral data.

In this paper we illustrate state-trace analysis in a paradigm used by Loftus et al. (2004) to investigate whether cognitive representations of unfamiliar faces provide a basis for accurate responding in a recognition memory task that is not available to other non-face stimuli. However, a variety of examples demonstrate the importance of the question of dimensionality to many other areas of psychology, both theoretical and applied.

Applications of state-trace analysis have been particularly prominent in memory research. In the domain of short-term memory, Lewandowsky, Geiger, Morrell and Oberauer (2010) used state-trace analysis to determine whether a single latent variable can account for the effects of different types of distracters on the accuracy of short-term recall in complex-span tasks. In the domain of long-term recognition memory, Dunn (2004, 2008) investigated whether accuracy for responses classified by

participants as being based on *remembering* or *knowing* (Tulving, 1985) are a function

of different latent variables (see also Henson, 2006, for a discussion of related issues

that arise with functional magnetic resonance imaging, fMRI, data). Brainerd, Wright,

Reyna and Payne (2002) used state-trace analysis to investigate whether separate

direct retrieval and reconstruction processes jointly determine free recall and

associative recall. These examples illustrate the relevance of state-trace analysis to

key theoretical debates in memory research; such as whether forgetting in short-term

memory is due to decay as well as interference (e.g., Oberauer & Lewandowsky,

2008), and whether memory has a single process or dual process architecture (e.g.,

Wixted, 2007).

Applications of state-trace analysis have not been limited to theoretical issues or

effects on the accuracy of memory. For example, state-trace analysis has been applied

to response time data to determine whether a single *general-slowing* factor can

explain age-related differences in location-based and identity-based negative priming

(Verhaeghen & De Meersman, 1998), selective and divided attention tasks

(Verhaeghen & Cerella, 2002) and single vs. dual task performance (Verhaeghen,

Steitz, Sliwinski & Cerella, 2003). Newell, Dunn and Kalish (2010) used a state-trace

analysis of accuracy data to investigate dual (explicit and implicit) system theories of

perceptual classification. In the domain of meta-cognitive skills, Jang and Nelson

(2005) applied state-trace analysis to confidence ratings about recognition memory

accuracy in order to investigate whether ratings made prospectively (i.e., during or

shortly after study) and retrospectively (i.e., during testing occurring some time after

study) have a common basis. Relevant to our main example, Loftus and Harley (2005)

discussed the implications for the accuracy of eyewitness testimony of determining

whether a common latent variable mediates the effects of priming, familiarity, viewing distance and spatial filtering on face recognition.

The forgoing examples explored differences in dimensionality as a function of tasks, stimuli and other experimental manipulations that can, at least in principle, be investigated at the individual participant level. State-trace analysis has also been used with both accuracy and response time measures in order to investigate differences between groups of participants. Several examples underline the relevance of this approach to applied areas.

In the domain of clinical psychology, Haist, Shimamura and Squire (1992) used state-trace analysis to investigate whether the same declarative memory process mediates recall and recognition accuracy for both normal and amnesic participants. In the areas of problem solving and development, De Brauwer, Verguts and Fias (2006) used a state-trace analysis of response time data to investigate whether the representation of multiplication facts differs among children of different ages, and discussed the importance of the answer to this question for the design and evaluation of educational curricula. In the area of perceptual category learning, Newell, Dunn and Kalish (2011) showed through state-trace analysis that deficits displayed by Huntington's and Parkinson's disease patients can be explained by a deficit in a single underlying factor. Brainerd, Reyna and Howe (2009) used reversed associations (a technique closely related to state-trace analysis discussed further below) to investigate the role of dual memory processes in recall over the lifespan (early development, adult and aged) and in neurocognitive impairments. In a recent development with the potential for application to a wide range of disorders, Van den Broeck and Geudens (submitted) showed that state-trace analysis is better able to test for specific deficits in reading than traditional methods making comparisons to matched control groups.

In the examples cited so far state-trace analysis was applied to the same dependent variable measured under different experimental conditions, or for different types of tasks, stimuli or groups of participants (e.g., reading performance for clinical vs. control groups). A less common type of application, which we will call *dependent-variable state-trace analysis*, investigates whether different types of dependent variables are functions of a common underlying latent variable. For example, long-term memory researchers have used state-trace analysis to investigate the relationship between retrospective confidence ratings and recognition accuracy (Busey, Tunnicliff, Loftus & Loftus, 2000; Heathcote, Freeman, Etherington, Tonkin & Bora, 2009), and between retrospective confidence ratings and judgments of the number of occasions on which an item was studied (Hintzman, 2004).

This type of dependent–variable state-trace analysis has also extended beyond recognition memory and behavioral measures. For example, Loftus and Irwin (1998) applied state-trace analysis to investigate whether a common perceptual process mediates different measures of visible and informational persistence. In the domain of neuroscience, Freeman, Dennis and Dunn (2010) applied state-trace analysis to investigate the relationship between the magnitudes of different evoked-response potential (ERP) components, as well as between the frequency of higher and lower confidence responses, in a recognition memory paradigm.

Taken together, these examples demonstrate the relevance of the question of dimensionality, and the applicability of state-trace analysis, to a variety of areas ranging from basic research in perception, attention, short-term and long-term memory, categorization, problem solving and meta-cognition to applications in aging, legal, clinical, educational, human factors and developmental psychology. These examples also make use of most of the dependent measures employed by

psychologists and neuroscientists (i.e., accuracy, ratings, response time, ERP and fMRI). In this paper we focus on the sort of design used by Loftus et al. (2004); within-subjects designs with accuracy as the dependent variable. However, we also discuss implications for other types of state-trace analysis.

The State-Trace Plot

State-trace analysis can be explained with reference to a state-trace plot (see Bamber, 1979, for a formal treatment). In dependent-variable state-trace analysis, where the question of interest is whether two different dependent variables are both a function of the same latent variable, each axis of the state-trace plot corresponds to one of the dependent variables. For example, one axis might represent accuracy and the other confidence ratings (e.g., Busey et al., 2000; Heathcote et al., 2009). When the question of interest is whether measurements of the same dependent variable under different conditions (e.g., for different types of participants, stimuli, or levels of an experimental manipulation) are a function of the same latent variable, each axis represents the results for the dependent variable in one of the conditions. For example, the axes might represent recognition accuracy for face stimuli and non-face stimuli, such as houses (e.g., Loftus et al., 2004).

The axes of the state-trace plot can be thought of as defining a space within which the states of the system under examination can be depicted (i.e., a *state space*). We describe the two conditions or two different dependent variables constituting the axes of the state-trace plot as forming a *state factor*. A point on the state-trace plot is defined by two measurements, one for each level of the state factor. Each measurement is based on responses from multiple experimental trials. For example, suppose in a particular experimental condition with two-alternative forced choice testing the studied face was selected on 67% of test trials and the studied non-face

item was selected on 58% of test trials. Results for this condition are represented as a single point, plotted at $\{x = .67, y = .58\}$ in a state-trace plot where face accuracy (i.e., the proportion of correct responses) is represented on the x-axis and non-face accuracy on the y-axis. Although the state space may have more than two dimensions (e.g., when three or more dependent variables or experimental conditions make up the levels of the state factor) we focus on the two-dimensional case as it has been most widely used in previous state-trace applications.

We illustrate state-trace analysis using data we collected in a recognition memory paradigm similar to that examined by Loftus et al. (2004). These experiments aimed to investigate whether faces are encoded on an extra dimension, commonly called a *configural* dimension (Maurer, Le Grand & Mondloch, 2002), not available to non-face stimuli. In particular, it was hypothesized that both face and non-face stimuli can be encoded in terms of their component features (i.e., on a *featural* dimension), but only faces can be encoded on a configural dimension (i.e., in terms of the relationships among features).

In the original experiment, and in our versions, houses were used as the non-face stimuli, because houses match faces on a range of characteristics, such as being mono-oriented (i.e., usually seen in one particular orientation), familiar and complex. The question addressed by state-trace analysis in this paradigm is whether measurements of memory accuracy for different types of items (houses and faces) arise from a single latent variable, sometimes called *memory strength*, or from two latent variables, which might be characterized as *featural memory strength* and *configural memory strength*.

In these experiments items were presented for study either upright or inverted. Inversion of a mono-oriented object usually results in decreased performance (the

inversion effect; Rock, 1974). However, a number of lines of evidence indicate that inversion particularly impedes configural encoding of faces (Rakover, 2002). Hence, inversion potentially changes the number of encoding dimensions for face stimuli from two (when upright) to one (when inverted). Consistent with these findings, inversion reduces memory performance for face and non-face stimuli, but the reduction is greater for faces (Valentine, 1988, Yin, 1969). We call the interaction between orientation and stimulus type the *differential face-inversion effect*. Figure 1a plots hypothetical data illustrating this effect; inversion causes a large decrease in performance for faces but only a small decrease for houses.

We call the second factor manipulated in Loftus et al.'s (2004) paradigm (i.e., inverted vs. upright study presentation) the *dimension factor*, as dimensionality is potentially influenced by its interaction with the state factor. The variation among points on a state-trace plot induced by manipulation of the dimension variable can be seen as providing a *trace* of the behavior of the system in the corresponding state space. As with the state factor, the dimension factor may have more than two levels, but again we focus on the two-level case, as it is most common in previous state-trace applications.

It is important to note that in state-trace paradigms examining only one dependent variable it is sometimes possible to exchange the attribution of experimental manipulations to state and dimension factors with no material effect on the outcome of the analysis. For example, in Loftus et al.'s (2004) paradigm we could plot accuracy for upright stimuli against accuracy for inverted stimuli. However, in some cases, of which Loftus et al.'s paradigm is one, this exchange threatens validity of the state-trace analysis, and so there is a unique best attribution that should always be used. We examine this issue in the second major section of the paper.

Scale dependent interactions. The differential face-inversion effect is usually tested by an interaction between the state (stimulus type: face vs. house) and dimension (orientation: upright vs. inverted) factors in a general linear model. When significant, such interactions are often labeled as *dissociations* and taken as support for a multi-dimensional process. A variety of types of dissociations have been enumerated (Shallice, 1988). For example, a single dissociation occurs when an experimental manipulation affects a dependent measure in one experimental condition but not in another condition. Figure 1a illustrates a single dissociation, where the face-house difference is apparent for inverted but not upright stimuli. Less commonly, double dissociations are reported, where two experimental manipulations affect measures of each process in opposite directions.

Double dissociations are usually assumed to support stronger inference about dimensionality than single dissociations, particularly when they result in a crossover interaction. However, Dunn and Kirsner (1988) showed that both crossed and uncrossed double dissociations can occur when the experimental manipulations have opposite effects on a single latent variable. They pointed out that a single latent variable could only be ruled out if a third experimental manipulation affects the dependent measures in the other conditions in the same direction. This outcome, which they labeled a reversed association, is equivalent to observing a state-trace plot indicative of more than one latent variable.

In order to provide a concrete illustration of these issues, we focus on one reason for uncertainty about dimensionality inference common to all types of dissociation; *scale-dependent* interactions, with the most commonly recognized causes being floor or ceiling effects. Scale dependent interactions can occur when the function mapping latent to dependent variables (sometimes called the *response*

function) is non-linear. Loftus (1978) details the effect of several types of response functions that account for bounds on dependent variables (e.g., an S shaped function accounting for the floor and ceiling in an accuracy measure, see Figure 2a). Figure 1a illustrates a scale-dependent interaction occurring because performance (i.e., accuracy) for both upright houses and faces is near ceiling.

One approach to floor and ceiling effects is to transform the dependent measure in a way that removes the bounds, such as a logit or probit transformation (i.e., inverse cumulative logistic or normal probability transformations) of a probability correct measure. Although easily implemented, this approach requires a strong assumption about the exact mathematical form of the response function, and that the form of the function is the same for different levels of the state factor. Another approach is to manipulate the overall difficulty of a task to move the observed data to the middle of the range of the dependent variable. However, even where it is possible to avoid floor and ceiling effects in this manner, scale dependent interactions may still occur. This is illustrated by the response function shown in Figure 2b, which adds to floor and ceiling effects lower sensitivity to changes in the latent variable in a region that maps to the middle of the response range.

Monotonicity. State-trace analysis avoids these problems because it tests an ordinal hypothesis that makes only a weak assumption about the response function: that it is monotonic. For example, monotonicity implies that an increase or decrease in the strength of a memory trace should lead to a corresponding increase or decrease in recognition accuracy, without making any assumptions about the magnitudes of these changes, or to no change in recognition accuracy (e.g., when performance is at ceiling or floor respectively). The ordinal hypothesis is based on the fact that, if both state and dimension effects are mediated by a single latent variable mapped monotonically

to the dependent variable, a plot of results for one level of the state factor against results for the other level of the state factor for each level of the dimension factor (i.e., a state-trace plot) must also be monotonic (Bamber, 1979).

Consequently, determining whether there is one or more than one mediating latent variable is accomplished by determining if the state-trace plot is monotonic. This determination can be accomplished graphically, as monotonicity implies that an always-increasing or always-decreasing line can join all points in the state-trace plot. Equivalently, monotonicity implies that points on one axis of the state-trace plot have the same order as the order of the points on the other axis (and so can be joined by an always-increasing line), or have the opposite order (and so can be joined by an always-decreasing line). Note that a state-trace plot must contain more than two points to be diagnostic of dimensionality, as these conditions are always true for two points. That is, the order of two points on one axis must always be the same as, or opposite to, the order of two points on the other axis.

The graphical approach to state-trace analysis has the weakness that it takes no account of measurement error: violations of monotonicity due to measurement error may be mistaken for evidence for more than one latent variable. Conversely, measurement error may cause observed monotonicity when there is mediation by more than one latent variable. Several statistical methods have been used to account for measurement error in the evaluation of state-trace monotonicity (see Newell & Dunn, 2008). In some cases these methods require further assumptions beyond those made by state-trace analysis. In the third major section of this paper we review some of these methods and propose a new method that aims to minimize added assumptions by taking the same ordinal approach as state-trace analysis.

The trace factor. One of the key methodological aspects differentiating state-trace (and reversed association) experiments from traditional dissociation experiments is the addition of a third factor, which we call a *trace factor*. In Loftus et al.'s (2004) design the trace factor was the amount of time spent studying each item. In designs with a two-level dimension factor a trace manipulation must be included to measure more than two points on the state-trace plot, and hence to enable inference about dimensionality. If the dimension factor has more than two levels a trace factor is not necessarily required, but it can be desirable for reasons we discuss in the next major section.

Figure 1b illustrates a state-trace plot for a design with a two-level trace factor. The figure was created by taking the data plotted in Figure 1a and adding data from new conditions that allow a shorter time for the study of each item (and, hence, result in poorer performance). The four pairs of coordinates defining the four points in Figure 1b correspond to results from eight experimental conditions created by a factorial crossing of the state factor (face vs. house stimuli), the dimension factor (inverted vs. upright orientation) and the trace factor (short vs. long study duration). Typically, lines are drawn on the state-trace plot to join results for conditions differing only on the trace factor in order to visually group them together. We will call such lines (e.g., the solid lines in Figure 1b) *data traces*.

The dotted line in Figure 1b is a monotonic curve that almost manages to connect all of the points in the state-trace plot. There is some error, because the curve does not quite pass through the lower right or upper left points (short-study-time upright items and long-study-time inverted items, respectively). These imperfections illustrate the statistical problem we address in the third major section of this paper: given that the data in Figure 1b contain some measurement error, should we conclude

that a monotonic curve describes the data well, and hence that a single latent variable

underlies memory for faces and for houses, or conclude that the monotonic curve

misses the data, and hence that more than one latent variable is required?

Note that in some applications of state-trace analysis the levels of the trace

factor correspond to individual participants (a random effect) rather than to

experimental conditions (a fixed effect). For example, Haist et al. (1992) plotted the

probability of recall against the probability of recognition (the state factor) for

amnesic and control groups (the dimension factor), with each point on the plot

representing results for one participant. Similarly, in De Brauwer et al. (2006) the

state factor was response time for different types of multiplication problems, the

dimension factor was age group and each point on the plot represented results for an

individual participant (for similar applications see also Verhaeghen & Cerella, 2002;

Verhaeghen & De Meersman, 1998, and Verhaeghen et al., 2003). Although we do

not focus on such applications in this paper, we do briefly address them in light of our

findings in the General Discussion.

State-Trace Experiment Design

In this section we focus on the design of state-trace experiments where the

trace-factor corresponds to a fixed effect. This makes the trace factor amenable to

experimental manipulations that can be used to refine an experimental design to suit

state-trace analysis. It is our experience that design refinements are often necessary to

produce state-trace results that are clearly diagnostic of dimensionality. The

experiments whose results we report here were run for just such a purpose, in an

attempt to refine Loftus et al.'s (2004) design.

Diagnosing Dimensionality

Even when a state-trace plot contains more than two points and there is no measurement noise it still may not be diagnostic of dimensionality, as we illustrate in Figure 3. The figure contains state-trace plots of hypothetical data from the same design as illustrated by Figure 1b (i.e., a design in which state, dimension and trace factors all have two levels). The points a and b are from the first level of the dimensional factor (e.g., inverted study items) for both levels of the trace factor (e.g., shorter and longer study time respectively), and their data trace is a thick solid line. Points A and B are from the second level of the dimension factor (e.g., upright study items), for both levels of the trace factor (e.g., shorter and longer study times respectively), and their data trace is a thick dashed line. In discussing this figure we assume that there is no measurement error, and so the data points fall exactly on their true values. We return to the issue of measurement error in the next major section.

If we were able to systematically vary all latent variables across all of their ranges, we would observe all possible values of the dependent variables corresponding to each axis of the state-trace plot. If this entire set of observations were represented on the state-trace plot and measured without error, the result would be the true *state-trace* of the system. Under very general assumptions, any data-generating process mediated by just one latent variable will always have a monotonic true state-trace. That is, the set of points created by moving the underlying latent variable across its whole range will make a curve in the state-trace plot that always increases or always decreases. For example, the thin dotted lines in Figures 3a and 3b show monotonic true state-traces. In Figures 3a and 3b, the data points are also monotonic, having the same order on both axes: in Figure 3a this order is {a,b,A,B}, and in Figure 3b it is {a,A,b,B}.

Note that in Figure 3b a monotonic line joining all data points does not follow the data traces. In terms of our recognition memory example, the monotonic line joins the short-study-time inverted data, followed by the short-study-time upright data, long-study-time inverted data, and then the long-study-time upright data. This illustrates that the data traces (i.e., the solid and dashed lines) are only a graphical convenience and do not necessarily reflect the true state-trace.

Note that neither the true state-trace of a one-dimensional system nor the curve joining points with the same order on both axes need be straight, or have constant curvature, or satisfy any other condition related to smoothness. Such conditions may be important due to theoretical considerations outside the scope of state-trace analysis. However, state-trace analysis itself makes no such assumptions; it is based purely on monotonicity.

If the data points in a state-trace plot are non-monotonic a one-dimensional system cannot have produced them. However, the converse does not necessarily hold; if the data points in a state-trace plot are monotonic this does *not* necessarily imply that a one-dimensional system generated them, even if there is no measurement noise. As illustrated by Figure 3, one further condition is required: *overlap of the data traces on at least one axis*.

Figures 3c and 3d show results for a two-dimensional system, illustrated by two true state-traces (thin dotted lines). The first true state-trace applies for the first level of the dimension factor (e.g., inverted stimuli) and the second applies for the second level of the dimension factor (e.g., upright stimuli). The data points in Figure 3d are non-monotonic; the order of points on the x-axis, {A,a,B,b} differs from the order on the y-axis, {a,A,b,B}. In contrast, the data points in Figure 3c are monotonic;

they have the same order {a,b,A,B} on both axes, despite the fact that the data in this figure come from a two-dimensional data generating process.

In fact, the data points in Figure 3c are identical to those in Figure 3a – just the interpretation is different. In Figure 3a, we interpreted the data points as belonging to a single monotonic true state-trace (supporting a one-dimensional system), but Figure 3c shows they can equally belong to two true state-traces, and so are consistent with a multi-dimensional system. Figures 3a and 3c illustrate that when data traces fail to overlap, state-trace analysis fails to diagnose dimensionality.

Such failures can usually be remedied by changing the levels of the trace factor. For example, suppose that the trace factor in Figure 3 was the duration for which items are studied, as it was in our experiments. Because increased study duration increases accuracy, the accuracy for point *b* in Figure 3a can be increased by increasing study duration, and the accuracy of point *A* can be decreased using a shorter study duration. This leads to the diagnostic state-trace plot shown in Figure 3b, in which the data traces overlap on both axes. Even though overlap on just one axis is sufficient, overlap on both is desirable as it increases the number of data points that can contribute to the detection of violations of monotonicity.

In Figure 3d the non-diagnostic result in Figure 3c was rectified by decreasing performance for both conditions within the second level of the dimension factor and increasing performance for both conditions within the first level of the dimension factor. Once again the resulting state-trace plot is strongly diagnostic of dimensionality, with data traces that overlap on both axes, and in this case it clearly indicates more than one dimension. If the trace factor were study duration in these examples, this result could be achieved by using longer study durations for level one than level two of the dimension factor (i.e., longer study for inverted than upright

items). Note that there is no requirement to use a fully factorial design in state-trace analysis. Indeed, as we discuss below, a fully factorial design can be very inefficient.

Choosing a Trace Factor

Much of the art of running a successful state-trace experiment is in choosing and calibrating an appropriate trace factor. We consider three design issues related to the trace factor. First, if non-monotonicity in a state-trace plot is to be unambiguously attributed to the interaction of the state and dimension factors, the trace factor must have a monotonic effect within each level of the dimension factor. In our recognition memory example this means that increased study time must lead to increased recognition accuracy in all four conditions. Other state-trace analyses of memory have used retention interval as the trace factor (Haist et al., 1992) or study-item repetitions (Bamber, 1979; Hintzman, 2004; Jang & Nelson, 2005), which are also assumed to have a monotonic effect on accuracy. The statistical procedures that we develop in the next major section provide a method of testing whether the trace factor effect is monotonic.

The second design issue requires trace levels to be chosen so they maximize overlap. Commonly, the dimension factor affects results for each state in the same direction. In such cases, different trace-factor levels can be used within each level of the dimension factor to compensate for the effect of the dimension manipulation, and hence to maximize overlap. For example, in our experiments inverted items were harder to remember than upright items. Hence, it makes sense to give generally longer study durations to the inverted items, so that accuracy on the inverted and the upright items becomes about equal, and so data traces overlap.

Thirdly, the effects of potential interactions between the trace factor and other factors must be considered. For example, it has been suggested that inversion and very

brief exposure have similar effects (Valentine, 1988), so that upright faces may be encoded only in terms of features (i.e., in a one-dimensional manner) below some minimum study duration. If upright and inverted data traces overlap only when study duration for the upright condition is below this minimum, the state-trace plot will be monotonic, even when face processing is multi-dimensional at longer durations. This issue may account for Loftus et al.'s (2004) finding in their first experiment of a monotonic state-trace plot, as only the shortest study duration (17ms) for upright items overlapped with the longest duration (250ms) for inverted items. Our alterations to Loftus et al.'s design aimed to address this issue by increasing the minimum study duration for upright items.

Refining a State-Trace Experiment: A Case Study

As previously discussed, a good state-trace design produces data traces that overlap on one, and ideally both, axes. When the dimension factor causes a large performance difference, using the same trace levels within each dimension level is inefficient, as only data in the overlapping region contributes evidence about dimensionality. In such cases it is better to use different levels of the trace factor for each dimension level (i.e., a non-factorial design).

For example, the design used by Loftus et al. (2004) in their first experiment produced a very strong dimension effect – items studied upright were much more accurately recognized than those studied inverted. In order to achieve data trace overlap, Loftus et al. used six study durations (from 17ms to 250ms) in a fully factorial design with the state and dimension factors. Although this range of study durations was sufficient to cause overlap for both houses and faces, the overlap was minimal.

We largely replicated this design with our first experiment, which we will call the *upright-test* experiment. However, in contrast to Loftus et al. (2004), we used longer study durations for inverted (256-2048ms) than upright (33-256ms) items. As shown in Figure 4a (see caption for further experimental details), the longer study times for inverted items overcame the performance decrement caused by inversion, and so produced almost perfect data-trace overlap. These durations were selected based on an earlier pilot study that was not so successful, illustrating the process of iterative refinement of a design.

Selection of an appropriate trace factor is subject to tradeoffs enforced by restrictions on the range and number of trace levels. If too many levels are used it is practically difficult to get enough trials at each level for precise measurement. For this reason we used four trace levels rather than Loftus et al.'s (2004) six. However, if too few levels are used it becomes difficult to detect non-monotonicity. For example, Figure 4c shows a state-trace plot from our third experiment (described further below) that is clearly non-monotonic with the four points per data trace that were measured. In contrast, if only the lowest and highest points in each data trace had been measured we might have come to a quite different conclusion, because a monotonic line can join these four extreme points.

Although it is difficult to specify how many trace levels are required in general, the example in Figure 4c suggests a minimum of three. Once the number is chosen, the spacing of the levels must be selected. The optimal spacing depends on the nature of any non-monotonicity that is present, but as this is not known in advance we recommend an iterative strategy, starting from a spacing that results in an approximately equal change in performance between each level. For example, Loftus et al. (2004) found that accuracy increased linearly with the logarithm of study

duration, so we selected durations that increased by a constant multiple in our experiment, resulting in the relatively even spacing displayed in Figure 4. Unequal changes in performance between data points are undesirable because very closely spaced points generally make a smaller contribution to exploring the behavior of the underlying system than more widely spaced points. Closer spacing also makes observing non-monotonicity within a data trace due to sampling error more likely.

Reducing the dimension effect. An alternative approach to maximizing data-trace overlap is to reduce the magnitude of the dimension factor effect. Confounding the dimension manipulation with another manipulation that has the opposite effect can do this. In fact, the reason that Loftus et al. (2004) obtained such a large difference between upright and inverted stimuli was that they confounded their inversion manipulation with the encoding-specificity effect (Tulving & Thomson, 1973), but in a way that increased rather than reduced the dimension effect. Encoding-specificity is a well-known and robust effect (Nilsson & Gardiner, 1993) whereby memory for a stimulus (including faces, Rakover & Teucher, 1997; Yin, 1969) improves when there is a better match between its study and test encodings. Loftus et al. had participants study stimuli either upright or inverted, but always tested them upright. Hence, the upright condition was advantaged even more than usual, because study and test encodings had a better match than in the inverted condition.

Loftus et al. (2004) applied inversion at study but not at test to investigate the hypothesis that evidence for a multi-dimensional face representation emerges only when inversion affects the processing of faces that are already stored in memory. As the faces they used in their first experiment were unfamiliar, they were not in memory before study, and so their hypothesis predicts a monotonic state-trace plot for this design. Arguably the same prediction should hold whether all images are tested

upright or all are tested inverted, so in our second experiment, which we call the *inverted-test* experiment, we tested all stimuli inverted. This reverses the confounding effect of encoding specificity, and so should reduce the difference in accuracy between upright and inverted stimuli.

We took advantage of the reduced inversion effect in the inverted-test experiment by using longer study durations for the upright condition (67-512ms) in order to avoid the potentially confounding effects of very brief study durations discussed earlier. We also kept the same study durations for the inverted condition (267-2048ms) as in our upright-test experiment. However, as shown by the state-trace plot of data from this experiment (Figure 4b), our choice of upright durations under-estimated the power of the encoding specificity effect relative to the inversion effect. With the durations we used, performance was actually *better* in the inverted than upright condition. These results suggested that the inversion and encoding specificity effects are about equal in magnitude in our design, and so the best overlap would have been obtained using equal durations for the upright and inverted conditions. Prince and Heathcote (2009) used this design and obtained greatly improved data-trace overlap, again illustrating the process of iterative design refinement.

Accuracy measures and multiple baselines. Within each level of the state factor in the upright-test and inverted-test experiments, accuracy was measured relative to a common baseline condition for all levels of the trace and dimension factors. That is, accuracy for houses was measured relative to a false alarm rate (FAR; i.e., the probability of wrongly saying a test item was studied) specific to houses, and accuracy for faces was measured relative to the FAR for faces.

In Figure 4 we used HR – FAR as the accuracy measure, where HR is the hit rate (i.e., probability of correctly saying a test item was studied). When there is a

common baseline the order of accuracy results will be the same for any reasonable measure, such as (HR – FAR) / (1 – FAR), which is similar to the measure used by Loftus et al. (2004), or the *d'* measure used in signal detection theory (i.e., z(HR) – z(FAR), Macmillan & Creelman, 2005, where z() is a probit transformation). Note that all of these measures are monotonically related. It is a notable strength of state-trace analysis that the dimensionality indicated by a state-trace plot can be unaffected by using different monotonically related accuracy measures.

However, state-trace analysis may not be invariant in this way for a design where accuracy is assessed against different baselines for different dimension factor levels. Our third experiment, with results shown in Figure 4c, is a case in point. As has been standard in the differential face-inversion effect literature since its inception (Yin, 1969), study and test orientations were matched in this experiment. That is, in this *matched-test* experiment, an item that was studied upright was tested upright and an item that was studied inverted was tested inverted. As the difference between upright and inverted stimuli is clear to participants, they might decide to use different criteria to make recognition decisions for upright and inverted test items.

The usual method of addressing such potential differences in response bias is to measure accuracy relative to separate baselines for upright and inverted conditions. That is, some test items that were not previously studied are presented inverted and some upright, and their separate false alarm rates are used to calculate accuracy for inverted and upright conditions respectively. However, different baselines for the dimension (or trace) factors mean that state-trace inference may no longer be invariant across different accuracy measures. In the matched-test experiment, for example, this can occur because the order of results among upright and inverted conditions can depend on the scale on which accuracy is measured.

As an existence proof, suppose FAR = 0.2 and HR = 0.85 for upright items and FAR = 0.35 and HR = 0.95 for inverted items. If we measure performance using HR – FAR, upright performance (0.65) is greater than inverted performance (0.6), but the order is reversed in *d'* (1.88 vs. 2.03 for upright and inverted respectively) and (HR – FAR) / (1 – FAR) (0.81 vs. 0.92). If the two baselines happen not to differ empirically this problem might not arise. Our third experiment was designed to check if this state of affairs held in the matched-test design; unfortunately, it did not, so we do not consider this experiment further. Note that a baseline that collapses over upright and inverted false-alarm rates does not address the problem, as a criterion shift would still affect hit rates.

Differing baselines for each state factor level (e.g., for houses and faces) do not compromise a state-trace analysis of accuracy. This is because the order of accuracy results within each state-factor level is unaffected by the baseline, and so monotonicity, which depends only on the order of conditions, is unaffected. Hence, when only one of the state and dimension factors has a common baseline, the factor with the common baseline should always be used as the dimension factor. Alternatively a different type of test that does not require a baseline, such as a two-alternative forced-choice test, can be used. The analysis methods we develop in the third major section of this paper work with both two-alternative forced choice and single item testing.

Summary and Recommendations for Experiment Design

In a state-trace experiment the trace factor is typically of lesser theoretical interest, as it is included in the design only to sweep out different levels of performance. However, it has an important role to play in obtaining overlapping data traces. Overlapping data traces are essential for enabling a state-trace plot to be

diagnostic of dimensionality. Hence, a successful state-trace experiment requires careful selection and calibration of the trace factor. Often, the most efficient way of ensuring data-trace overlap is to use different levels of the trace factor between levels of the dimension factor in order to counteract the effect of the dimension factor.

The number and spacing of the levels of the trace factor must also be calibrated in order to best detect any non-monotonicity in the state-trace plot. At least three trace levels per dimension level, with each set of trace levels having approximately evenly spaced effects on performance, are recommended. In order to unambiguously interpret any non-monotonicity (and hence evidence of multi-dimensionality) as a result of the interaction of state and dimension factors it is also desirable that the trace factor is known to have, or can be shown to have, a monotonic effect on performance.

Statistical Analysis of State-Trace Data

In this section we propose a new method for analyzing state-trace plots of accuracy data, using Klugkist, Kato and Hoijtink's (2005) Bayes-factor method of selecting among models defined by inequalities. Technical details are provided in an Appendix. We also provide an example of how the method can be applied, using data from our upright-test and inverted-test experiments. First, however, we motivate our new approach by discussing the potential shortcomings of some existing methods.

Existing Statistical Methods

Perhaps the most common method for state-trace analysis is visual inspection of state-trace plot based on data averaged over participants. Measurement error is quantified by confidence intervals, with small intervals, and hence a clear conclusion, obtained by averaging over a large number of participants. For example, Loftus et al., (2004) averaged over 366 participants in their first experiment. They also quantified

evidence for monotonicity using Spearman's Rho (ρ), a rank based measure of

association (monotonicity implies $\rho = 1$). They showed that a criterion of $\rho = 1$

reliably identifies a one-dimensional system when results are averaged over

participants simulated by varying the system's parameters. Their method results in

state-trace plots that are monotonic for every simulated participant.

However, when we repeated their analysis with added measurement error,

which creates more realistic data where some participant's simulated state-trace plots

are non-monotonic, this test became unreliable. A value of $\rho = 1$ rarely occurred, and

even when an optimal ρ criterion was used, the monotonic vs. non-monotonic

classification was unreliable unless measurement error was small. Moreover, even

when measurement error was small, optimal criteria on ρ varied substantially across

the simulations, indicating that this test might be difficult to apply in practice.

Figure 4 exemplifies the type of analysis used by Loftus et al. (2004), as

applied to our three experiments. To facilitate congruency between *inference-by-eye*

and conventional standards for null-hypothesis statistical testing (NHST), we added to

Figure 4 *least-significant difference* intervals (see figure caption for details), so that

non-overlapping bars indicate a significant difference at the .05 level. Based on this

method, a one-dimensional system cannot be rejected for the upright-test (Figure 4a)

and inverted-test (Figure 4b) experiments. However, the trends in the upright-test

experiment suggest non-monotonicity, so perhaps the non-significant results are

attributable due to a small sample size (see figure caption), and hence a lack of power.

In the matched-test experiment (Figure 4c) this method supports a multi-dimensional

system, although, as we have noted, interpretation of accuracy results for this

experiment are problematic. All of these conclusions are consistent with the

Spearman's ρ values in the figure, which are relatively close to one for the upright-

test and inverted-test experiments and substantially less than one for the matched-test experiment.

Averaging and monotonicity. Unfortunately, the averaging that underpins the reduction in measurement error necessary for the preceding types of analyses can be problematic. First, averaging can produce results that depend on the type of accuracy measure used. More fundamentally, neither monotonicity nor non-monotonicity is necessarily preserved under averaging. For example, Figure 5, which uses the same design and notation as Figure 3, shows hypothetical data from two participants (left and middle columns) and the average over those two participants (right column). The first row shows that, even when both participants have non-monotonic state-trace plots, the average plot can be monotonic. The second row shows the opposite; both participants are monotonic but the average is non-monotonic (see Newell et al., 2010, for an example of this phenomenon). In all cases the individual participant data are plausible – data traces overlap on at least one axis and an increase in the trace factor within each level of the dimension factor causes an increase on both axes. Similarly plausible examples can be constructed where a mixture of monotonic and non-monotonic participants leads to either a monotonic or non-monotonic average.

These examples do not prove that a state-trace analysis performed on average data is always, or even very often, misleading. However, they do show that averaging has the potential to be misleading. In our experiments, where all factors are within-subjects, a complete state-trace analysis of each individual participant's data provides a possible solution. Results of individual analyses can then be safely aggregated, except where there is evidence for strong individual differences, such as individual participants with outlying results or larger sub-groups with different dimensionality. Eliminating outlying participants from aggregates or aggregating over only sub-

groups of participants can address such problems. In the following *Example Analysis* section we present results for individual participants using figures that make it easy to detect such issues.

Although individual analysis avoids distortion due to averaging, it can introduce other issues. First, it cannot be used for state-trace designs with between-subjects factors. Second, higher levels of measurement noise usually attend individual analysis, necessitating more observations per participant to get precise estimates. Finally, within the NHST framework, a non-significant test result due to a lack of power can be confused with support for a null hypothesis. Although this confusion must be avoided in all contexts, it is particularly salient for individual analysis, due to higher levels of measurement noise. For NHST analyses of state-trace data, where a one-dimensional model is the null model, this confusion leads to a bias against a multi-dimensional finding.

Null hypothesis statistical testing. A range of model-selection techniques that have been applied to average state-trace data can also be applied to individual data. Newell and Dunn (2008) used monotonic regression (Barlow, Bartholomew, Bremner, & Brunk, 1972), with goodness-of-fit tests based on bootstrap estimates of null-hypothesis distributions (McLachlan & Peel, 2000). Other NHST approaches, such as comparing maximum likelihood fits amongst order-constrained models, can also be applied to individual data. For example, a particular one-dimensional model (i.e., a model that enforces a particular order for all points on both axes) could be used as the null, and its fit compared to a model without order constraints, or to models that enforce only the order dictated by the trace factor within each data trace.

We do not pursue the NHST approach here because it is not clear how it can take account of differences in model complexity. The class of multi-dimensional

models is clearly much more flexible than the class of one-dimensional models, as it has fewer order (inequality) restrictions, and so is likely to fit data better by chance. Methods of accounting for model complexity due to differences in number of parameters (see Pitt & Myung, 2002) cannot be applied, as one-dimensional and multi-dimensional models have the same number of parameters. As Klugkist, Laudy and Hoijtink (2005) point out: "For inequality constrained hypotheses the number of parameters does not reflect the complexity or size of the parameter space of the model" (p. 478).

A second reason that we did not pursue an NHST approach is that we believe it is inherently unsuited to state-trace analysis. Loftus (2002) suggested that state-trace analysis should be used to determine the simplest explanation of a phenomenon. Hence, state-trace analysis needs a statistical method that provides estimates of the probability that the simpler model (e.g., a one-dimensional model) is sufficient, rather than focusing only on estimating the probability that a more complex model (e.g., a multi-dimensional model) is required. NHST casts the simpler model in the role of the null hypothesis. A significant result can provide evidence against the null model. However, NHST cannot provide evidence for the null model, as a failure to reject the null hypothesis might either be due to the null being true or to a lack of power, which is particularly likely in individual participant analysis (see Wagenmakers, 2007, for other problems with NHST). Gallistel (2009) summarizes these and related issues and suggests that Bayesian analysis provides a potential method for solving both problems – confirming the null and addressing model complexity.

Proposed Method

We propose a method of state-trace analysis based on quantifying evidence for different models of orderings of results over experimental conditions. The models

assume independent binomial distributions for the binary choice data within each condition, and all have the same number of parameters. That is, there is one binomial probability parameter for each condition. For our upright-test and inverted-test experiments with four levels of study duration, for example, there are 18 parameters, corresponding to the $2 \times 2 \times 4 = 16$ hit rates in the conditions testing studied items and two false-alarm rates for conditions testing unstudied items (i.e., false-alarm rates for faces and houses).

The models differ in terms of inequality constraints (i.e., orders) placed on the set of binomial probability parameters (θ). In our application, where each state has a separate baseline condition, $\theta = \{\theta^F_i, \theta^H_{i,j,k}\}$, where the superscript indicates a false-alarm (F) or hit (H) rate parameter, and the subscripts indicate the state ($i = 1, 2$), dimension ($j = 1, 2$) and trace ($k = 1\ldots4$) factor levels. Although the labeling of the state and dimension factor levels is arbitrary, the trace factor levels are assumed to be labeled in the order predicted to produce increasingly accurate performance (e.g., in order of increasing study duration). In a design without baselines (e.g., two-alternative-forced choice testing), $\theta = \theta^H_{i,j,k}$.

We examine four ordinal models that are diagnostic either of dimensionality or of the need for design refinement. The first diagnostic model, the non-trace (NT) model, is defined as violating the order dictated by the trace factor. In our application, for example, the order predicted by the trace factor (study duration) is $\theta^F_i < \theta^H_{i,j,1} < \theta^H_{i,j,2} < \theta^H_{i,j,3} < \theta^H_{i,j,4}$. The non-trace model violates this order for one or more of the state and dimension conditions. Evidence for the non-trace model suggests that the trace factor manipulation did not have a monotonic effect, and so the trace factor manipulation may be in need of revision. The same logic applies to designs with no baseline, except the constraint involving θ^F_i is dropped.

The remaining three diagnostic models are special cases of the complement of the non-trace model, which we call the trace (T) model (i.e., a model in which trace-factor order applies for all state and dimension factor levels). As the trace model implies that $\theta^F_i < \theta^H_{i,1,1}$ and $\theta^F_i < \theta^H_{i,2,1}$ we need only consider constraints on the order of the hit-rate parameters for the remaining three diagnostic models. Hence, designs with and without baselines for each state can be treated the same. The second diagnostic model, the multi-dimensional (MD) model, is defined as having a different order for the parameters associated with each state-factor level. Evidence for this model suggests a system with more than one latent variable.

The final two diagnostic models are special cases of the complement of the multi-dimensional model within the trace model, which we call the monotonic (M) model (i.e., a model in which parameters have the same order within each state-factor level). In particular, the third diagnostic model, which we call the no-overlap (NO) model, is defined as having one of the two orders indicating that there is no overlap between data traces (e.g., $\theta^H_{i,1,4} < \theta^H_{i,2,1}$ or $\theta^H_{i,2,4} < \theta^H_{i,1,1}$). Evidence for this model suggests that monotonicity is not diagnostic of dimensionality, and hence the need for a refinement in the experimental design to increase overlap. The fourth diagnostic model, which we call the uni-dimensional (UD) model, is the complement of the no-overlap model within the monotonic model, so is defined as having any of the remaining monotonic orders. Evidence for this model suggests a system with one latent variable.

The union of the four diagnostic models contains all possible parameter orders. Each model differs tremendously in its complexity, and hence in its ability to fit data by chance. One way of quantifying the complexity of ordinal models is by counting the number of orders they contain. A model with no order constraints on its

p parameters, which we call the unrestricted (U) model, contains *p*! orders. For example, consider a simple design with no baseline and two levels for the state, dimension and trace factors, in which *p* = 8. There are 40,320 possible orders in the unrestricted model, but only 36 of these satisfy the constraints of the trace model. Of those 36, only four satisfy the uni-dimensional model, 30 satisfy the multi-dimensional model, and the remaining two satisfy the no-overlap model. In this case the non-trace model is more complex than the multi-dimensional model by a factor of greater than 1000. The multi-dimensional model is 7.5 times more complex than the uni-dimensional model, which in turn is twice as complex as the no-overlap model. These factors increase and diverge for more complicated designs (see Appendix for general formulae), underlining the need to take account of model complexity.

Bayes-factor model selection. A (BF; Kass & Raftery, 1995) is the ratio of the marginal probability of the observed data given one model (M_i) divided by the marginal probability of the observed data given another model (M_k): $BF_{i,k} = m(\mathbf{D}|M_i) / m(\mathbf{D}|M_k)$. The marginal probability equals the likelihood of the data given a model with parameters θ, $f(\mathbf{D}|M, \theta)$, integrated over the prior distribution of the parameters, $p(\theta|M)$: $m(\mathbf{D}|M) = \int f(\mathbf{D}|M, \theta) \, p(\theta|M) \, d(\theta)$. We use Bayes factors to select among the four diagnostic models because, as Myung, Karabatsos and Iverson (2008) state: "Bayes factor-based model selection automatically adjusts for model complexity" (p.314). The aim of Bayes factor model selection is to find the model with the highest posterior model probability. This is the probability, given observed data **D,** that one model (M_i) amongst a set of two or more models, is the true model: $p(M_i|\mathbf{D})$.

For a set of models M_i, $i = 1 \ldots m$, that are assumed to have an equal prior probability of being the true model[1], the posterior model probability for M_i is calculated using each model's Bayes factor all relative to the same model, M_k:

$p(M_i \mid \mathbf{D}) = BF_{i,k} / \sum_{j=1}^{m} BF_{i,k}$. Note that, as the marginal probability for M_k cancels in this ratio, the posterior model probability is the same for different choices of M_k. Note also that, unlike NHST probabilities, both high and low posterior model probabilities are interpretable, as evidence for and against a model respectively.

Bayes factors can be difficult to estimate because the integration required to calculate marginal probabilities is over a high-dimensional space (e.g., 18 dimensions for our experiments). However, when the models being compared have the same parameters, differing only in that one (M_i) is an order-constrained special case of the other (M_k), Klugkist, Kato and Hoijtink (2005) showed that the Bayes factor can be estimated based only on Monte-Carlo (MC) methods that generate random samples from the prior and posterior of the less restricted or *encompassing* model (see Andrieu, de Freites, Doucet & Jordan, 2003, for an overview of methods used to obtain MC samples). All that is required are estimates of the proportion of the prior ($\hat{\pi}$) and posterior ($\hat{\Pi}$) MC samples from the encompassing model (i.e., M_k) that obey the order constraints dictated by the more restricted model (i.e., M_i). The Bayes factor is approximated by the ratio of these two proportions: $BF_{i,k} \approx \hat{\Pi} / \hat{\pi}$.

This method of estimating Bayes factors is made both conceptually and computationally simpler by assuming independent uniform priors for each parameter

---

[1] This assumption does not imply that differences in model complexity (i.e., in the ability of one model to fit data generated by other models) are not taken into account. An increase in the prior probability of a model being true (e.g., based on theoretical grounds or previous findings) causes an increase in its posterior probability, whereas an increase in the probability that a model fits the data by chance (given the prior) causes a reduction in its posterior model probability, all other things being equal.

in the unrestricted model. Such a uniform prior corresponds to the assumption that, before the data are observed, all orders are equally likely. Other priors could be used, but as long as they allow some non-negligible probability for all orders, they will produce the same Bayes-factor estimates (Klugkist, Kato & Hoijtink, 2005) making results insensitive to the choice of prior (see Liu & Aitkin, 2008, for a discussion of the problem of prior sensitivity in other applications of Bayesian methods).

A uniform prior has the computational advantage that prior proportions can be computed analytically (see Appendix), avoiding the computational expense of using MC samples. A second computational advantage follows from the fact that a uniform distribution is a special case of the beta distribution, which is the conjugate prior for the binomial distribution. This implies that the posterior also has a beta distribution (see Appendix for details). Fast algorithms are available in most statistical packages for obtaining independent MC samples from beta distributions, minimizing the computational expense of estimating posterior proportions based on MC samples from the posterior distribution of the unrestricted model.

Figure 6 illustrates the procedure by which we estimate posterior proportions for the four diagnostic models. First, MC samples are generated from the posterior of the unrestricted model (represented by a unit rectangle in Figure 6a). A count is made of the number that violate the trace-model order, enabling estimation of the posterior proportion $(\hat{\Pi}_{NT,U})$, and hence the Bayes-factor ($\widehat{BF}_{NT,U} = \hat{\Pi}_{NT,U} / \pi_{NT,U}$), for the non-trace model relative to the unrestricted model (note that by definition $\pi_{NT,U} = \Pi_{NT,U} = BF_{U,U} = 1$). In theory, estimates of the Bayes factors for the remaining three diagnostic models can be based on counts of the number of trace-model MC samples (represented by the oval in Figure 6a) that follow the orders

dictated by each. However, this method is usually impractical, because posterior

trace-model MC samples may be obtained only rarely.

To avoid this inefficiency we use a type of Markov Chain Monte Carlo

(MCMC) algorithm, an order-constrained Gibbs sampler (Gelfand, Smith & Lee,

1992, see Appendix for details), to approximate a sequence of MC samples from the

posterior of the trace model. The trace model is represented by the unit rectangle in

Figure 6b, with the proportion of monotonic posterior trace-model MC samples

represented by the oval. The area outside the oval represents the proportion of

posterior multi-dimensional model MCMC samples relative to the trace model

($\hat{\Pi}_{MD,T}$). The two regions within the oval represent the proportions of posterior no-

overlap model ($\hat{\Pi}_{NO,T}$) and uni-dimensional model ($\hat{\Pi}_{UD,T}$) MCMC samples.

These three posterior proportions relative to the trace model are multiplied by

the posterior trace-model proportion obtained from unrestricted posterior model MC

samples (Figure 6a) to obtain estimates of proportions relative to the unrestricted

model ($\hat{\Pi}_{MD,U}$, $\hat{\Pi}_{NO,U}$ and $\hat{\Pi}_{UD,U}$), and hence Bayes factors relative to the

unrestricted model ($\widehat{BF}_{MD,U} = \hat{\Pi}_{MD,U} / \pi_{MD,U}$, $\widehat{BF}_{NO,U} = \hat{\Pi}_{NO,U} / \pi_{NO,U}$ and

$\widehat{BF}_{UD,U} = \hat{\Pi}_{UD,U} / \pi_{UD,U}$).

Klugkist, Laudy and Hoijtink (2005) describe Bayes factors relative to the

unrestricted model as measuring model fit in a way that takes into account model

complexity. Intuitively, the adjustment for complexity occurs because, as a model's

prior proportion (π) increases (indicating that the model is a priori more complex

because it is able to fit a greater variety of data by chance), the Bayes factor

decreases. Similarly, the Bayes factor increases as the model's posterior proportion

increases, indicating that it provides a better fit to the data. When the Bayes factor for

a model relative to the unrestricted model is greater than one it provides evidence that the model is preferred over the unrestricted model with associated posterior model probability $p_{m,\{m,U\}} = BF_{m,U}/(1 + BF_{m,U})$.

Model-selection strategies. Posterior model probability estimates for each diagnostic model ($m = 1...4$) within the set of four diagnostic models ($\hat{p}_{m,\{U\}}$) are given by $\widehat{BF}_{m,U}/\left(\widehat{BF}_{NT,U} + \widehat{BF}_{MD,U} + \widehat{BF}_{NO,U} + \widehat{BF}_{UD,U}\right)$, where {U}={NT, MD, NO, UD}. These posterior model probabilities quantify the relative evidence for and against each diagnostic model when no order is assumed more likely before observing the data and one of these models is assumed to be the true model. We will refer to model selection based on these probabilities as using a simultaneous exhaustive strategy.

The expensive computations (sampling and counting orders) implementing the simultaneous exhaustive strategy can be re-used to quickly explore many other model-selection strategies. For example, if the trace model is assumed to be certainly true, posterior model probabilities can be calculated relative to the remaining three diagnostic models, $\hat{p}_{m,\{T\}} = \widehat{BF}_{m,U}/\left(\widehat{BF}_{MD,U} + \widehat{BF}_{NO,U} + \widehat{BF}_{UD,U}\right)$, where {T}={MD, NO, UD}. We will describe model selection based on these probabilities as following a simultaneous trace-true strategy. Simultaneous trace-true probabilities can also be calculated from only posterior trace-model MCMC samples,

$$\hat{p}_{m,\{T\}} = \widehat{BF}_{m,T}/\left(\widehat{BF}_{MD,T} + \widehat{BF}_{NO,T} + \widehat{BF}_{UD,T}\right), \text{ where } \widehat{BF}_{m,T} = \hat{\Pi}_{m,T}/\pi_{m,T}.$$

Selection could also be applied sequentially, based on posterior probabilities for pairs of models at each step. For example, first trace and non-trace models are compared, then the multi-dimensional and monotonic models and finally uni-dimensional and no-overlap models. The model selection sequence terminates if the

non-trace model probability ($p_{NT,\{NT,T\}}$) is high, supporting a need for design refinement. Termination also occurs if the multi-dimensional model probability ($p_{MD,\{MD,M\}}$) is high, supporting more than one latent variable. If the monotonic model is selected the final comparison determines whether there is support for a single latent variable (higher $p_{UD,\{UD,NO\}}$) or the need to calibrate the trace factor to improve overlap (higher $p_{NO,\{UD,NO\}}$).

Different selection strategies can answer similar questions, but when they are based on different assumptions they can give different answers. For example, the simultaneous trace-true strategy typically produces more decisive selection among the multi-dimensional, uni-dimensional and no-overlap models than the simultaneous exhaustive strategy, reflecting the certainty with which it assumes that the non-trace model is not the true model. Similarly, posterior probabilities calculated in the sequential strategy reflect the added assumptions for each step after the first (i.e., the non-trace model is false, then both non-trace and multi-dimensional models are false).

Although the three methods we have outlined often produce consistent results we generally prefer the simultaneous exhaustive strategy, both because of its weaker prior assumptions and the usefulness of each of its potential outcomes. However, where prior experimental evidence in favor of the trace model (i.e., evidence for a monotonic effect of the trace factor) is deemed sufficiently strong, the simultaneous trace-true strategy has the advantage of typically being more decisive, although the difference may be small. The same is often the case for later steps in the sequential strategy, but the conditional nature of these probabilities can also sometimes be confusing.

Group Bayes factors. Bayes factors for individual participants can be combined by multiplication to provide evidence about the best model for a group of participants. For example, individual participant Bayes factors for the trace and non-

trace models ($BF_{T, U, i}$ and $BF_{NT, U, i}$ for $i = 1 \ldots n$ participants) yield corresponding

group Bayes factors $GBF_{T,U} = \prod_{i=1}^{n} BF_{T,U,i}$ and $GBF_{NT,U} = \prod_{i=1}^{n} BF_{NT,U,i}$. These group

Bayes factors can be combined to yield a posterior probability comparing a model in

which the trace model is true for all participants to a model in which the non-trace

model is true for all participants: $gp_{T,\{T,NT\}} = GBF_T / (GBF_T + GBF_{NT})$. Similar

calculations produce group Bayes factors and posterior model probabilities that can be

used in any of the model selection strategies just outlined.

It is important to note that the group Bayes factors do not provide evidence

about the population of participants, as they assume that participants are unrelated,

rather than being samples from a common population distribution. Group Bayes

factors may not be appropriate when the group of participants is heterogeneous with

respect to the model comparison (i.e., when the true model differs among sub-groups

of participants), as they make the assumption that *all* participants are of one type or

another. However, they are preferable to analysing state-trace plots averaged over

participants, which not only share the limitations of group Bayes factors, but which

can also potentially be misleading due averaging distortion, as discussed previously.

Example Analysis

We illustrate the proposed analysis method by applying it to our inverted-test

and upright-test experiments. We use letters to identify individual participant results;

upper case for the inverted-test experiment and lower case for the upright-test

experiment. Figure 7 graphically represents, for one participant in each experiment,

the results of MC sampling from the unrestricted model posterior. The large points in

the plots are posterior modes (i.e., points of highest posterior density) for the HR –

FAR accuracy measure. Unrestricted model posterior MC samples can also be used to

calculate other measures of central tendency, such as the mean or median (all three

measures are very similar in our experiments). The plots also show 50% credible regions (the Bayesian analogue of NHST confidence regions), indicating the degree of uncertainty about posterior parameter estimates (see Figure 7 caption for details).[2]

Figure 7 shows that posterior MCMC samples can be used to visualize order-restricted models. For example, Figure 7a shows posterior modes for the trace model MCMC samples (small symbols), with separate lines joining points from different dimension-factor levels. Although the trace model plot for participant G is consistent with multi-dimensionality the large credible regions suggest this may be attributable to estimation error.

Posterior MCMC samples can also be used to choose the monotonic order that best describes a state-trace plot and the level uncertainty about this choice. For example, Figure 7b shows the modes of the best monotonic model for participant j, where *best* is defined as the most frequently occurring monotonic model in posterior MCMC samples. In this case the best monotonic model is non-overlapping, as all inverted-condition modes are less than all upright-condition modes on both axes. Visual inspection of the unrestricted model modes in Figure 7a suggests that the monotonic model reversing the middle two points (i.e., triangle 1 before circle 4) is almost as good as the non-overlapping model. Reflecting this uncertainty, the Bayes factor for the non-overlapping model vs. the model with the middle points reversed, as estimated by the ratio of posterior counts for each model (as both models have equal

---

[2]Note that the posterior means of the binomial probability parameters with a uniform prior can be directly estimated as they have a simple analytic form, $\hat{\theta} = (n+1)/(N+2)$, where, for example, n is the number hits and N the number of trials with studied items. However, this is not the case for accuracy measures, even in the simplest case, the HR – FAR measure. In contrast, estimates of posterior central tendency statistics for any accuracy measure are easily estimated by applying the accuracy measure to posterior MC samples. Similarly, credible regions are easily obtained from the appropriate quantile of the posterior accuracy distributions.

prior probability), is only 1.05, with a corresponding posterior model probability of 0.53 for the no-overlap model.

Figure 8 plots estimated simultaneous exhaustive posterior model probabilities for individual participants. Figure 8a plots inverted-test experiment estimates and Figure 8b upright-test experiment estimates. Each of the four panels within a plot displays results for one of the four diagnostic models. In each panel participants are sorted in order of their probability estimates to make it easy to identify extreme cases. The titles for each panel give group posterior model probabilities for the corresponding diagnostic model.

In order to aid interpretation of posterior model probabilities horizontal dotted lines in each panel indicate ranges which Raftery (1995) characterized as providing: a) equivocal evidence (against the model when $.25 \leq p < .5$ and for the model when $.5 < p \leq .75$), b) positive evidence (against the model when $.05 \leq p < .25$ or for the model when $.75 < p \leq .95$) or c) strong evidence (against the model when $p < .05$ or for the model when $p > .95$). For example, Figure 8a shows the evidence for both the uni-dimensional and multi-dimensional models is equivocal for participant G, confirming the impression given by the large credible regions in Figure 7a.

Figure 8a shows positive or better individual-participant evidence for the trace model in the inverted-test experiment. The exception is participant O, who has positive evidence for the non-trace model. Because there is generally little evidence for the non-trace model, simultaneous trace-true posterior model probabilities for the other three diagnostic models are little changed from the values displayed in Figure 8a, except for O, where $\hat{p}_{NO,\{T\}} = .8$, $\hat{p}_{MD,\{T\}} = .106$ and $\hat{p}_{UD,\{T\}} = .094$.

The failure of the trace model for participant O is likely due to them having the lowest overall accuracy of any participant (an average HR – FAR = 0.13 averaged

over conditions), which means the trace factor necessarily has only a small effect.

This conclusion is consistent with small Bayes factors for participant O

($\widehat{BF}_{NT,\{U\}} = 1.00$, $\widehat{BF}_{NO,\{U\}} = .121$, $\widehat{BF}_{UD,\{U\}} = .014$ and $\widehat{BF}_{MD,\{U\}} = .016$), indicating

that no order-restricted model is preferred over the unrestricted model. In contrast,

other participants have at least one order-restricted model with a Bayes factor of 8 or

greater, and in most cases much greater.

Taking into account their low accuracy, positive evidence for the non-trace

model, and positive evidence for a failure of overlap (and hence non-diagnostic

dimensionality results) even when the trace model is assumed true, censoring of

participant O is indicated. However, when this is done, there is little change in the

group posterior model probabilities for the inverted-test experiment ($\widehat{gp}_{UD,\{U\}} = .94$,

$\widehat{gp}_{MD,\{U\}} = .06$, and $\widehat{gp}_{NT,\{U\}} = \widehat{gp}_{NO,\{U\}} = 0$). Individual analyses also favour a multi-

dimensional model for all participants (except O), with participants A and D being

possible exceptions. In contrast to other participants, these participants have positive

evidence against the uni-dimensional model, and equivocal evidence favouring the

multi-dimensional model. With participants A and D, as well as participant O,

censored, $\widehat{gp}_{UD,\{U\}} = .996$, strongly supporting the homogenous multi-dimensional

model for this subset of participants.

Figure 8b shows positive or better evidence supporting the trace model for all

participants in the upright-test experiment. In contrast to the average results in Figure

4a, a failure of data-trace overlap is indicated for participant j (see Figure 7b), d and

possibly f. However, censoring these participants has little effect on the group

posterior model probabilities ($\widehat{gp}_{MD,\{U\}} = .999$, $\widehat{gp}_{UD,\{U\}} = .001$, and $\widehat{gp}_{NT,\{U\}} = 0$,

$\widehat{gp}_{NO,\{U\}} = 0$). Overall, a strong conclusion in favor of a multi-dimensional model for all participants is supported for the upright-test experiment.

In order to further examine the effect of censoring, Figure 9 displays results averaged over uncensored participants. Modes and confidence regions are based on participant averages obtained by bootstrap methods (Efron & Tibshirani, 1998) applied in a way that reflects the uncertainty about each participant's parameters quantified by their posterior distributions (see Figure 9 caption for details). The results shown in Figure 9 are similar to results obtained when this method is applied to all participants (not shown). Both of these results are similar to Figure 4, which is to be expected, as Bayesian estimates based on a uniform prior are asymptotically equivalent to the maximum-likelihood estimates (i.e., $\hat{\theta} = n/N$) used to construct Figure 4. That is, given the number of test trials in our experiments the estimates are almost numerically equivalent.

Summary and Recommendations for Statistical Analysis

We recommend the use of Bayes-factor based analyses of individual participants for appropriate experiments (binary choice data and a within-subjects design). Generally the simultaneous exhaustive strategy is recommended, although when there is strong prior evidence for the trace model the trace-true strategy can be both more appropriate and diagnostic. Although we have emphasized posterior model probabilities, users uncomfortable with the underlying assumption that one model is the true model may prefer to rely on Bayes factors, which quantify the relative evidence for two models without making this assumption. In any case, we recommend examination of Bayes factors relative to the unrestricted model as a way of assessing absolute goodness-of-fit taking into account model complexity.

We recommend plots of individual Bayes factors and posterior model probabilities (e.g., Figure 8) be examined to check for outliers of sub-groups that differ in terms of evidence for different models. We recommend group Bayes factors and/or posterior model probabilities and plots such as Figure 9 for presenting summary results. However, group statistics should be used with caution[3], with the reasonableness of aggregation checked and censoring applied, or separate group statistics based on sub-groups reported, as indicated by plots like Figure 8. With accuracy data it is also important to check for performance at floor or ceiling. Although state-trace analysis can address potential *confounding* of dimensionality estimation attending these effects, such data contain little useful information, making model selection equivocal (e.g., participant O in the example analysis).

All of the types of statistics and figures presented in the example analysis, and extensions, can be conveniently obtained using software freely available at http://www.newcl.org (see Prince et al., submitted). To explore extensions, users may also, for example, access raw posterior MC and MCMC sample-order counts to compute the Bayes factors required to explore alternative model-selection strategies, plot posterior means and medians as well as modes, and construct average state-trace plots by bootstrap methods that treat participants as samples from a population.

General Discussion

A key methodology in many areas of psychology involves studying the interplay between two factors in order to determine whether their effect can be

---

[3]A simulation study reported by Hawkins, Prince, Brown and Heathcote (2010) found that group Bayes factors can fail to identify the dimensionality correctly when the number of trials in each condition for each participant is small. Hence, we recommend that experiments be designed to maximize the number of trials in each condition (see the last major section for recommendations about how to achieve this outcome). Note, however, that the simulation study assumed a worst-case scenario, where all participants were samples from an identical population model with no between-subjects variation, whereas group Bayes factors assume participants are unrelated. In ongoing work we are testing the operating characteristics of group Bayes factors in more realistic situations where participants are samples from a population distribution with between-subjects variation.

explained by a single latent variable. The traditional approach to this issue, rejecting a one-dimensional model when a significant interaction is found in an ANOVA, requires strong assumptions that are difficult to check (Dunn & Kirsner, 1988) and the fortuitous finding of a particular data pattern such as a crossover interaction (Loftus, 1978). State-trace analysis (Bamber, 1979) is an alternative approach that places control back in the hands of the experimenter. Its assumptions are minimal and a diagnostic pattern of data can usually be guaranteed through systematic manipulation of a third (trace) factor. Although the experimental methodology for state-trace analysis differs in its requirements from familiar factorial designs, once the differences are recognized it is straightforward to develop and refine a state-trace experiment. The present paper provides guidelines for this process, which apply to almost all of the types of state-trace experiments in the existing literature.

Although we advocate the wider use of state-trace analysis it is important to recognize its limitations. Although state-trace analysis can provide strong evidence for more than one latent variable that does not necessarily imply the involvement of more than one cognitive module, representation, process, or brain region if these entities have a multivariate nature. For example, parametric single-process models of recognition memory, which assume recognition decisions are based on a single memory-strength dimension, are usually multivariate as memory strength is characterized by both mean and variance parameters. If experimental manipulations have independent effects on both parameters state-trace plots will be non-monotonic because the two parameters of the model act as two latent variables and so, in Loftus et al.'s (2004) terminology, there are two dimensions.

In general, we see state-trace analysis and parametric model fitting as complimentary approaches. State-trace analysis is a relatively assumption free method

of identifying the number of model parameters that should be freely estimated in order to fit data. For example, Heathcote, Bora and Freeman (2010) found that quite different (single and dual process) models of recognition memory required the same number of free parameters as indicated by Heathcote et al.'s (2009) state-trace analysis of the same paradigm. As another example, Brainerd et al. (2009, Figure 12) applied state-trace analysis to direct access and reconstruction parameter estimates from a fitted model of recall in order to check whether they measure distinct processes.

We believe that inference about dimensionality requires an alternative to traditional null-hypothesis statistical testing, which can provide evidence against but not for a one-dimensional (null) model. Further, we believe it is important to take into account the greater flexibility (i.e., the ability to fit data by chance) of higher-dimensional models. Bayes factors represent one method to address such differences in functional form complexity (Pitt & Myung, 2002) and the limitations of null-hypothesis based analyses. When combined with Klugkist, Kato and Hoijtink's (2005) method for approximating Bayes factors, the result is a statistical methodology compatible with the relatively assumption-free nature of state-trace analysis. We developed and implemented this approach for experiments with a binary dependent variable (e.g., accuracy), enabling users not only to assess dimensionality, but also to help refine experimental methodology. However, further work remains to extend this approach to the full range of existing state-trace applications.

Future Directions

One set of extensions relates to experiments with ratings or response time as the dependent variable. These extensions require the assumption of different data distributions, but are otherwise conceptually straightforward. Dunn (2008) describes

an application of dependent-variable state-trace analysis to three-level rating data, which works with the binary-response analysis developed here, where the state factor is made up of cumulative probabilities for two of the three ratings levels. A second set of extensions is to designs with a dimension factor that has more than two levels or multiple dimension factors. Once again this extension is conceptually straightforward, although computational cost increases rapidly with the number of experimental conditions.

Kleigel, Maayr and Krampe (1994) state that: 'state-trace analysis can be thought of as a factor-analysis for experimental research; it yields the minimum number of mechanisms required for description of ordinal interactions' (p.153). For example, Heathcote et al. (2009) used state-trace analysis in way analogous to identifying manipulations that load on different factors. They showed that the effect of one manipulation was non-monotonic, indicating the need for at least two latent variables, but that the effects of two other manipulations were monotonic within each level of the non-monotonic manipulation, suggesting the effect of the latter two manipulations load on the same latent variable.

However, state-trace analysis with only two state levels is not fully analogous to factor analysis as it only determines if one or more than one latent variable is required. In order to determine the number of dimensions above one, extra axes must be added to the state-trace plot. That is, to identify the number of dimensions up to D, a state factor with $D + 1$ levels is required (see Dunn & James, 2003, for a general formulation of this problem). The potential importance of pursuing this extension is indicated by typical results from structural equation modeling, where reliable estimation of latent mediators usually requires converging evidence from several indicator variables. It seems likely that state-trace analysis will also more reliably

identify dimensionality using a state factor with more than two levels (where levels are analogous to indicator variables).

A key area for future extension concerns addressing differences between participants. We applied our model-selection procedures to individual participant data because of concerns about averaging. Multilevel models provide an alternative method of group analysis that does not require averaging, and which can also be applied to designs with between-subjects factors. For example, Verhaeghen and Cerella (2002, see also Verhaeghen et al., 2003) used maximum-likelihood multilevel modeling, assuming a linear relationship between levels of a dependent-variable state factor, to a meta-analysis where the trace factor represented different studies, and Van Den Broeck and Geudens (submitted) when the trace factor represented different participants. Dimensionality was diagnosed by selection between a model with different linear relationships for levels of the dimension factor (i.e., a multi-dimensional model) and a model with a common relationship (i.e., a uni-dimensional model).

Further discussion and development of this approach is beyond the scope of the present article, but we note that Klugkist, Kato and Hoijtink's (2005) Bayes factor method can be implemented in a multilevel framework (Myung et al., 2008), avoiding the need for assumptions about a particular parametric form for the relationship between levels of the state-factor. Multilevel models also have an advantage for purely within-subjects designs in providing group-level summaries based on the plausible assumption that participants come from a common population. The individual level methods proposed here provide a basis for the development and validation of such multilevel models (Borsboom, Mellenbergh & van Heerden, 2003).

References

Andrieu, C., de Freites, N., Doucet, A. & Jordan, M.I. (2003). An introduction to MCMC for machine learning. *Machine Learning*, *50*, 5-43.

Bamber, D. (1979). State-trace analysis: A method of testing simple theories of causation. *Journal of Mathematical Psychology, 19*, 137-181.

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., & Brunk, H. D. (1972). *Statistical inference under order restrictions*. Wiley: London.

Bogartz, R. S. (1976). On the meaning of statistical interactions. *Journal of Experimental Child Psychology*, *22*, 178-183.

Borsboom, D., Mellenbergh, G. J. & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*, 203-219.

Brainerd, C. J., Reyna, V. F., & Howe, M. L. (2009). Trichotomous processes in early memory development, aging, and neurocognitive impairment: A unified theory. *Psychological Review*, *116*, 783-832.

Brainerd, C. J., Wright, R., Reyna, V. F., & Payne, D. G. (2002). Dual retrieval processes in free and associative recall. *Journal of Memory and Language*, *46*, 120–152.

Busemeyer, J. R. & Jones, L. E. (1983). Analysis of multiplicative combination rules when the causal variables are measured with error. *Psychological Bulletin, 93*, 549-562.

Busey, T. A., Tunnicliff, J., Loftus, G. R., & Loftus, E. F. (2000). Accounts of the confidence-accuracy relation in recognition memory . *Psychonomic Bulletin & Review, 7*, 26-48.

De Brauwer, J., Verguts, T., & Fias, W. (2006). The representation of multiplication

facts: Developmental changes in the problem size, five, and tie effects.

Journal of Experimental Child Psychology, 94, 43–56.

Dunn, J. C. (2004). Remember–know: A matter of confidence. *Psychological Review,*

111, 524–542.

Dunn, J. C. (2008). The dimensionality of the remember–know task: A state-trace

analysis. *Psychological Review, 115*, 426-446.

Dunn, J. C. & James, R. N. (2003). Signed difference analysis: Theory and

application. *Journal of Mathematical Psychology, 47,* 389-416.

Dunn, J. C. & Kirsner, K. (1988). Discovering functionally independent mental

processes: The principle of reversed association. *Psychological Review, 95,*

91-101.

Efron, B. & Tibshirani, R. J. (1998). *An Introduction to the Bootstrap*. CRC Press,

New York.

Freeman, E., Dennis, S., & Dunn, J. (2010). An examination of the ERP correlates of

recognition memory using state-trace analysis. In S. Ohlsson & R. Catrambone

(Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science*

Society. Austin, TX: Cognitive Science Society.

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review*,

116, 439–453.

Gelfand, A. E., Smith, A. F. M. & Lee, R.-M. (1992). Bayesian analysis of

constrained parameter and truncated data problems. *Journal of the American*

Statistical Association, *87*, 523-532.

Haist, F., Shimamura, A.P. & Squire, L.R. (1992). On the relationship between recall and recognition memory. *Journal of Experimental Psychology: Learning, Memory & Cognition, 18,* 691-702.

Hawkins, G., Prince, M., Brown, S.D. & Heathcote, A. (2010)**.** Designing state-trace experiments to assess the number of latent psychological variables underlying binary choices. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Heathcote, A., Bora, B. & Freeman, E. (2010). Modeling choice-similarity effects in episodic recognition. *Journal of Memory and Language*, *62*, 183-203.

Heathcote, A., Freeman, E., Etherington, J., Tonkin, J. & Bora, B. (2009). A dissociation between similarity effects in episodic face recognition. *Psychonomic Bulletin & Review*, *16*, 824-831.

Henson, R. (2006). Forward inference using functional neuroimaging: Dissociations versus associations. *Trends in Cognitive Sciences*, *10*, 64-69.

Hintzman, D. L. (2004). Judgment of frequency versus recognition confidence: Repetition and recursive reminding. *Memory & Cognition, 32,* 336-350.

Jang, Y. & Nelson, T. O. (2005). How many dimensions underlie judgments of learning and recall? Evidence from state-trace methodology. *Journal of Experimental Psychology: General, 134*, 308-326.

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

Kliegl, R., Maayr, U., & Krampe, R. T. (1994). Time-accuracy functions for determining process and person differences: An application to cognitive aging. *Cognitive Psychology, 26,* 134-164.

Klugkist, I., Kato, B. & Hoijtink, H. (2005). Bayesian model selection using

encompassing priors. *Statistica Neerlandica*, *59*,57–69.

Klugkist, I., Laudy, O. & Hoijtink, H. (2005). Inequality constrained analysis of

variance: A Bayesian approach. *Psychological Methods*, *10*, 477-493.

Lewandowsky, S., Geiger, S. M., Morrell, D. B., & Oberauer, K. (2010). Turning

simple span intocomplex span: Time for decay or interference from

distractors? *Journal of Experimental Psychology: Learning, Memory, and

Cognition, 36,* 958-978.

Liu, C. C. & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model

generalizability. *Journal of Mathematical Psychology*, *52*, 362-375.

Loftus, G. R. (1978). On interpretation of interactions. *Memory & Cognition, 6,* 312-

319.

Loftus, G. R. (2002). Analysis, interpretation, and visual presentation of experimental

data. In H. Pashler (Ed.), *Stevens' handbook of experimental psychology* (Vol.

4, pp. 339-390). New York: John Wiley and Sons.

Loftus, G. R. & Harley, E. M. (2005). Why is it easier to identify someone close than

far away? *Psychonomic Bulletin & Review, 12*, 43-65.

Loftus, G. R. & Irwin, D. E. (1998). On the relations among different measures of

visible and informational persistence. *Cognitive Psychology, 35,* 135–199.

Loftus, G. R. & Masson, M. E. J. (1994). Using confidence intervals in within-subject

designs. *Psychonomic Bulletin & Review*, *1*, 476-490.

Loftus, G. R., Oberg, M. A. & Dillon, A. M. (2004). Linear theory, dimensional

theory, and the face-inversion effect. *Psychological Review, 111*, 835-865.

Macmillan, T. A., & Creelman, C. D. (2005). *Detection theory: A users guide*

(2ndEdn.). New York: Cambridge University Press.

Maurer, D., Le Grand, R. & Mondloch, C. J. (2002). The many faces of configural processing. *Trends in Cognitive Sciences*, *6*, 255-260.

McLachlan, G. & Peel. D. (2000). *Finite mixture models*. New York: Wiley.

Myung, J., Karabatsos, G. & Iverson, G. J. (2008). A statisticians view on Bayesian evaluation of informative hypotheses. In H. Hoijtink, I. Klugkist, & P. Boelen (Eds.), *Bayesian evaluation of informative hypotheses* (pp. 309-328). Berlin: Springer.

Newell, B. R., & Dunn, J. C. (2008). Dimensions in data: Testing psychological models using state-trace analysis. *Trends in Cognitive Sciences, 12,* 285-290.

Newell, B. R., Dunn, J. C., & Kalish, M. (2010). The dimensionality of perceptual category learning: A state-trace analysis. *Memory & Cognition, 38,* 563-581.

Newell, B. R., Dunn, J. C., & Kalish, M. (2011). Systems of category learning: Fact or fantasy? In B.H. Ross (Ed.) *The Psychology of Learning & Motivation, 54*, 167-215.

Nilsson, L. & Gardiner, J. M. (1993). Identifying exceptions in a database of recognition failure studies from 1973 to 1992. *Memory & Cognition, 21*, 397-410.

Oberauer, K., & Lewandowsky, S. (2008). Forgetting in immediate serial recall: Decay, temporal distinctiveness, or interference? *Psychological Review, 115, 544-576*.

Pitt, M.A. & Myung, I. J. (2002).When a good fit can be bad. *Trends in Cognitive Sciences*, *6*, 421-425.

Poldrack, R.A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Science*, *10*, 59-63.

Prince, M., Hawkins, G., Love, J. & Heathcote, A. (submitted). StateTrace: An R package for Bayesian ordinal analysis of binomial state-trace data.

Prince, M. & Heathcote, A. (2009). State-trace analysis of the face-inversion effect. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

R Development Core Team (2007). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Raftery, A.E. (1995). Bayesian model selection in social research. *Sociological Methodology, 25,* 111-163.

Rakover, S. S. (2002). Featural vs. configurational information in faces: A conceptual and empirical analysis. *British Journal of Psychology*, *93*, 1-30.

Rakover, S. S. & Teucher, B. (1997). Facial inversion effects: Parts and whole relationship. *Perception & Psychophysics*, *59*, 752-761.

Rock, I. (1974). The perception of disoriented figures, *Scientific American*, *230*, 78-85.

Rouder, J., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, *12*, 573–604.

Saville, D.J. (2003). Basic statistics and the inconsistency of multiple comparison procedures. *Canadian Journal of Psychology*, *57*, 167-175.

Shallice, T. (1988). *From Neuropsychology to Mental Structure*, Cambridge University Press.

Sheather, S. J. & Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B, 53*, 683-690.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology*, *26*, 1–12.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, *80*, 352-373.

Valentine, T. (1988). Upside-down faces: A review of the effect of inversion upon face recognition. *British Journal of Psychology, 79,* 471–491.

Van den Broeck, W. & Geudens, A. (submitted). Old and new ways to study characteristics of reading disability: The case of the nonword reading deficit.

Verhaeghen, P. & Cerella, J. (2002). Aging, executive control, and attention: A review of meta-analyses. *Neuroscience and Biobehavioral Reviews, 26,* 849–857.

Verhaeghen, P. & De Meersman, L. (1998). Aging and the negative priming effect: A meta-analysis. *Psychology and Aging*, *13*, 435-444.

Verhaeghen, P., Steitz, D. W., Sliwinski, M. J. & Cerella, J. (2003). Aging and dual-task performance: A meta-analysis. *Psychology and Aging*, *18*, 443-460.

Wagenmakers, E. -J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*, 779-804.

Wand, M. P. & Jones, M. C. (1995). *Kernel Smoothing.* Chapman and Hall: London.

Wixted, J.T. (1990). Analyzing the empirical course of forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 927-935.

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review*, *114*, 152-176.

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental*

Psychology, 81, 141-145.

Appendix

Prior Proportions

We assume both state and dimension factors have two levels and the trace factor has t levels. The no-baseline design is denoted as B0 and a design with one baseline for each state, as B2. Analytic prior proportions (π) are obtained based on an encompassing prior that assumes identical and independent uniform distributions for all binomial probability parameters (θ), making every ordering of these parameters equally likely. In the following, θ for a baseline condition is denoted θ_0 and for the remaining conditions θ_i, $i = 1,2, \ldots t$. For brevity we refer to draws from the prior and posterior distributions of an order-constrained model, whether obtained by MC or MCMC methods, as prior and posterior samples.

Proportions within the unrestricted model. The trace model assumes that the same order holds within a data trace for both states. Given there are t levels within a trace and $t!$ possible orders:

$$\pi_T(B0) = [\, t!\,]^{-4} \qquad\qquad (1a)$$

For a B2 design we include the extra constraint that $\theta_0 < \theta_i$ within each state $\times$ dimension condition. This constraint has probability $1/(\, 2t+1\,)$ as it applies simultaneously to all $2t$ non-baseline θ_i. In combination with the probability of obtaining the required order of the remaining θ parameters within each state $\times$ dimension condition:

$$\pi_T(B2)=[(t!)^2(2t+1)]^{-2} \qquad\qquad (1b)$$

Proportions within the trace model. To determine the prior probability of the monotonic model within the trace model, we first require the number of possible orderings for trace models. In general for k traces of equal length there are $(k\times t)!\, /$

$(t! \times (k \times t - t)!)$ (i.e., $^{kt}C_t$) ways of choosing a trace model order, so for $k = 2$ the number of orders is:

$$O_T(t) = (2t)! \, (t!)^{-2} \qquad (2)$$

Given there are $2t\theta$ values for each state in the B0 design, the probability of any one trace model order is $1/(2t)!$, and so the proportion of prior samples with a monotonic ordering is the product of this value for each state with the number of trace models: $O_T(t) \times [(2t)!]^{-2}$.

Similarly, given $2t + 1$ values of θ for each state (including θ_0), for the B2 design the proportion of prior samples with a monotonic ordering is: $O_T(t) \times [(2t+1)!]^{-2}$. Note that in this case, although the order is on the $d = f(\theta_I) - f(\theta_0)$ differences measuring accuracy (where f is a monotonic function), we need only consider the ordering on the θ_i as this necessarily entails the same ordering on d given the common baseline θ_0.

The proportion of prior MCMC samples from the trace model that have a monotonic order is given by dividing the proportion of unrestricted prior samples with a monotonic ordering by the proportion that conform to the trace model: this proportion is the same for B0 and B2 designs:

$$\pi_M = (t!)^2/(2t)! = 1/O_T(t) \qquad (3)$$

The proportion of monotonic but non-overlapping data traces in prior MC samples from the trace model (i.e., samples not diagnostic of dimensionality), π_{NO}, is obtained from the fact that, for any number of trace levels, there are always exactly two non-overlapping orders and that each of the O_T monotonic model orders is equally likely:

$$\pi_{NO} = 2[\, O_T(t)\,]^{-2} \qquad (4)$$

This proportion is the same for B0 and B2 designs and can be used to get the proportion of monotonic MC samples that are overlapping, and thus diagnostic of a uni-dimensional system: $\pi_{UD} = \pi_M - \pi_{NO}$.

Posterior Proportions

The uniform distribution is a special case of the Beta distribution, Beta (a, b), where $a = b = 1$ for the uniform distribution. The Beta distribution is conjugate to the binomial distribution, meaning that the marginal posteriors also have a Beta distribution. For example, if s studied responses are observed from S recognition test trials, the marginal posterior has a Beta $(s + a, S - s + b)$ distribution. Samples drawn from the latter Beta distribution are used to estimate posterior model proportions. One possible proportion estimate for a given model with a count of n out of N posterior samples is $\hat{\Pi} = n/N$. However, in practice the estimate corresponding to the posterior mean proportion under a uniform prior, $\hat{\Pi} = (n+1)/(N+2)$, is preferable, as it has lower estimation variance (Rouder & Lu, 2005).

Computation of the trace model proportion can be done efficiently by taking advantage of the independence of accuracy measures between state and dimension factor conditions. Proportions are determined separately for independent parts of the posterior sample then combined by multiplication. Where accuracy is measured by proportion correct (e.g., for two-alternative forced choice testing), all four conditions are independent, so $\hat{\Pi}_T = \hat{\Pi}_{T,1,1} \times \hat{\Pi}_{T,1,2} \times \hat{\Pi}_{T,2,1} \times \hat{\Pi}_{T,2,2}$, where $\hat{\Pi}_{T,i,j}$ is the proportion of MC samples from the i'th state level and j'th dimension level following the trace order. Where accuracy is measured relative to a baseline (e.g., the HR-FAR measure) that is common to all conditions with a level of the state factor (e.g., one FAR for houses and one for faces) only the state levels are independent, as the trace model

contains both dimension levels relative to their shared baseline (i.e., $\theta_{i,0} < \theta_{i,2,1}$ and $\theta_{0,i}$ $< \theta_{i,2,1}$). Hence, $\hat{\Pi}_T = \hat{\Pi}_{T,1} \times \hat{\Pi}_{T,2}$, where $\hat{\Pi}_{T,i}$ is the estimate for the proportion of posterior MC samples in state i conforming to the trace order.

Posterior trace model proportions. Samples can be obtained directly from the posterior of the trace model using Gibbs sampling (Gelfand et al., 1992). The computational cost of Gibbs samples is around an order of magnitude greater than directly sampling from the Beta distribution. Also, samples from the initial iterations of the MCMC algorithm must be discarded because they do not provide a good approximation to the posterior distribution. Further inefficiency occurs because the sequence of MCMC samples is not independent, so an MCMC sample contributes less information than a direct sample. However, in practice we have found that this method remains computationally superior to estimating trace-model proportions by directly sampling from the unrestricted model posterior.

The Gibbs algorithm can be specified in terms of a set of marginal Beta cumulative distribution functions, F_k, and their inverses, F^{-1}_k. Denote the ith sample by a vector, $\mathbf{x_i}$, with n elements $(x_{i1}, x_{i2} \dots x_{in})$ – these elements represent posterior probability samples over each of the $2t$ conditions in the B0 design and $2(t+1)$ conditions in the B2 design.

For the B0 design let $k = 1 \dots t$ correspond to increasing trace conditions within a level of the dimension factor. Marginal posteriors are distributed as Beta ($n_k + a$, $N_k - n_k + a$), where n_k and N_k are the number of successes (e.g., correctly responding that the test item was studied) and the number of trials, respectively, for the k'th condition. Given an arbitrary initial sample $\mathbf{x}_1$ which respects $(x_{11} < x_{12} < \dots < x_{1n})$ a sequence $\mathbf{x_i}$, $i = 1 \dots S$, of samples from the trace model is obtained, after sufficient

burn-in samples are discarded, by randomly selecting and updating an element k of $\mathbf{x}_i$ using:

$$x_{(i+1)k} = F^{-1}_k(F_k(x_{i(k-1)}) + u(F_k(x_{i(k+1)}) - F_k(x_{i(k-1)})) \qquad (5)$$

In (5) u is a uniform random deviate on the unit interval and we define $F_k(x_{i0}) = 0$ and $F_k(x_{i(t+2)}) = 1$. The Gibbs sampler is run four times to produce four independent sets of samples, one for each state $\times$ dimension condition.

In the B2 design a single Gibbs sampler must be defined to obtain samples from both levels of the dimension factor at once, as both levels of the dimension factor share a baseline. For generality we define a Gibbs sampler for the case where there are possibly different numbers of trace levels for each state, $t1$ and $t2$ and let $t = t1 + t2 + 1$. Let $k = 1 \ldots t$, where $k = 1$ corresponds to the baseline condition, $k = 2 \ldots ($ $t1 + 1$) corresponds to increasing trace conditions in the first dimension level and $k = (t1 + 2) \ldots t$ corresponds to increasing trace conditions in the second dimension level.

Let x_{ik} represent a sample for condition k on iteration i. The samples must respect the joint order: $(x_{i1} < x_{i2} < \ldots < x_{i(n1+1)})$, $(x_{i1} < x_{i(n1+2)} < \ldots < x_{i(n1 + n2 + 1)})$. This is achieved by the following update rule:

$$x_{(i+1)k} = F^{-1}_k(p + u(P - p)) \qquad (6)$$

Where:

$p = 0$	if $k = 1$
$p = F_k(x_{i1})$	if $k = t1 + 2$
$p = F_k(x_{i(k-1)})$	otherwise

and:

$P = \min[F_k(x_{i2}), F_k(x_{i(n1+2)})]$	if $k = 1$
$P = 1$	if $k = t1 + 1$ or $k = t$
$P = F_k(x_{i(k + 1)})$	otherwise

Finally, we define a sequence of accuracy measures for each dimension as $d_{1ij} = d(x_{i(j+1)}, x_{i1})$, where $j = 1 \ldots n1$ and $d_{2ij} = d(x_{i(n1 + j + 2)}, x_{i1})$, where $j = 1 \ldots n2$ (e.g.,

$d(\text{x, y}) = x - y$). The sampler is run twice to produce two independent sets of accuracy samples, one for each state.

Having obtained a sufficiently large set of samples from the posterior of the trace model, Π_M can be estimated by counting the number which obey the constraints of the monotonic model, which can be expressed as rank (d_1) = rank (d_2), where rank ($\mathbf{y}$) is a function that returns the ranks of the elements of the vector $\mathbf{d}$ and the equality is in the sense of corresponding vector elements being equal. The issue of ties can be ignored as $\mathbf{d}$ is real valued, so ties occur with probability zero. When ties do occur to the limits of the machine's precision, they can be broken randomly with little influence on the outcome.

Adjacent samples generated by (5) and (6) are highly correlated, as they share all but one value. Where independent posterior samples are required they can be obtained by keeping only one sample in every T when T is sufficiently large. We set T equal to the length of a sample (i.e., on average every element had been updated between samples that were kept) and found this produced sufficiently independent samples for our purposes (in particular, so that we could estimate numerical accuracy as described in the next section). Given these specifications a burn-in period discarding only the first 100 samples was also found to be sufficient.

Numerical Accuracy

Monte-Carlo estimates become more precise as the sample size, S, on which they are based increases. We choose S such that the ($1 - \alpha$) credible interval for an estimate of a posterior proportion, Π, was less than δ, i.e., $F^{-1}(1 - \alpha / 2, a, b) - F^{-1}(\alpha / 2, a, b) < \delta$, where $F^{-1}(x, a, b)$ is the inverse of the cumulative distribution function for the Beta distribution with $a = \Pi S + 1$ and $b = S - \Pi S + 1$. Calculation of this interval requires knowledge of Π. We bootstrapped this knowledge by taking a

sample of size S_0, estimating Π, calculating the credible interval. We then iterated this procedure until the credible interval criterion was fulfilled.

This process was carried out separately with posterior samples from the unrestricted and trace models. For the unrestricted model credible intervals were calculated for each of the independent proportion estimates used to calculate $\hat{\Pi}_T$. For the trace model credible intervals were calculated for estimates of Π_{MD}, Π_{NO} and Π_{UD}. In both cases sampling was terminated when all estimates fulfilled the criteria. Monte-Carlo error in the Π_T estimate affects calculation of Bayes factors for the multi-dimensional, no-overlap and uni-dimensional models relative to the unrestricted model (as the trace model proportions are multiplied by $\hat{\Pi}_T$). Hence, we used a stricter criterion for the unrestricted model ($\delta = .0005$) than the trace model ($\delta = .005$) sampling, with $\alpha = .05$ in both cases.

Author Notes

Figure Captions

Figure 1. (a) Hypothetical results from a traditional dissociation experiment on the differential face-inversion effect (inversion has a large effect on performance for faces, but only a small effect for houses). (b) Hypothetical results from a corresponding state-trace experiment. This experiment takes the four conditions from the dissociation experiment (represented by the upper two points) and four new conditions, formed by manipulating the overall task difficulty (the lower two points). The state-trace plot in (b) graphs performance for house stimuli vs. performance for face stimuli, with lines (data traces) joining points from the upright condition and points in the inverted conditions. The dotted line shows that a monotonic (always-increasing) line can join all of the data points with small amount of misfit.

Figure 2. Two hypothetical mappings (response functions) between a latent variable (on an arbitrary scale) and a response variable on a 0-1 scale (e.g., hit rate).

Figure 3. State-trace plots showing illustrative data patterns from a 2×2×2 design. Letters indicate data points. Thick lines are *data traces*, which join groups of data points that differ only on the trace factor. Thin dotted lines show the state-trace of the underlying system; panels in the upper row show a one-dimensional system and panels in the lower row show a multi-dimensional system.

Figure 4. State-trace plots of accuracy as measured by hit rate minus false alarm rate (HR – FAR) for each participant then averaged over participants, for the (a) upright-test (18 participants), (b) inverted-test (16 participants) and (c) matched-test (14 participants) experiments. Each participant's FAR was based on 128 test trials and each HR on 32 test trials. Data points are plotted with least significant ($p=.05$) difference bars: $\pm(SE\times1.96\times\sqrt{2})/2$, (Saville, 2003), where SE is a within-subject

standard error (Loftus & Masson, 1994). Spearman's rank based measure of association (ρ) for each condition is given above its panel. The lines are *data-traces*, which join groups of data points that differ only on the trace factor.

Figure 5. State-trace plots illustrating averaging distortions. Letters indicate data points and the lines are *data-traces*, which join groups of data points that differ only on the trace factor. The dimension factor has two levels, with data from one level (*a* and *b*) joined by a solid data trace and data from the other level (*A* and *B*) joined by a dashed data trace.

Figure 6. Illustrations of the nesting relationship between samples from (a) the unrestricted (U) and trace (T) models, and (b) the trace model and the monotonic (M) model, which is represented by an oval divided into areas corresponding to monotonic orders with overlapping (O) and non-overlapping (NO) data traces. The Π values indicate proportions of samples corresponding to the model indicated by the subscript.

Figure 7. Modes of the posterior binomial probability parameter estimates from the unrestricted model (large symbols) with 50% credible regions (ellipses) for (a) participant *G* in the inverted-test experiment and (b) participant *j* in the upright-test experiment. The numbers 1…4 indicate shorter to longer study durations. The lines in (a) join posterior modes of MC samples from the trace model for each dimension factor level. The line in (b) joins posterior modes of MC samples from the best (i.e., most frequently sampled) monotonic model. Modes and credible regions were obtained using linear binned two-dimensional kernel density estimates (see, e.g., Wand & Jones, 1995) with a bivariate normal kernel, with covariance matrix given by Sheather and Jones' (1991) direct plug-in algorithm.

Figure 8. Posterior model probability estimates for each participant (denoted by letters) in (a) the inverted-test and (b) the upright-test experiments. Each panel shows estimates for one of the four diagnostic models indicated by the title above the panel, which also provides the group posterior probability for each model. Within each panel participants are sorted in order of increasing probability estimates.

Figure 9. Modes of the posterior binomial probability parameter estimates from the unrestricted model averaged over participants (large symbols) with 50% credible regions (ellipses) for (a) the inverted-test experiment with participants *A*, *D* and *O* censored and (b) the upright-test experiment with participants *d*, *f* and *j* censored. Averages over participants were obtained by bootstrapping based on 10,000 posterior estimates from the unrestricted model for each participant. For each participant one of the 10,000 posterior samples was randomly selected and the selected samples averaged over participants. This procedure was repeated 10,000 times, and the methods described in Figure 8's caption applied to obtain posterior modes and credible regions.
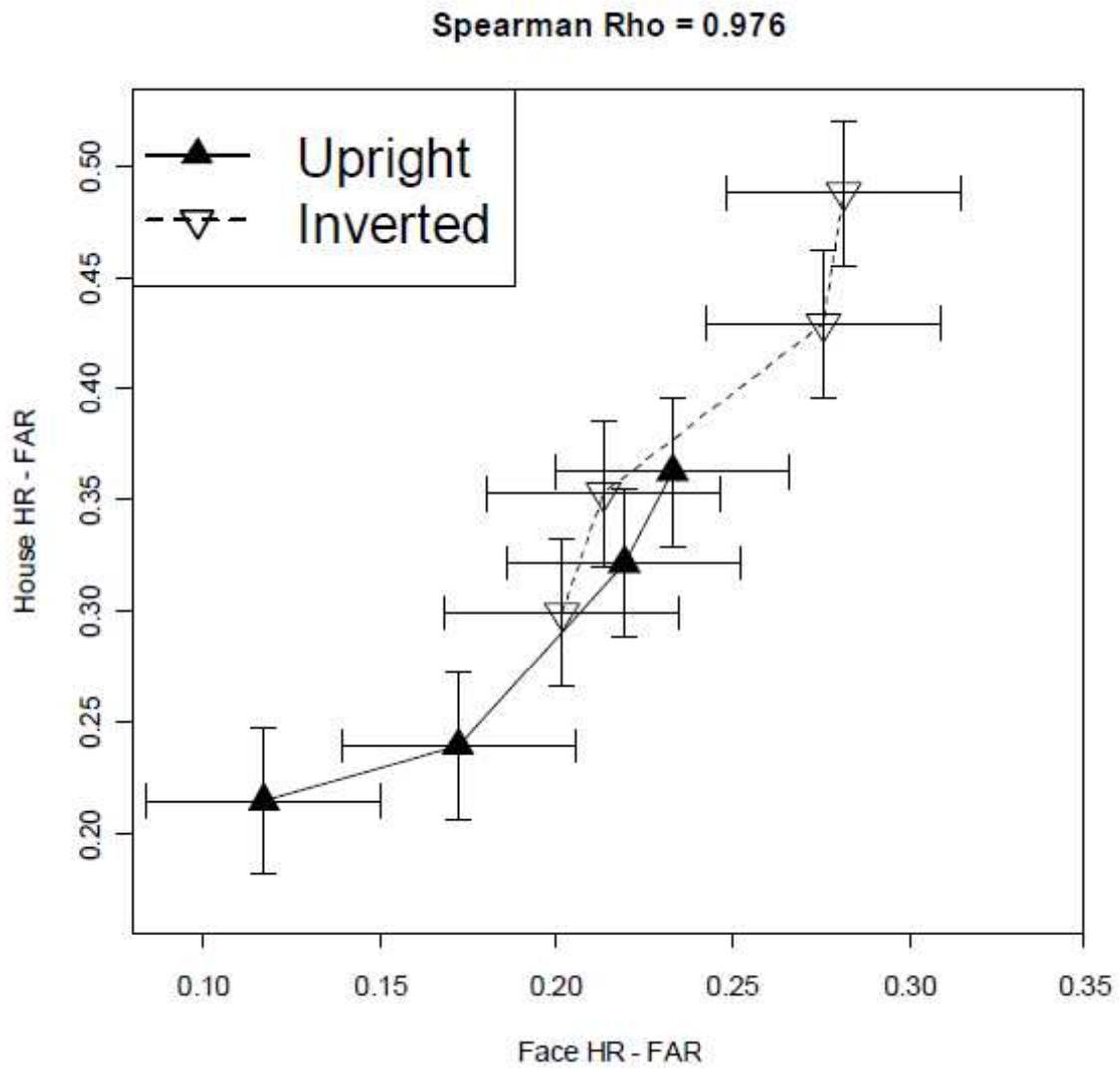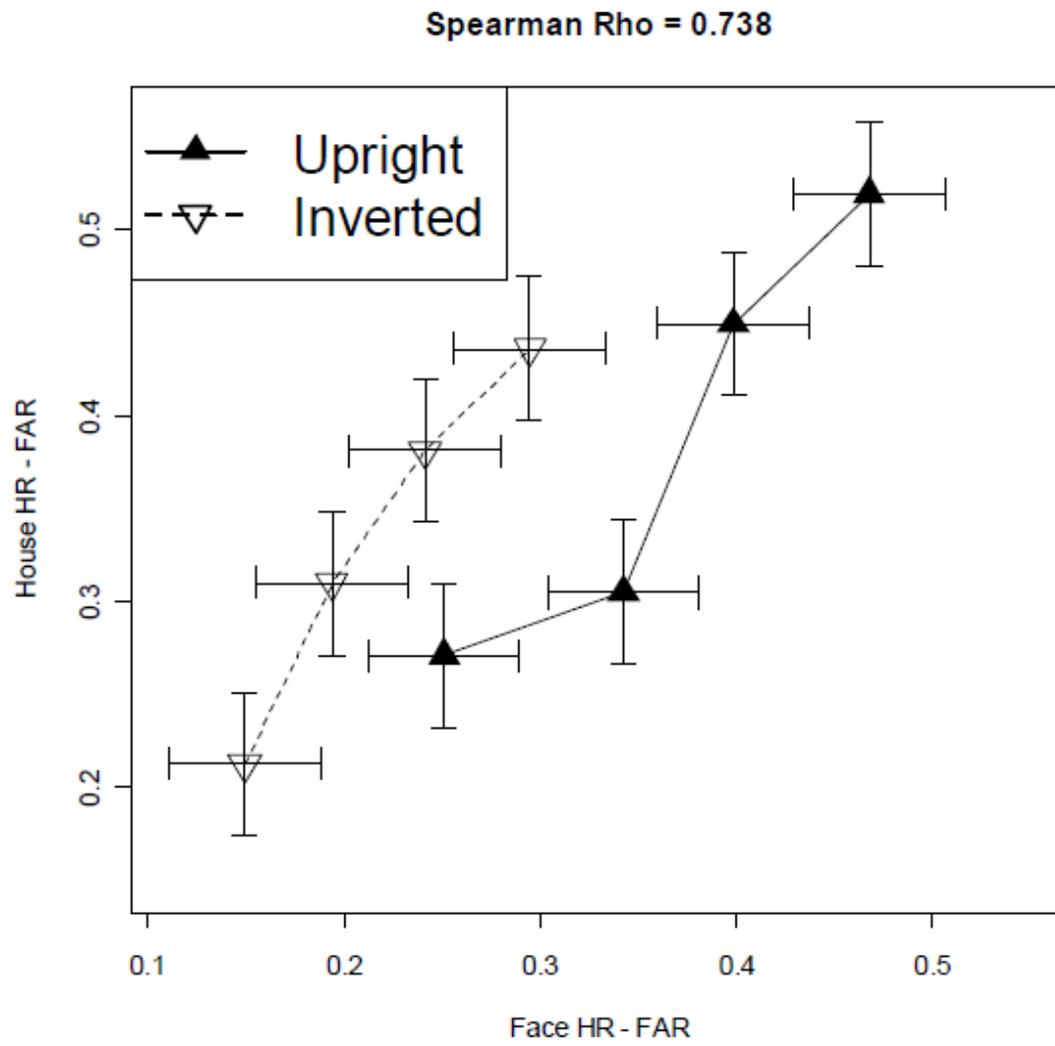
(a)

(b)

Figure 1

(a)
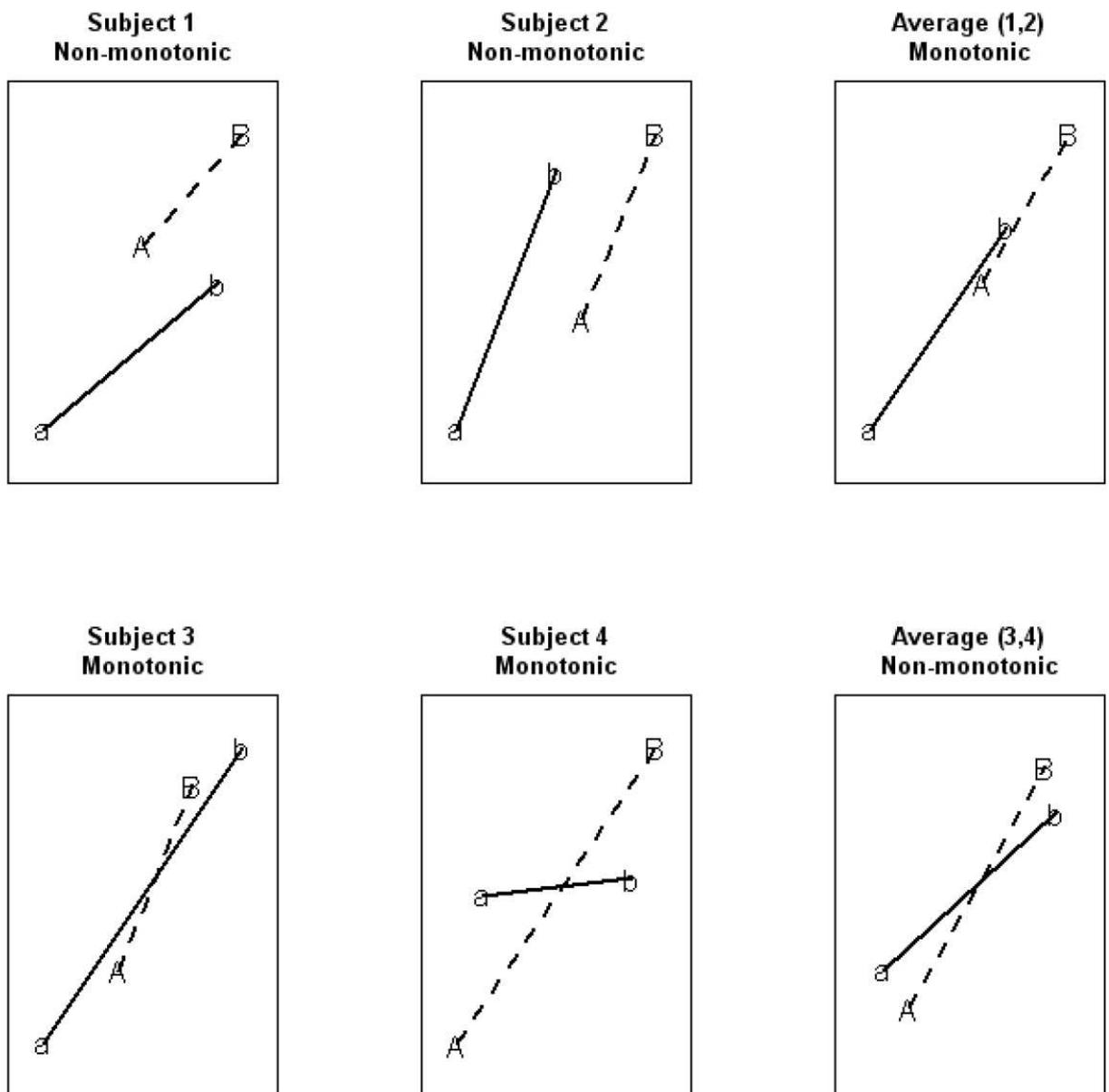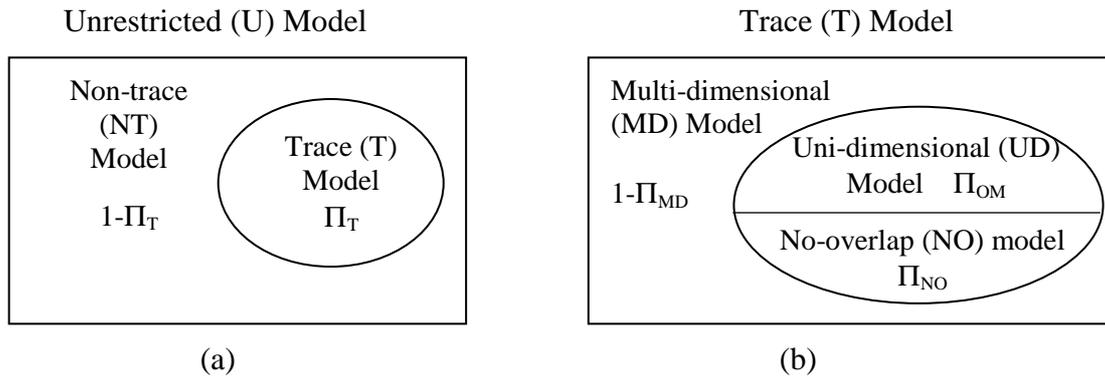
(b)
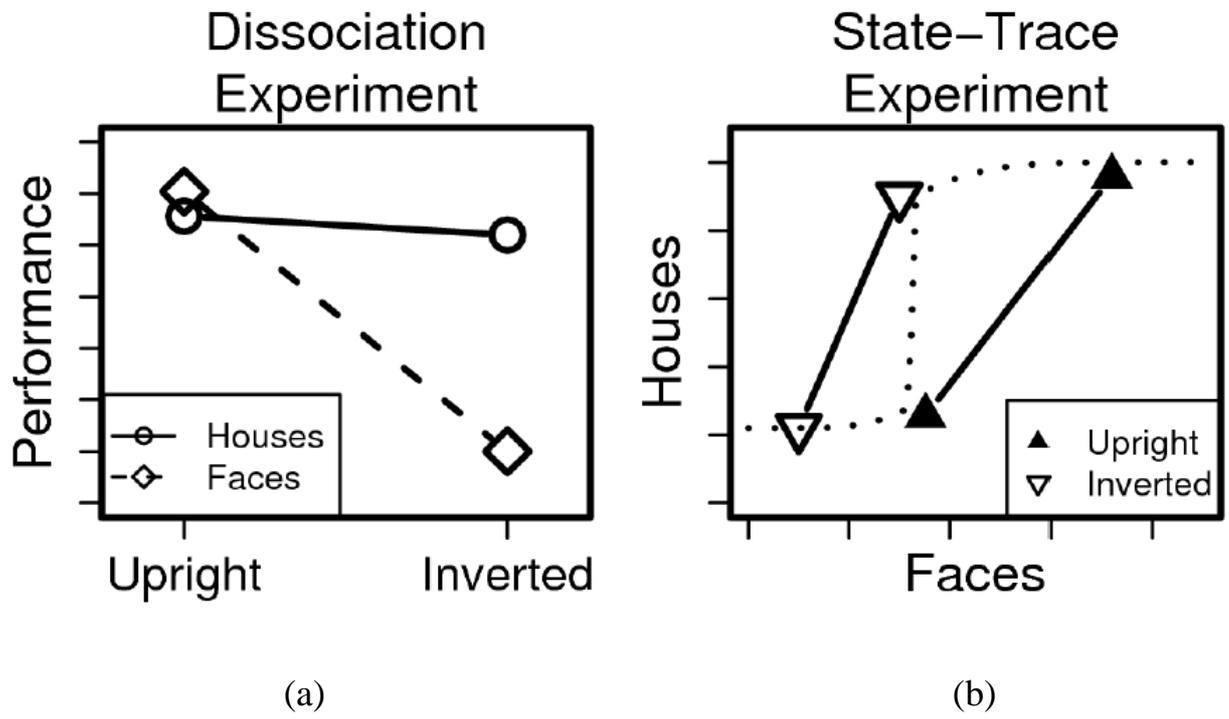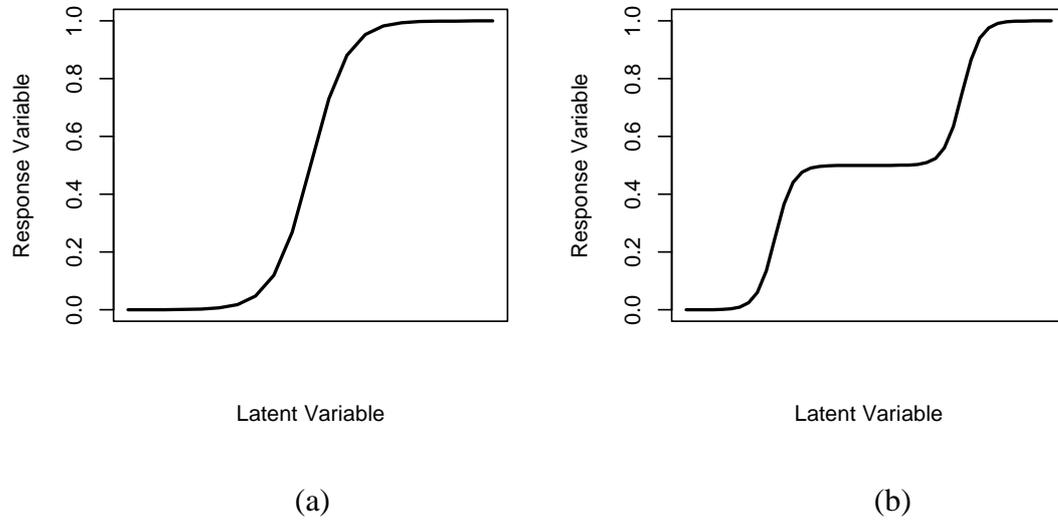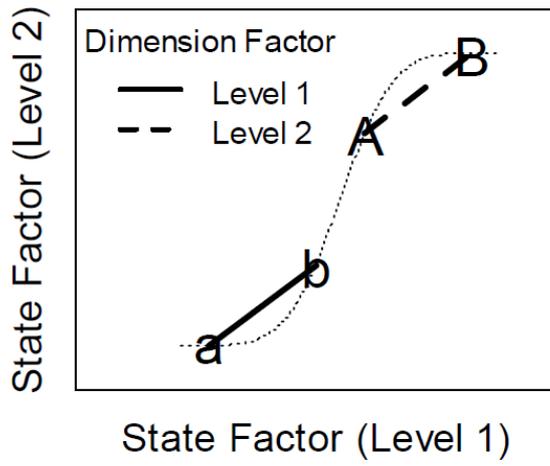
Figure 2

(a)
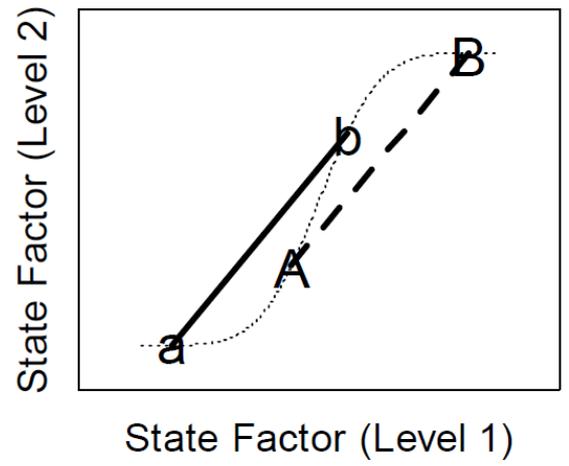
(b)
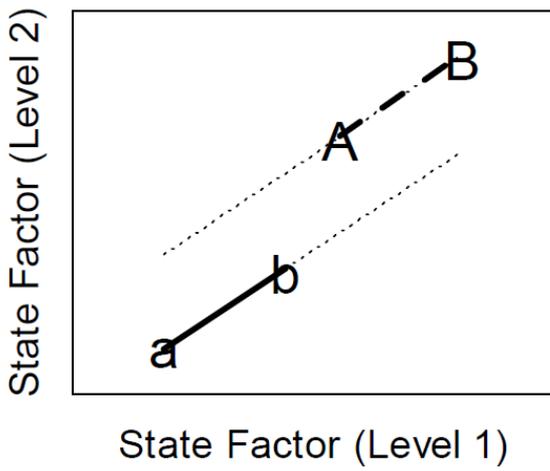
(c)
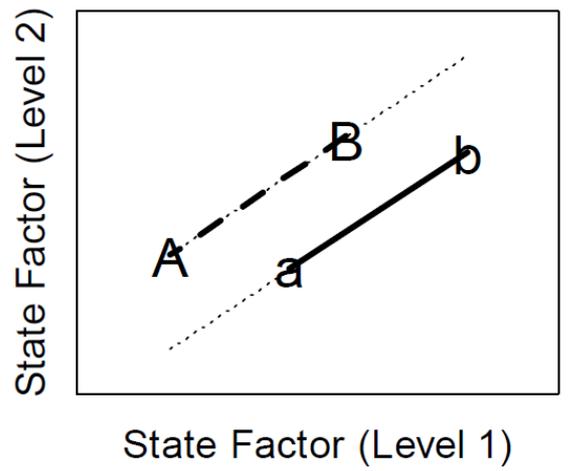
(d)

Figure 3

(a)

(b)

(c)

Figure 4

Figure 5

Unrestricted (U) Model

Non-trace
(NT)
Model

Trace (T)
Model

$1-\Pi_T$

Π_T

(a)

Trace (T) Model

Multi-dimensional
(MD) Model

$1-\Pi_{MD}$

Uni-dimensional (UD)
Model Π_{OM}

No-overlap (NO) model
Π_{NO}

(b)

Figure 6

Participant G Inverted Test (mode)
Pr(posterior regions) = 0.5

(a)

Participant j Upright Test (mode)
Pr(posterior region) = 0.5

(b)

Figure 7

(a)

(b)

Figure 8

Participant average posterior modes
p(posterior region)=0.5

(a)

Participant average posterior modes
p(posterior region)=0.5

(b)

Figure 9