A statistical test of the equality of latent orders

Michael L. Kalish[a,*], John C. Dunn[b], Oleg P. Burdakov[c], Oleg Sysoev[c]

[a] *Syracuse University, USA*

[b] *University of Adelaide, Australia*

[c] *Linköping University, Sweden*

Abstract

It is sometimes the case that a theory proposes that the population means on two variables should have the same rank order across a set of experimental conditions. This paper presents a test of this hypothesis. The test statistic is based on the coupled monotonic regression algorithm developed by the authors. The significance of the test statistic is determined by comparison to an empirical distribution specific to each case, obtained via non-parametric or semi-parametric bootstrap. We present an analysis of the power and Type I error control of the test based on numerical simulation. Partial order constraints placed on the variables may sometimes be theoretically justified. These constraints are easily incorporated into the computation of the test statistic and are shown to have substantial effects on power. The test can be applied to any form of data, as long as an appropriate statistical model can be specified.

Keywords: state-trace analysis, monotonic regression, hypothesis test

*Corresponding address: Prof. M. L. Kalish, Department of Psychology, Syracuse University, Syracuse, NY, 13244

Email addresses: `mlkalish@syr.edu` (Michael L. Kalish), `john.c.dunn@adelaide.edu.au` (John C. Dunn), `oleg.burdakov@liu.se` (Oleg P. Burdakov), `oleg.sysoev@liu.se` (Oleg Sysoev)

Introduction

Consider an experiment in which data are obtained on two different variables across k different conditions. We would like to know if these data are drawn from populations whose means on the two variables have different orders. That is, we ask if the variables have unequal *latent orders*. This question arises in the theory of *state trace analysis* (STA) where inferences concerning the number of latent variables underlying changes in two or more dependent variables depend on the ordinal arrangements of their respective population means (Bamber, 1979; Prince et al., 2012a). STA contrasts a *one-dimensional model*, in which changes in the dependent variables are mediated by one latent variable, and a *two-dimensional model*, in which changes are mediated by more than one latent variable (Loftus et al., 2004; Newell & Dunn, 2008). Under the assumption of the one-dimensional model that each dependent variable is a (distinct) monotonic function of the single latent variable, this model predicts that the latent orders of the two variables are equal. It follows that if the variables have different latent orders across a set of experimental conditions then the effects must be mediated by more than one latent variable.

Implementation of STA requires a statistical procedure to test whether two sets of population means have the same order across a set of conditions. To our knowledge, at least three previous approaches to this problem have been proposed in the psychological literature. The first of these, described by Loftus et al. (2004), relies on reducing sampling error to near zero thereby using the observed sample means as a proxy for the population means. Clearly, this approach cannot be applied in situations with non-negligible sampling error and it lacks a means of quantifying when the sampling error is small enough to be ignored. The second approach, described by Pratte & Rouder (2012), quantifies the effects of sampling error but is limited to particular theory-dependent dependent variables and to a fixed two-by-two factorial design. The third approach, described by Prince et al. (2012a), uses Bayesian model selection to test whether two sets of population means have the same or different orders. While the approach is in principle quite general, the particular implementation described by Prince et al. (2012a) applies only to binomial data and to a relatively constrained factorial design. We discuss this approach in greater detail below and compare it to the test that we develop.

The test we present here is a null hypothesis statistical test (NHST), based on the computation of an empirical p-value of the data given the null hypothesis. Despite the well known problems with p-values (Wagenmakers, 2007), the evidence provided by them remains useful; e.g., it predicts future replicability (Open Science Collaboration, 2015).

The outline of the paper is as follows. First, we describe more fully the logic of our statistical test, based on an extension of monotonic regression (Burdakov et al., 2012). In so doing, we introduce the concept of partial order constraints and foreshadow how they may be used to increase statistical power. Second, we describe a null hypothesis significance test of the equality of latent orders based on a bootstrap resampling procedure for estimating the empirical sampling distribution of the test statistic. Third, we examine the statistical power of our procedure for a fully randomized design with and without partial order constraints. Finally, we extend the procedure to binomial data and compare it to the Bayesian model selection approach developed by Prince et al. (2012a).

The orders of sample and population means

Consider two different dependent variables, x and y, observed across k different experimental conditions. Let $x_1, \ldots, x_k$, $y_1, \ldots, y_k$, be the k population means of each variable and let $X_1, \ldots, X_k$, $Y_1, \ldots, Y_k$, be the corresponding sample means. We define the (latent) order of x as a permutation, $O(x) = (i_1, i_2, \ldots, i_k)$, such that, $x_{i_1} \leq x_{i_2} \leq \ldots \leq x_{i_k}$. We wish to test the hypothesis that $O(x) = O(y)$, given the data. A desirable feature of such a test is that it should be sensitive to both the number and magnitude of differences in the two orders. Intuitively, given equal latent orders, numerically small violations of equality of the orders of the observed means are more likely than numerically large violations. This property is a feature of monotonic (or isotonic) regression (Robertson et al., 1988). Our test is based on this method.

Monotonic Regression

Monotonic regression addresses the problem of finding the best approximation, $\hat{X}$, to a set of observed values, X, under the constraint that $O(\hat{X})$ is known, either completely or partially. Let K be the set of integers, $\{1, 2, \ldots, k\}$. We represent a partial (or total) order on K by means of a subset of ordered pairs $(i, j) \in E \subseteq K \times K$[1]. An order, $O(\hat{X})$, is *consistent* with E if $\hat{X}_i \leq \hat{X}_j, \forall (i, j) \in E$. Formally, let X be a set of k values, let v be a set of corresponding weights, and let E be a partial order. Then monotonic regression finds a set of values, $\hat{X}$, consistent with E, that best approximates X in a weighted least-squares sense. That is, $\hat{X}$ solves the monotonic regression (MR) problem,

$$\min \sum_{i=1}^{k} v_i (X_i - \hat{X}_i)^2 \text{ , subject to } \hat{X}_i \leq \hat{X}_j \text{, for all } (i, j) \in E \tag{1}$$

The choice of weights is critical for obtaining a meaningful 'best' $\hat{X}$. In this respect, we are guided by the property that the solution of Equation (1) is the maximum likelihood estimate if the observations in each condition are independent and normally distributed with weights given by the precision of the data weighted by the number of observations in each condition (Robertson et al., 1988). That is,

$$\begin{aligned} v_i &= \frac{n_{x_i}}{S^2_{X_i}} \\ w_i &= \frac{n_{y_i}}{S^2_{Y_i}} \end{aligned} \tag{2}$$

[1]Unless otherwise stated, a partial order, E, is assumed to be transitively closed.

where $S^2_{X_i}$ is the sample variance of variable x in condition i and $S^2_{Y_i}$ is the sample variance of variable y in condition i.

In many situations the observations in each condition are not independent, as when conditions are manipulated within participants rather than between. In this case the maximum likelihood estimate depends on the entire covariance matrix and the sets of weights, v_i and w_i, are replaced by appropriate matrices. For this reason, we generalize Equation (2) in the following way. Suppose there are g groups of participants of size n_i, $i = 1, \ldots, g$, each measured under m different conditions on variable x. The total number of conditions is thus $k = gm$. Let $\mathbf{S}_i$ be the $m \times m$ covariance matrix for group i. Then the corresponding weight matrix is given by the following block-diagonal matrix,

$$V = \begin{bmatrix} n_1 {\mathbf{S}_1}^{-1} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & n_g {\mathbf{S}_g}^{-1} \end{bmatrix} \tag{3}$$

The weight matrix, W, for variable y is similarly defined[2]. $\mathbf{S}_i^{-1}$ approximates the inverse of the population covariance matrix, $\mathbf{\Sigma}_i^{-1}$. A better estimate of $\mathbf{\Sigma}_i^{-1}$ can be obtained by first 'shrinking' $\mathbf{S}_i$, which reduces the unreliable off-diagonal elements but does not necessarily set all of them to zero (Ledoit & Wolf, 2004). We use Ledoit-Wolf method to adjust the weight matrices in our current approach.

Let X be a vector of k sample means and let $\hat{X}$ be a vector of values. Then, with the weight matrix V defined by Equation (3), the MR problem is given by,

$$\min \left(X - \hat{X}\right)^T V \left(X - \hat{X}\right), \text{ subject to } \hat{X}_i \leq \hat{X}_j, \text{ for all } (i,j) \in E \tag{4}$$

We write the problem corresponding to Equation (4) as $\mathrm{MR}(X, V, E)$ and

[2] We assume that observations on x and y are themselves independent.

the minimum value as $\omega(X, V, E)$, or, in shorthand form, as ω_X. Finding the solution to the MR problem is not trivial, but fast algorithms have been developed. If E is a total order then the MR problem can be solved using the *pool-adjacent-violators algorithm* (PAVA), a version of which was used in the original development of non-metric multidimensional scaling (Kruskal, 1964). Otherwise, the problem as posed in Equation (4) can be solved using quadratic programming algorithms (de Leeuw et al., 2009). The functions *lsqlin* (equivalently, *quadprog*) and *lsei* implement this algorithm in MATLAB® and R (R Core Team, 2013) respectively. In addition, a rapid approximate solution may also be obtained using the *generalized pool-adjacent-violators* (GPAV) algorithm developed by Burdakov et al. (2006).

Coupled monotonic regression

Monotonic regression can be extended to incorporate the additional constraint that the fitted values on two variables are themselves monotonically ordered. That is, $O(\hat{X}) = O(\hat{Y})$. This defines the following *coupled monotonic regression* (CMR) problem: Given two sets of values X and Y, corresponding weight matrices, V and W, and a partial order[3], E, we wish to find $\hat{X}$ and $\hat{Y}$ that are solutions to $\mathrm{MR}(X, V, E)$ and $\mathrm{MR}(Y, W, E)$, respectively, while satisfying the additional coupled monotonicity constraint,

$$\begin{aligned} \hat{X}_i < \hat{X}_j \Rightarrow \hat{Y}_i \leq \hat{Y}_j \\ \hat{Y}_i < \hat{Y}_j \Rightarrow \hat{X}_i \leq \hat{X}_j \end{aligned} \tag{5}$$

This constraint can also be expressed succinctly as follows. If Equation (5) holds then there is no (i, j) such that,

$$(\hat{X}_i - \hat{X}_j)(\hat{Y}_i - \hat{Y}_j) < 0. \tag{6}$$

If there is an (i, j) such that Equation (6) is true then the corresponding pair of points is called *infeasible* and the sets, $\hat{X}$ and $\hat{Y}$, are called infeasible solutions.

[3]Note that in the CMR problem, but not in the MR problem, E can be empty.

To formalize the CMR problem, we note that for a given partial order, E, there is a set of all total orders, $\mathcal{L}(E) \supset E$, called the *linear extensions* of E. The CMR problem can then be stated as the problem of finding sets, $\hat{X}$, $\hat{Y}$, and $\hat{E}$, that solve,

$$\begin{gathered}\min \left[\left(X-\hat{X}\right)^T V\left(X-\hat{X}\right)+\left(Y-\hat{Y}\right)^T W\left(Y-\hat{Y}\right)\right] \\ \text{subject to, } \hat{X}_i \leq \hat{X}_j\,, \hat{Y}_i \leq \hat{Y}_j \text{ for all } (i,j) \in \hat{E},\ \hat{E} \in \mathcal{L}(E)\end{gathered} \tag{7}$$

We write the problem corresponding to Equation (7) as CMR(X, Y, V, W, E), shorthand CMR(E), and the minimum value as $\omega(X, Y, V, W, E)$, shorthand ω_{XY}.

One way of solving the CMR problem defined by Equation (7) is by direct search. While this is guaranteed to find a global minimum, it can be exceptionally slow, as it requires evaluation of a potentially very large number of total orders. For example, for $k = 10$ and $E = \varnothing$, there are $k! = 3,628,800$ orders to search. To circumvent this problem, Burdakov et al. (2012) recently devised the *CMR algorithm* that finds a solution in approximately exponential rather than factorial time. We briefly describe that algorithm here and provide pseudo-code in the Appendix.

The CMR algorithm is a branch-and-bound algorithm that can be viewed as an intelligent search through the set of linear extensions of a specified partial order, E. Given E, which may be empty, it progressively adds additional order constraints until an optimal solution is reached.

On each iteration, a new extension, $E' \supset E$, is considered. For this E', if the corresponding MR solutions, X' and Y', are feasible, i.e. they satisfy monotonicity constraint (5), then the fit of these values provides an upper bound on ω_{XY} (improved, if possible, on each iteration). If the sets X' and Y' are infeasible, however, the corresponding fit provides a lower bound on ω_{XY} for any extension $E'' \supset E'$. The algorithm then chooses an infeasible $(i, j) \notin E'$, and branches by generating two new extensions, $E' \cup \{(i,j)\}$ and $E' \cup \{(j,i)\}$. These extensions inherit the lower bound associated with E' and are added to the set of those to be further considered (tested). This set forms a queue because its elements, all extensions of E, are sorted in increasing value of their inherited lower bounds and the solution, $\hat{E}$, is guaranteed to be an extension of at least one member of the queue. In addition, on each iteration, the algorithm generates a feasible solution based on extending E' in several ways and choosing the one with the best fit. This fit is used for possible improvement of the currently available upper bound on ω_{XY}.

If the obtained fit for any E' is greater than the current upper bound then E', as well as all its extensions, can be eliminated from the search. This leads to the improvement in performance over direct search. The algorithm continues branching and eliminating until the queue is empty or if the inherited upper bound of the first member in the queue is greater than the current best upper bound. The final upper bound is the fit of the optimal least-squares solution, ω_{XY}.

In a worst case scenario involving uncorrelated variables and $E = \emptyset$, simulations confirm that the CMR algorithm converges to the optimal solution as a function of $\exp(k)$ rather than $k!$. Even in this case, the relative speed up is substantial. For example, for $k = 10$, the CMR algorithm evaluates on average about 25 sub-problems in contrast to a direct search of over three million sub-problems. In addition, to the extent that the variables are correlated over conditions and order constraints are specified in E, the algorithm will converge at an even faster rate.

Insert Figure 1 here

Example application of the CMR algorithm

Figure 1 shows a state-trace plot based on results found by Nosofsky et al. (2005) in their Experiment 1. The axes correspond to performance on two different categorization tasks (called “RB” and “II”, respectively). The experimental conditions consisted of a sequence of eight blocks of training trials followed by two blocks of re-training trials that differed between the two groups: a button-switch group who exchanged the position of the response buttons between training and re-training, and a control group who did not. The plot shown in Figure 1 was first generated by Dunn et al. (2012) who used it to discuss whether these data constituted evidence for the existence of more than one latent variable. The first step in answering this question is to solve the CMR problem and determine the fit of the best-fitting monotonically-related set of points. Dunn et al. were unable to solve this problem previously for two reasons. First, they only had direct search method available to them which was unable to find a solution in a practical period of time[4]. Second, the relevant data is a mixture of conditions, one of which was varied within-participants (trial block), the other between-participants (response switch vs. no switch). This requires use of the corresponding weight matrices defined by Equation (3).

Figure 1 also shows the optimal CMR solution, connected by dashed lines to aid visibility. The actual fit value, ω_{XY}, corresponding to the solution of Equation (7), was 3.514. This value depends upon the sample means, X and Y, the weight matrices, V and W, computed according to Equation (3), and the pre-defined partial order, E (empty in this case).

The partial order, E, may be used to specify prior knowledge concerning an expected order of the population means over a sub-set of conditions. In the present case, each group participated in 10 blocks of learning trials with the

[4] On a standard desktop, finding the CMR solution for the current problem by direct search would take approximately 10 hours. In contrast, the CMR algorithm produced the solution in approximately 0.1 seconds.

first eight corresponding to successive blocks of training on the same task. It is reasonable to assume that the population means should not decrease over these blocks. It may be similarly argued that the population means should not decrease across the two post-switch blocks, 9 and 10, in each group. Based on these considerations, it is possible to impose a partial order constraint on the solution to the CMR problem. Note that within this partial order, although the first eight blocks and the last two blocks are ordered for each group and task, there is no constraint on the order of blocks 8 and 9. Indeed, the possibility of different orders between these conditions on the RB and II variables in the button-switch group was the main theoretical question posed by Nosofsky et al.

If no partial order is specified, the fit value is 3.514 (as noted above). If the partial order constraint is specified then the fit value cannot decrease, and may increase. In the present case, the model fit increases slightly to 3.774 suggesting that the observed means, X and Y, conform closely to the assumed partial order. One reason for imposing a partial order constraint on the solution is that it may lead to a more powerful test of the hypothesis of equal orders. In this case, the test statistic is the difference in fit between a model that assumes only the partial order constraint and a model that assumes both the partial order constraint and coupled monotonicity. We discuss this in the next section.

Hypothesis test

While the CMR algorithm allows us to find a value for ω_{XY}, a substantial problem remains in determining whether this value is large enough to reject the null hypothesis that the population means have the same order. To do this, we first define two models of interest. The one-dimensional model (conditional on E) is defined as follows:

$$M_1: \;\; O(x) = O(y) \quad \& \quad O(x), O(y) \in \mathcal{L}(E)$$

This states that the order of the population means on x is the same as the order on y and that this order is a linear extension of the specified partial order, E.

This model is nested within a two-dimensional model (conditional on E) defined as follows:

$$M_2: \ O(x), O(y) \in \mathcal{L}(E)$$

This states only that the orders on x and y are both (potentially different) linear extensions of the specified partial order, E. Fitting M_2 does not require the CMR algorithm as it consists of two standard MR problems, one in X and one in Y. Further, if $E = \varnothing$, the fit of M_2 is necessarily equal to zero.

At present there is no statistical test of the loss in fit from M_2 to M_1. In the simpler case of (ordinary) monotonic regression, some work has been done on developing a test of the hypothesis, $O(x) \in \mathcal{L}(E)$, against an unconstrained alternative based on the sampling distribution of ω_X. It is known that under this hypothesis, the test statistic follows a $\bar{\chi}^2$ (chi-bar squared) distribution (Robertson et al., 1988). This is a mixture of χ^2 distributions of different degrees of freedom with mixture weights, called level probabilities, which depend in complex ways on the number of conditions, the number of participants, and the partial order, E. As a result, $\bar{\chi}^2$ distributions have been calculated for only a few relatively simple cases. While it may be possible to extend this approach to coupled monotonic regression, we have not attempted this, as it seems likely that calculation of the theoretical distribution would encounter even greater difficulties.

Our test of the fit of M_1 against the fit of M_2 is constructed by empirically estimating the sampling distribution of the difference in respective fits under the assumption that M_1 is the true model. The method is adapted from the bootstrap re-sampling procedure described by Wagenmakers et al. (2004). As these authors point out, their procedure cannot be directly applied when the models to be compared are nested. Since M_1 is nested in M_2, M_2 always fits better than M_1. For this reason, the fit of M_1 can only be compared against the fit of M_2. The steps in this procedure are as follows:

Let $\mathbf{X}$ and $\mathbf{Y}$ and be two data sets, let X and Y be vectors of the corresponding sample means, and let V and W be the corresponding weight matrices. Let E be a specified partial order.

1. Using the CMR algorithm, find the observed fit of M_1, $\omega_{XY} = \omega(X, Y, V, W, E)$. Using any suitable MR algorithm, find $\omega_X = \omega(X, V, E)$ and $\omega_Y = \omega(Y, w, E)$, and calculate the observed fit of M_2, $\omega_{X+Y} = \omega_X + \omega_Y$. If $E = \varnothing$ then $\omega_X = \omega_Y = 0$. Calculate the observed difference in fits, $\delta = \omega_{XY} - \omega_{X+Y}$.
2. Generate two non-parametric bootstrap samples, $\mathbf{X}'$ and $\mathbf{Y}'$, from the corresponding data sets. This step is undertaken in order to incorporate sampling error in parameter estimation. Calculate the corresponding sample means, X' and Y', and weight matrices, V' and W'.
3. Solve the CMR problem for the bootstrap samples and, using X', Y', V' and W', find the best-fitting values, $\hat{X}'$ and $\hat{Y}'$.
4. Transform the original data so that the means are now equal $\hat{X}'$ and $\hat{Y}'$. That is, form new samples, $\mathbf{X}_T = \mathbf{X} - X + \hat{X}'$ and $\mathbf{Y}_T = \mathbf{Y} - Y + \hat{Y}'$, and, from these, draw a second set of non-parametric bootstrap samples, $\mathbf{X}'_T$ and $\mathbf{Y}'_T$. Calculate the corresponding sample means, X'_T and Y'_T, and weight matrices, V'_T and W'_T, respectively.
5. Using the CMR algorithm, find the observed fit of M_1, $\omega'_{XY} = \omega(X'_T, Y'_T, V'_T, W'_T, E)$. Using any suitable MR algorithm, find $\omega'_X = \omega(X'_T, V'_T, E)$ and $\omega'_Y = \omega(Y'_T, W'_T, E)$, and calculate the observed fit of M_2, $\omega'_{X+Y} = \omega'_X + \omega'_Y$. Calculate and store the sample difference in fits (for current iteration i), $\delta'_i = \omega'_{XY} - \omega'_{X+Y}$.
6. Repeat Steps 2-5 N times where N is a sufficiently large number (e.g., 10,000).
7. Calculate, p, the proportion of values of δ'_i that are greater than or equal to δ. If $p < \alpha$ then reject the null hypothesis.

The above procedure can also be adapted to test the fit of M_2 for $E \neq \varnothing$. In this case, the procedure is modified by replacing M_1 by M_2 and replacing

M_2 by the unconstrained model, the fit of which is necessarily zero. The steps of this procedure are as follows:

Let $\mathbf{X}$ and $\mathbf{Y}$ and be two data sets, let X and Y be vectors of the corresponding sample means, and let V and W be the corresponding weight matrices. Let E be a specified partial order.

1. Using any suitable MR algorithm, find $\omega_X = \omega(X, V, E)$ and $\omega_Y = \omega(Y, W, E)$, and calculate the observed fit of M_2, $\omega_{X+Y} = \omega_X + \omega_Y$. Calculate the observed difference in fits[5], $\delta = \omega_{X+Y} - 0$.
2. Generate two non-parametric bootstrap samples, $\mathbf{X}'$ and $\mathbf{Y}'$, from the corresponding data sets. Calculate the corresponding sample means, X' and Y', and weight matrices, V' and W'.
3. Solve the MR problems for each of the bootstrap samples and, using X', Y', V' and W', find the best-fitting values, $\hat{X}'$ and $\hat{Y}'$.
4. Form new samples, $\mathbf{X}_T = \mathbf{X} - X + \hat{X}'$ and $\mathbf{Y}_T = \mathbf{Y} - Y + \hat{Y}'$, and, from these, draw a second set of non-parametric bootstrap samples, $\mathbf{X}'_T$ and $\mathbf{Y}'_T$. Calculate the corresponding sample means, X'_T and Y'_T, and associated weight matrices, V'_T and W'_T, respectively.
5. Using any MR algorithm, find $\omega'_X = \omega(X'_T, V'_T, E)$ and $\omega'_Y = \omega(Y'_T, W'_T, E)$, and calculate the fit of M_2, $\omega'_{X+Y} = \omega'_X + \omega'_Y$. Calculate and store the sample difference in fits, $\delta'_i = \omega'_{X+Y} - 0$.
6. Repeat Steps 2-5 N times where N is a sufficiently large number (e.g., 10,000).
7. Calculate, p, the proportion of values of δ'_i that are greater than or equal to δ. If $p < \alpha$ then reject the null hypothesis.

Each of the hypothesis tests outlined above rely on two principal elements, the CMR algorithm and bootstrap re-sampling. Because both of these are quite

[5] We include the notional subtraction of zero, the fit of the unconstrained model, to highlight the parallels between the two procedures.

general, the procedure can be applied to a wide variety of research designs. The experimental conditions can be fully randomized across participants, applied entirely within-participants, or any combination of between- and within-participant treatments. The procedure may also be adapted for discrete data although, in this case, the model-consistent data, $\mathbf{X}_T$ and $\mathbf{Y}_T$, are derived from a parametric bootstrap of the observed data (in step 3 above). However, this is not a substantial concern as this relevant distribution is entirety specified by the data so parametric and non-parametric re-sampling are equivalent. We discuss the application of the method to binomial data in a later section.

Insert Figure 2 here

Example application

To illustrate the application of the hypothesis testing procedure, we return to the state-trace plot shown in Figure 1. Figure 2 shows two empirical distributions of δ', each based on 10,000 iterations, and two observed values of δ. The dashed line and unfilled triangle are based on the assumption of no partial order, $E = \varnothing$. In this case, the two-dimensional model fits perfectly (as it is unconstrained) and δ is equal to the observed fit of the one-dimensional model and has the value of 3.514 (as noted earlier). The corresponding empirical p-value is 0.77 from which it is concluded (for $\alpha = .05$) that the hypothesis $O(x) = O(y)$ cannot be rejected.

If the partial order, E, described earlier in relation to the data shown in Figure 1, is implemented then the testing procedure differs. The first step is to test the fit of M_2 which has an observed fit of 0.929. The corresponding empirical p-value is 0.72 from which it is concluded that the hypothesis $O(x), O(y) \in \mathcal{L}(E)$ cannot be rejected. Following this, the next step is to test the difference in fit between M_1 and M_2. The solid line in Figure 2 shows the estimated empirical distribution of δ' and the filled triangle shows the observed value of δ. As stated earlier, the observed fit of the one-dimensional model (M_1) is fractionally increased to 3.774. However, the value of δ is now $3.774 - 0.929 = 2.845$, and the corresponding empirical p-value is 0.57. We again conclude that the hypothesis, $O(x) = O(y)$, given $O(x), O(y) \in \mathcal{L}(E)$, cannot be rejected.

Although in this case, both analyses (with and without assuming a prior partial order) lead to the same conclusion, inspection of Figure 2 illustrates the increase in statistical power that may result from the addition of an appropriate partial order constraint. Although δ has decreased from the no-partial-order to the partial-order case, this difference is relatively small compared to the difference in the shapes of the corresponding empirical distributions. Specifically, the distribution of δ', when the partial order is specified (filled curve), is contracted leftwards compared to the sampling distribution of δ', when no partial order is specified (dashed curve). As a result, relatively less mass falls to the right of the observed value of δ leading to a lower p-value and an associated increase in statistical analysis. The reason for this is that, if the population means satisfy the partial order constraint, E, then the fit of M_2 will be close to zero. However, many of the bootstrap samples of M_1 may violate the partial order in which case the fit of M_2 will be substantially greater than zero, thereby contracting the distribution of δ'.

Analyzing power

It is desirable that our proposed test have sufficient power to reject the null hypothesis of equal latent orders when it is false. We address this issue in the present section where our goals are; (1) to define a measure of effect size in post-hoc power analyses, (2) to show how power can be estimated for any given effect size, (3) to discuss the problem of estimating effect size for proactive power analyses, and finally (4) to demonstrate the effect on power of imposing partial order constraints.

We consider in detail a measure of effect size for a fully-randomized between-participant experiment with n participants in each of k conditions. In this case the true effect size is the fit, ω_{xy}, of the solution to the following CMR problem:

$$\begin{aligned} &\omega_{xy} = \min\left[(x-\hat{x})^T \Upsilon\,(x-\hat{x}) + (y-\hat{y})^T \Psi\,(y-\hat{y})\right] \\ &\text{subject to, } (\hat{x}_i - \hat{x}_j)(\hat{y}_i - \hat{y}_j) \geq 0, \text{ for all } (i,j) \end{aligned} \tag{8}$$

where, $\Upsilon = \operatorname{diag}\left(\sigma^2_{x_1}, \ldots, \sigma^2_{x_k}\right)^{-1}, \Psi = \operatorname{diag}\left(\sigma^2_{y_1}, \ldots, \sigma^2_{y_k}\right)^{-1}$.

For convenience we set $\sigma^2_{x_i} = \sigma^2_{y_i} = \sigma^2$ for all i. Both the number of violations of monotonicity and the size (relative to the population precision) of each violation determine the value of ω_{xy}, so in order to explore the power of the CMR test we varied both of these over a wide range. We adopted the case where $k = 8$, and set $x = \{1, \ldots, 8\}$. We manipulated the number of violations of monotonicity from 1 to 28 by choosing y as a permutation of $\{1, \ldots, 8\}$ to produce the desired number of violations. For each number of violations, we then varied σ^2 in order to generate effect sizes ranging from 0.1 to 10. This process resulted in a set of 398 combinations of means, variances, and associated effect sizes which were used to estimate power for various sample sizes.

Insert Figure 3 here

For each of these 398 sets of population parameters we drew a sample data set consisting of k independent, normally distributed, samples, each of size n, for $n = \{10, 20, 30, 40, 50\}$ for each variable, x and y. For each data set, we followed the 7-step procedure presented earlier to determine whether M_1 could be rejected for two levels controlling the Type I error rate, $\alpha = .05$ and $\alpha = .01$. Because each data set was drawn from a population in which the monotonic component of M_1 is false, the observed proportion of correct rejections is an estimate of the power, $(1 - \beta)$, of the test. The results of these simulations are shown in Figure 3. Each power curve corresponds to the best fitting logistic function of the effect size, ω_{xy}, for each value of n.

The curves shown in Figure 3 can be used to estimate the number of participants that an experimenter may need in order to achieve a given level of power for the fully randomized design considered above. To do so, it is necessary to estimate ω_{xy}. For the present equal-n design, an obvious estimates of that ω_{xy} is given by ω_{XY}/n. For designs with unequal n between groups, the corresponding estimate is $\omega_{XY}/\bar{n}$, where $\bar{n}$ is the mean n over groups[6] These curves allow a researcher to make a rough claim about the scale of the observed effect size. In the case of Cohen's (1988) d, the scale relates to power as follows: a small effect has a power of .1 with$n = 20$, medium has a power of .2, and large about .4. For the CMR test, this corresponds to δ of 0.1, 0.2, 0.4 as power is nearly linear at that level with $n = 20$. A very large effect (power .8) would be $\delta = 0.50$.

Power under partial order constraints

In this section, we re-examine the potential increase in power due to the addition of a partial order constraint. Because there are a very large number of possible partial order constraints, we focus on one that naturally arises in a factorial design. Consider an experiment with two between-participant factors, A and B, such that A has two levels and B has 4 levels (i.e., $k = 8$). A prior belief may exist concerning the orders of the dependent variables on each factor. We suppose that for each level of B, level 1 of A will produce smaller values on both dependent variables (e.g., less accurate responding, lower response times) than will level 2. We further suppose that for each level of A, the levels of B will conform to a particular total order. By way of an example, an experiment may examine the effect on recognition memory of a change in the format of visually presented words and study duration. In this case, factor A is presentation format (two levels: same format at study and test, different formats) and B is study duration (4 levels: say, 0.25 sec, 0.5 sec, 1 sec, 2 sec). Based on prior knowledge, it is plausible to assume that memory for words presented in the same format

[6]Although an obvious approach, it is likely that reliance on ω_{XY} may underestimate ω_{xy}. Further research on this question is required.

will be no worse than memory for words presented in different formats, while, for each format, memory will not decrease with increasing study duration. In order to illustrate the effect of this prior partial order on statistical power, we simulated the case of $n = 10$ for each group, using the procedure described previously.

Insert Figure 4 here

Figure 4 reveals the gain in power that results from imposing the proposed partial order. The addition of this constraint leads to a nearly five-fold increase in the rate of increase of the power curve compared to the no-partial order case. The relevant measure of effect size when there is a partial order is the difference between M_1 and M_2, δ_{xy}. In order to achieve power equal to 0.8 at α =.05, we found that the observed effect size in the partial order case was $\delta_{xy} = 0.13$, a value substantially less than the observed effect size in the no-partial order case, $\delta_{xy} = 0.78$. The corresponding population variances were 0.18 and 0.03, respectively. In order to give some sense of how this may appear in the data, we drew a random data set from populations with each of these variances and summarized these in the state-trace plots shown in Figure 5. The larger variance in the partial-order case is striking. In our experience, measurements with variability of this magnitude are not difficult to find in psychological experiments.

Insert Figure 5 here

As noted earlier, the imposition of a partial order reduces the variance of the distribution of δ, the difference in fit between M_1 and M_2, as long as the population means conform to the partial order. On the other hand, if the population means do not conform to the partial order then both M_1 and M_2 are false. Because power is necessarily limited, Type II errors are always possible.

The test of the partial order model, M_2, is at best a check that the experiment has been correctly designed. Furthermore, a partial order should not be adopted merely to facilitate rejection of M_1. In order to be logically coherent, any partial order should be defined prior to conducting the experiment and be based on a compelling and universally accepted motivation.

The power analysis presented above is useful for post-hoc analyses, where the effect size can be estimated from data. However, its use in prospective power estimation is limited because the estimate of the effect size depends on the particular design. For example, in the previous simulations, we assumed a uniform spacing of x and y which may be unlikely to occur in practice. In the context of state-trace analysis, the optimal design is one which maximizes δ_{xy} given a particular two-dimensional manifold of possible latent means in the state space. This, in turn, will depend upon the configuration of latent means selected from the manifold through selection of the experimental factors and the number and nature of their levels. Similarly, repeated measures will affect power in ways that are dependent on the particulars of the variance-covariance matrix. A prospective power analysis will thus require the experimenter to essentially replicate a sub-set of our procedure for the design under consideration.

Control of Type I error

Our method is based on bootstrap resampling. An advantage of this approach is that no assumption is required concerning the nature of the distribution of observations[7]. However, bootstrap samples may underestimate variance for small n (Chernick, 2007) which can lead to a corresponding inflation of the Type I error rate. For this reason we conducted a series of simulations in which we replaced the bootstrap samples with samples from the known distribution

[7] Although, of course, if the data are not normally distributed the obtained values of ω and δ will not be maximum likelihood estimates.

from which the data were drawn (in this case, a normal distribution). In each simulation, the population means were monotonically related; they were, for each variable, simply the integers 1 to 8, and no partial order was assumed. We manipulated the variance of each distribution and the sample size, both of which were assumed to be constant over conditions and variables. On each simulation, for a given variance and sample size, a sample data set was drawn and the CMR procedure applied to generate an empirical distribution of fits (based on 10,000 samples). The procedure was applied both in its bootstrap form (as described earlier) and in a form in which the bootstrap step was replaced by re-sampling from the normal distributions used to generate the data. We then used the latter, parametric, empirical distribution to identify cut-offs for different percentiles including the 95th and 99th percentiles corresponding to $\alpha = 0.05$ and $\alpha = 0.01$, respectively. We then calculated the proportion of cases that exceeded these cut-offs in the empirical distribution derived from the bootstrap method. So long as resampling did not produce degenerate cases (which did not occur with $n > 8$ in our simulations) the percent of the cases that exceeded the cut-off deviated very little from the expected proportions.

Extension of the CMR procedure to binomial data

In this section, we describe how the CMR procedure can be extended to binomial data structures. We also take the opportunity to compare this procedure to the Bayesian model selection approach developed by Prince et al. (2012a), highlighting their similarities and differences.

Some notations are introduced first. Let n_x be a (column) k-vector of the number of Bernoulli trials for variable x on each of k conditions. Let a_x be the (column) k-vector of the number of successes in each condition and let b_x be the corresponding vector of the number of failures, where $n_x = a_x + b_x$. Let X be the vector of the observed mean proportion of successes for variable x across k conditions, i.e. $X = a_x/n_x$, where the division is understood to be element-wise. The same kind of notation can be introduced for variable y. We seek to solve the CMR problem given by Equation (7).

With $V = \text{diag}(n_x)$ and $W = \text{diag}(n_y)$, the least-squares solution to the problem given by Equation (7) is also the maximum likelihood solution. This follows from Theorem 12 of Robertson et al. (1988, p. 32) which states that the solution, $\hat{X}$, to the least-squares monotonic regression on X with weights, n_x, is also the maximum likelihood solution. Because the solution to Equation (7) is the sum of two monotonic regression problems for some $\hat{E}$, it follows that it is also the maximum likelihood solution. The only difference in applying it to binomial data is that evaluation of sub-problems in the CMR algorithm is based on the actual likelihood function rather than evaluation of Equation (7). Equivalently, it can be based on the following negative log-likelihood function:

$$f(\hat{X}, \hat{Y}) = -(a_x^T \ln(\hat{X}) + b_x^T \ln(1 - \hat{X}) + a_y^T \ln(\hat{Y}) + b_y^T \ln(1 - \hat{Y}))$$

Because the value of this function is non-zero when the fit is perfect, it is convenient to subtract the corresponding value of the perfect fit, $f(X, Y)$. This leads to an equivalent formulation in terms of the G^2-statistic:

$$G^2 = 2[f(\hat{X}, \hat{Y}) - f(X, Y)]$$

Application to binomial data

Prince et al. (2012b) analyzed a set of binomial data using the Bayesian model selection procedure described by Prince et al. (2012a). These data were obtained from a two-alternative forced-choice recognition memory experiment that investigated the face-inversion effect, based on a similar study by Loftus et al. (2004). The stimuli were pictures of faces or houses which defined the dependent variables of interest (i.e., memory accuracy for faces and memory accuracy for houses). Performance was tested under the orthogonal combination of two factors; stimulus orientation (upright vs. inverted), and study duration (short, medium, and long). All experimental factors (stimulus type, orientation, and duration) were manipulated within-participants. The data for each participant (as well as data aggregated over participants) consists of counts of successes (i.e., selecting the correct test item) and counts of failures (i.e., selecting the incorrect test item) for each stimulus type under each of the six experimental conditions[8].

The three different study durations imply a partial order on performance. Namely, the proportion of successes should not decrease from short to medium and from medium to long durations for both upright and inverted presentation formats for both face recognition and house recognition. For consistency with Prince et al. we did not place a partial order on the upright and inverted conditions, although this could readily be included.

Insert Figure 6 here

Figure 6 shows the state-trace plot based on the mean proportion of successes averaged over all participants. The dashed line shows the best fitting monotonic curve. It is clear that for each dependent variable, the effect of study duration is consistent with the assumed partial order. These data may be analyzed in three different ways using CMR. First, the mean scores of proportion correct (corresponding to the points plotted in Figure 6) can be analyzed using the original CMR procedure described earlier, assuming a normal distribution of means across participants. In this case, the empirical p-value based on 10,000 iterations is 0.044, which implies rejection of the monotonic model, M_1, at $\alpha = 0.05$. Second, the counts of successes and failures can be aggregated over participants and these data analyzed using the binomial CMR procedure. In this case, the empirical p-value of δ based on 10,000 iterations is 0.017, also implying rejection of M_1. However, as Prince et al. have pointed out, aggregation over

[8]The authors are grateful to Melissa Prince and colleagues for making these data available.

participants has the potential to distort the underlying pattern of the data. For this reason, they analyzed each participant separately, which leads to the third way in which the data can be analyzed using binomial CMR. In this case, consistent with the analysis of the aggregated data, none of the p-values for M_2 were significant (minimum $p = 0.079$). On the other hand, none of the p-values for M_1 against M_2 reached significance (minimum $p = .062$). This is to be expected given the low power associated with the smaller number of observations for each participant. Given this, it is desirable to combine this evidence in a manner that does not lead to distortions due to averaging (Davis-Stober et al., In Press). This can be done by conducting a test of the *sum* of the individual fits. Such a test is equivalent to using the binomial CMR procedure to fit M_1 and M_2 to a concatenated set of kn conditions with a partial order constraint and a monotonicity constraint applied to each set of k conditions for each of the n participants. In practice, the relevant statistics can be obtained from the individual analyses already conducted - the sum of the model fits across participant is compared against the distribution of the sum of random samples drawn from the individual empirical distributions obtained from the bootstrap procedure. Consistent with the aggregated data which exactly conform to the partial order constraint, the combined p-value for the test of M_2 is not significant ($p = 0.817$). However, the combined p-value for the test of M_1 against M_2 fell short of significance ($p = 0.084$)[9].

Insert Figure 7 here

Comparison to Bayesian model selection approach

In order to compare the results of the binomial CMR procedure with the Bayesian model selection developed by Prince, et al. (2012), is necessary to

[9]Based on 100,000 combined samples each corresponding to the sum of 18 individual random samples from the individual empirical distributions.

explain their approach in some detail and to identify the points of similarity and difference with the CMR approach. Figure 7 summarizes the main features of the two approaches. The left hand side of Figure 7 shows a binary tree generated by the sequential addition of order constraints. The top-most model is the unconstrained model (called the *encompassing model* by Prince et al.), which, by definition, fits the observed data perfectly. The second level contrasts two models defined by the addition of the partial order constraint, $O(x), O(y) \in \mathcal{L}(E)$, where $\mathcal{L}(E)$ is the set of linear extensions of the specified partial order, E. The model for which this constraint is true is called the *trace model* by Prince et al., and the model for which it is false is called the *non-trace model*. The Bayesian procedure directly compares these models and selects the one with the greater posterior model probability. In contrast, the CMR procedure tests if the addition of the partial order constraint leads to a statistically significant decrease in goodness of fit. Following the Bayesian procedure, if the trace model is selected[10] then two additional models are contrasted at the third level, defined by the addition of the monotonicity constraint, $O(x) = O(y)$. The model for which this constraint is true is called the *monotonic model* by Prince et al., and the model for which it is false is called the *multidimensional model*. Again, the Bayesian procedure directly compares these two models while the CMR procedure tests the loss of fit caused by the additional monotonicity constraint.

Finally, Prince et al. proposed a binary contrast at a fourth level, between two complementary models called the *overlap* and *non-overlap models*. In the experimental design used by Prince et al., non-overlap means that the effect of stimulus orientation (upright vs. inverted) is sufficiently large that there is no overlap between the sets of data points corresponding to the three stimulus durations. If this occurs, the resulting state-trace is trivially monotonic and Prince et al. advised that the experiment should be re-designed. Let $\mathcal{L}'(E)$ and

[10]Prince et al. describe both a sequential and simultaneous model evaluation procedure. We describe the sequential approach for expository purposes.

$\mathcal{L}''(E)$ be a partition of $\mathcal{L}(E)$ such that $\mathcal{L}'(E)$ is the set of linear extensions of E consistent with overlap and $\mathcal{L}''(E)$ is the set of extensions inconsistent with overlap. The final constraint is therefore that, $O(x), O(y) \in \mathcal{L}'(E)$.

An apparent advantage of the Bayesian procedure is that it allows the weight of evidence for pairs of disjoint models at each level of constraint to be directly compared. In contrast, a null hypothesis statistical test, which forms the heart of our procedure, tests whether the addition of a constraint leads to a statistically significant loss of fit. Offsetting this advantage is the necessity of assuming a prior distribution over the set of all possible orders of conditions. Depending on the context, different priors are possible and each choice will lead to a different outcome in model selection. Prince et al. assumed that this prior is uniform.

Analogous to the combined p-value, Prince et al. calculated a group posterior model probability based on combined Bayes factors, essentially the product of individual Bayes factors, and found the probability of the trace model compared to the non-trace model was greater than 0.95. This is analogous to our test of M_2 (against the unconstrained model) which had a combined p-value of 0.85. Consistent with this, the rank order of individual participants' posterior probabilities of the non-trace model is similar (but not identical) to the rank order of the individual fits of M_2, Kendall's tau $= 0.73$, $p < 0.0001$. Prince et al. also found that the group posterior model probability of the monotonic model compared to the multidimensional model was less than 0.05. In contrast, our analogous test of M_1 against M_2 had a combined p-value of 0.070 which fell short of significance ($\alpha = 0.05$). However, the rank order of participants' posterior probabilities for the multidimensional model is similar (but not identical) to the rank order of the difference in fit between M_1 and M_2, Kendall's tau $= 0.42$, $p = 0.007$. Thus, while the two methods are based on different theoretical orientations and procedures, and technically test different models, their commonalities are such that they may well lead to similar conclusions.

Unlike Prince et al., we do not incorporate a test of overlap into our procedure. We have not pursued this option for three reasons. First, it is not essential to the principal question of testing the model of equal orders. Second, the concept appears to be most relevant to the kind of factorial design investigated by Prince et al. It is not clear how it might be relevant to other designs, such as that used by Nosofsky et al. (2005). Finally, it is not clear that the concept of non-overlap is sufficiently inclusive. Given a set of populations that have different orders (i.e., where M_1 is false), there are many configurations of sample means that will be trivially monotonically ordered[11]. Non-overlap is but one example. In our view, the failure to reject M_1 requires further analyses of the data to determine whether this is due to the configuration of sample means. Such follow-up analysis is analogous to inspection of the scatterplot to aid interpretation of a correlation coefficient. If the data are trivially monotonic, the pattern of points will suggest possible changes to the levels of the experimental factors to increase the chance of rejecting M_1 (assuming it is false). Prince et al. made similar recommendations and suggested that, in attempting to maximize power, it may be useful to adopt non-standard factorial designs.

We endorse consideration of non-standard factorial designs. In such designs, the levels of one factor may differ across levels of the other factor. For example, in the face-inversion study conducted by Prince et al., stimulus durations for the more difficult inverted condition may be longer than corresponding durations for the easier upright condition. Such choices maximize the chance that some pairs of points in the state-trace plot will violate monotonicity. It must be remembered that even if the underling state-trace is two dimensional (with unequal latent orders), this will only be revealed in the observed data if the configuration of points contains violations of monotonicity. This, in turn, will depend in complex ways on the levels of the factors that have been manipulated. Depending on these levels, violations may or may not be observed.

[11]For the design used by Prince et al., other examples include the lack of an effect of either or both experimental factors, or a 'staircase' arrangement of points in the state space which suggest two-dimensionality but fail to produce any violations of monotonicity.

Conclusion

We have presented a comprehensive procedure for testing for the equality of latent orders. The procedure consists of two main parts: (1) The CMR algorithm that finds the best single order on two dependent variables over k conditions and returns a measure of the lack-of-fit of that order to the data; (2) a significance test for this lack-of-fit, based on bootstrap resampling. Consistent with experience of the bootstrap (Chernick, 2007), we showed that this test controls Type I error rate for sample sizes greater than eight. We also showed that the power of the test was a function of effect size and sample size for a fully randomized, equal n, design and that it obtained reasonably high levels of power (> 0.80) for data that could plausibly occur in typical psychology experiments. We also demonstrated the role of partial orders, or pre-experimental order constraints on conditions, in substantially increasing power in the case where the partial order is true.

Although we presented the CMR procedure principally in relation to continuous data, we showed how it can be readily extended to discrete data and discussed the binomial case in some detail. A feature of the procedure for continuous data is that it permits a non-parametric bootstrap. Thus, it is not necessary to make any distributional assumptions. Nor is it necessary to assume equal variances or equal n, at least in a fully randomized design, as unequal precisions are explicitly built into the monotonic regression weights.

No discussion of hypothesis testing should ignore the crucial differences between Bayesian and frequentist approaches. Our bootstrap method provides a frequentist estimate of the variability of the CMR fit estimate. It should be possible to construct an alternative Bayesian approach to examining latent orders using CMR, and Bayesian hypothesis tests for state-trace applications of latent order testing without CMR already exist (Davis-Stober et al., In Press; Prince et al., 2012b). One critical feature that divides the Bayesian and frequentist approaches is the treatment of model complexity. The equal-order model is less complex than the alternative, where each variable follows its own (partial) order. Our frequentist approach does not penalize the separate-order model, because its complexity is unknown. Because the common-order model is nested within the separate-order model, the latter will always fit better than the former. We recommend rejection of the common-order model when the probability of the fit being as bad as is observed is small. The Bayesian approach does penalize for complexity, by specifying priors for both models. The separate-order model will have a more diffuse prior than the common-order model, making it possible to compare the models to each other and accept either one. This bi-directional decision is enabled only by making specific assumptions about what the appropriate prior should be for both models. Such priors equate to theories about the data generating processes. On the one hand, such theories are critical to advancing our understanding of the process that give rise to observed data. On the other hand, disagreement about what theories are reasonable will necessarily extend to the results of Bayesian hypothesis testing. We have argued that there is a role for a procedure that makes minimal assumptions about the distribution of latent orders, and we believe that our NHST approach is informative within that context.

We motivated the development of the CMR procedure by reference to its relevance to state-trace analysis where the presence of different latent orders implies that the dependent variables are functions of more than one latent variable. For this reason, we discussed the application of the CMR procedure to two dependent variables, as commonly used in STA. However, the procedure can also be readily generalized to test the equality of latent orders over any number of dependent variables.

A further, intriguing, challenge is to consider the more complex case in which the latent orders of N dependent variables conform to a linear space of $d < N$ dimensions (Dunn & James, 2003). For $N = 2$ dependent variables, equal latent orders implies that $d = 1$. For $N > 2$ and $d > 1$, different constraints will apply to generate sets of permitted N-tuples of orders. While this problem poses a number of significant difficulties, its solution would lead to a general test of latent orders beyond simple equality.

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

Burdakov, O. P., Dunn, J. C., & Kalish, M. L. (2012). An approach to solving decomposable optimization problems with coupling constraints.

Burdakov, O. P., Sysoev, O., Grimvall, A., & Hussian, M. (2006). An o(n2) algorithm for isotonic regression. In G. di Pillo, & M. Roma (Eds.), *Large-Scale Nonlinear Optimization* (pp. 25–33). New York: Springer volume 83 of *Nonconvex Optimization and Its Applications*.

Chernick, M. R. (2007). *Bootstrap methods: A guide for practitioners and researchers*. New York: Wiley.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences (2nd Edition)*. Routledge.

Davis-Stober, C., Morey, R. D., Gretton, M., & Heathcote, A. (In Press). Bayes factors for state-trace analysis. *Journal of Mathematical Psychology*, .

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

Dunn, J. C., Newell, B. R., & Kalish, M. L. (2012). The effect of feedback delay and feedback type on perceptual category learning: The limits of multiple systems. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *38*, 840–859.

Kruskal, J. (1964). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, *29*, 115–129.

Ledoit, O., & Wolf, M. (2004). Honey, i shrunk the sample covariance matrix. *The Journal of Portfolio Management*, *30*, 110–119.

de Leeuw, J., Hornik, K., & Mair, P. (2009). Isotone optimization in r: Pool-adjacent-violators algorithm (pava) and active set methods. *Journal of Statistical Software*, *32*, 1–24.

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*, 835–863.

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

Nosofsky, R. M., Stanton, R. D., & Zaki, S. R. (2005). Procedural interference in perceptual classification: Implicit learning or cognitive complexity? *Memory & Cognition*, *33*, 1256–1271.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, *349*. URL: `http://www.sciencemag.org/content/349/6251/aac4716.abstract`. doi:`10.1126/science.aac4716`. `arXiv:http://www.sciencemag.org/content/349/6251/aac4716.full.pdf`.

Pratte, M. S., & Rouder, J. N. (2012). Assessing the dissociability of recollection and familiarity in recognition memory. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *38*, 1591–1607.

Prince, M., Brown, S., & Heathcote, A. (2012a). The design and analysis of state-trace experiments. *Psychological Methods*, *17*.

Prince, M., Hawkins, G., Love, J., & Heathcote, A. (2012b). An r package for state-trace analysis. *Behavior Research Methods*, *44*, 644–655.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria. URL: `http://www.R-project.org/`.

Robertson, T., Wright, F. T., & Dykstra, R. L. (1988). *Order restricted statistical inference*. Chichester, UK: John Wiley & Sons.

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

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

Appendix

CMR algorithm

The following pseudo-code describes the CMR algorithm (Burdakov et al., 2012). Here, X and Y vectors of means, V and W are corresponding weight matrices, E is a specified partial order partial order and $F(\hat{X}, \hat{Y})$ is the objective function value in (3) computed for the vectors $\hat{X}$ and $\hat{Y}$. L is a list of pairs of the form (e, f) where e is a partial order and f is the value of the corresponding inherited lower bound.

Input: X, Y, V, W, E. Output: $\hat{X}$, $\hat{Y}$, $F(\hat{X}, \hat{Y})$.
$L = \{(E, -\infty)\}$, $F_U = \infty$, $F_L = -\infty$
while $(|L| > 0)$ & $(F_L < L_U)$ **do**
    $(E', F_L) \leftarrow L(1)$
    if $F_L < F_U$ **then**
        find X' that solves $\mathrm{MR}(X, v, E')$ and Y' that solves $\mathrm{MR}(Y, w, E')$
        compute $F(X', Y')$
        if $F(X', Y') < F_U$ **then**
            if (X', Y') is feasible **then**
                $F_U \leftarrow F(X', Y'),\ (\hat{X}, \hat{Y}) \leftarrow (X', Y')$
            else
                generate feasible solution (X'', Y'') and compute $F(X'', Y'')$
                if $F(X'', Y'') < F_U$ **then**
                    $F_U \leftarrow F(X'', Y''),\ (\hat{X}, \hat{Y}) \leftarrow (X'', Y'')$
                end
                find (i, j) such that $(X'_i - X'_j)(Y'_i - Y'_j) < 0$
                $E'_{ij} \leftarrow E' \cup \{(i, j)\}\,,\ E'_{ji} \leftarrow E' \cup \{(j, i)\}$
                append $(E'_{ij}, F(X', Y'))$ and $(E'_{ji}, F(X', Y'))$ to L
                reorder $L = \{\ldots, (e, f), \ldots\}$ in increasing values of f
            end
        end

```
    end
end
```

Acknowledgements

The authors wish to thank Ben Newell, EJ Wagenmakers, Andrew Heathcote, John Kruschke, Laura Anderson, and Don Bamber for their helpful discussions on many related topics. The authors gratefully acknowledge the continued support of the Australian Research Council (Discovery Grants: 0877510, 0878630, 110100751, and 130101535), the National Science Foundation (Award 1256959) to Kalish, and two visiting fellowships from Linköping University to Dunn.

Figure Captions

1. Data from Nosofsky, Stanton, and Zaki (2005, Experiment 1). State-trace plot of mean proportion correct on RB and II category structures for each block of trials in the learning or pre-switch phase (Blocks 1-8 only) and in the post-switch or transfer phase (final two blocks for each group). In the control condition, the same response assignment was maintained across the two phases. In the button switch condition, the response assignment was switched between learning and transfer phases. Error bars indicate standard errors. Filled symbols correspond to performance in the pre-switch phase. Unfilled symbols correspond to performance in the post-switch phase. Dashed line and crosses indicate the best-fitting monotonic model. Adapted from Figure 1b in *The effect of feedback delay and feedback type on perceptual category learning: The limits of multiple systems*, by J. C. Dunn, B. R. Newell, & M. L. Kalish, 2012, Journal of Experimental Psychology: Learning, Memory, & Cognition, 38(4), pp. 840-859. Copyright 2012 by the American Psychological Association.

2. Empirical distributions of statistic, δ, based on analysis of data from Nosofsky, Stanton, and Zaki (2005, Experiment 1). In the partial order condition, a non-decreasing order is assumed over blocks 1 to 8 and over blocks 9 to 10 in both the control and button-shift groups. Also shown are the observed fit statistics for the data with and without the above partial order, filled and unfilled triangles, respectively.

3. Power plots for the CMR effect size statistic, ω_{xy}, with no partial order constraints and $k = 8$ conditions. (a) Power, $(1 - \beta)$, as a function of effect size, ω_{xy}, and sample size, n_i, for $\alpha = 0.05$. (b) Power, $(1 - \beta)$, as a function of effect size, ω_{xy}, and sample size, n_i, for $\alpha = 0.01$. Note the different scales on the ordinates.

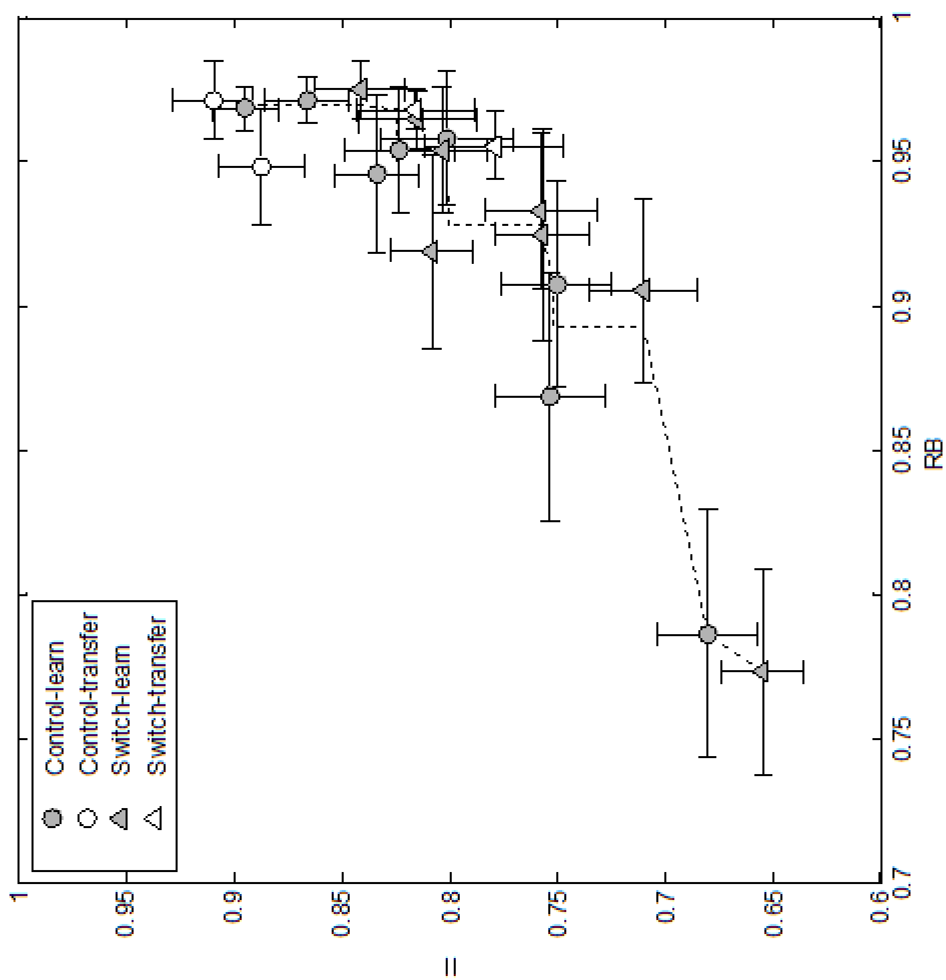
4. Power plots for the CMR effect size statistic, ω_{xy}, with a partial order constraint on $k = 8$ conditions (see text for constraint) compared to without

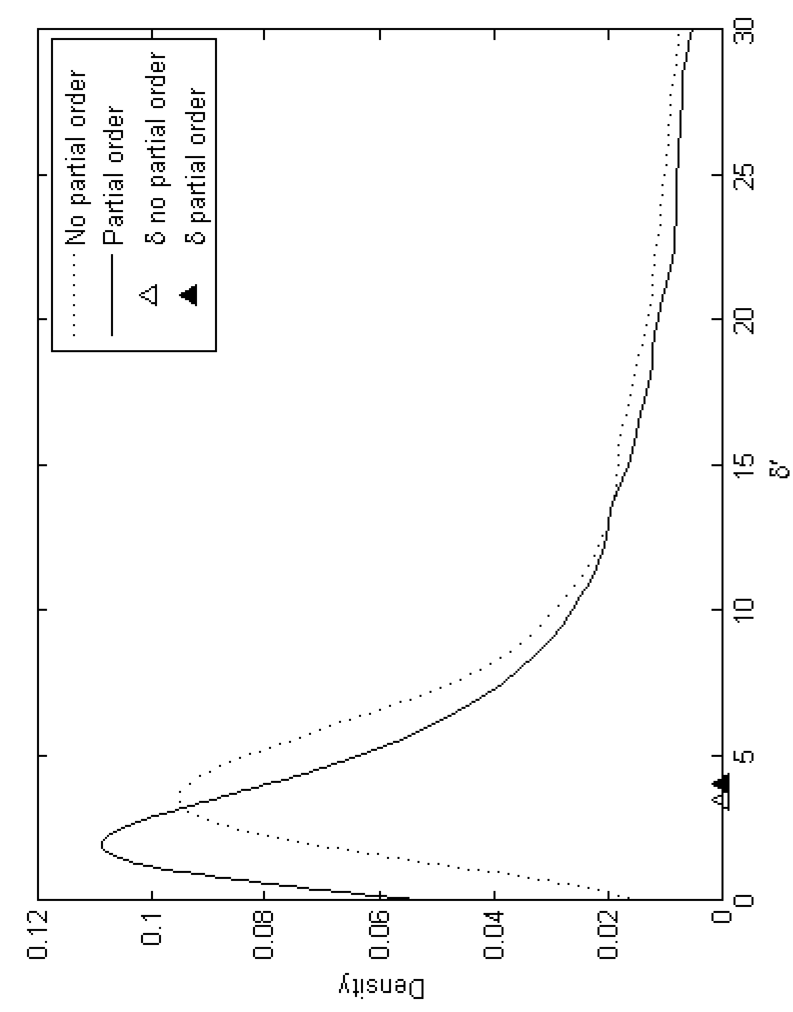
a partial order constraints. (a) Power, $(1-\beta)$, as a function of effect size, ω_{xy}, and sample size, n_i, for $\alpha = 0.05$. (b) Power, $(1-\beta)$, as a function of effect size, ω_{xy}, and sample size, n_i, for $\alpha = 0.01$.

5. State-trace plots of 4 x 2 factorial design corresponding to power of 0.80. (a) Sample means and standard errors under no partial order. (b) Sample means and standard errors under partial order defined on both factors.

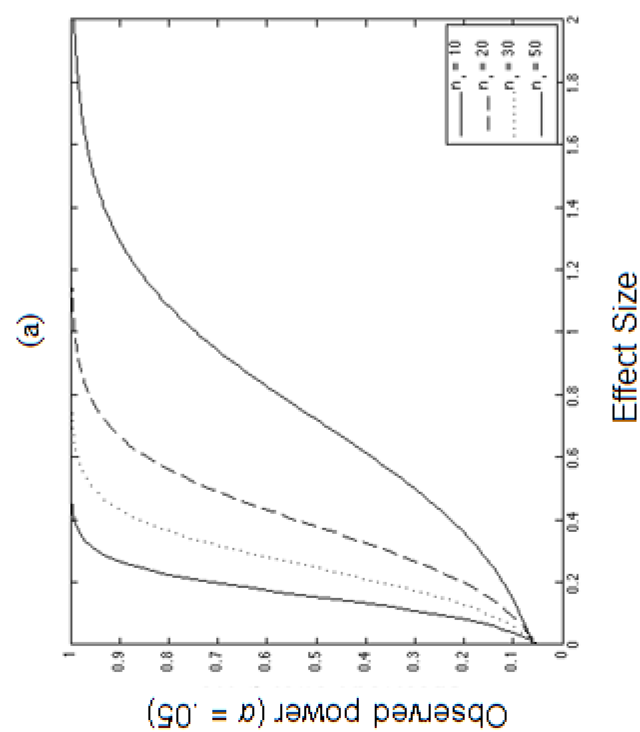
6. State-trace plot of mean proportion correct (averaged over participants) from Prince, Hawkins, Love and Heathcote (2012). The dashed line indicates the best-fitting monotonic curve based on the CMR procedure. Error bars indicate within-participant standard errors calculated according to the Loftus-Masson procedure (Loftus & Masson, 1994).

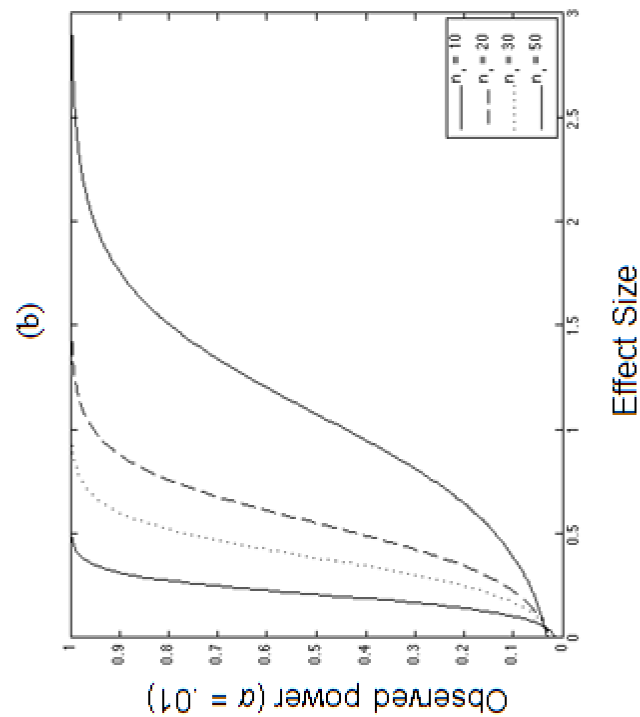
7. Model structure tested by the CMR and Bayesian procedures. The left hand side shows the model tree proposed by Prince, Brown and Heathcote (2012) and tested by their Bayesian model selection procedure. The two models at each level are the complements of each other and the Bayesian procedure selects which of each pair is more strongly supported by the data. The right hand side shows the constraints that added at each level of the tree. The CMR procedure tests if the addition of each constraint leads to a significant decrease in model fit. See text for a definition of each term.

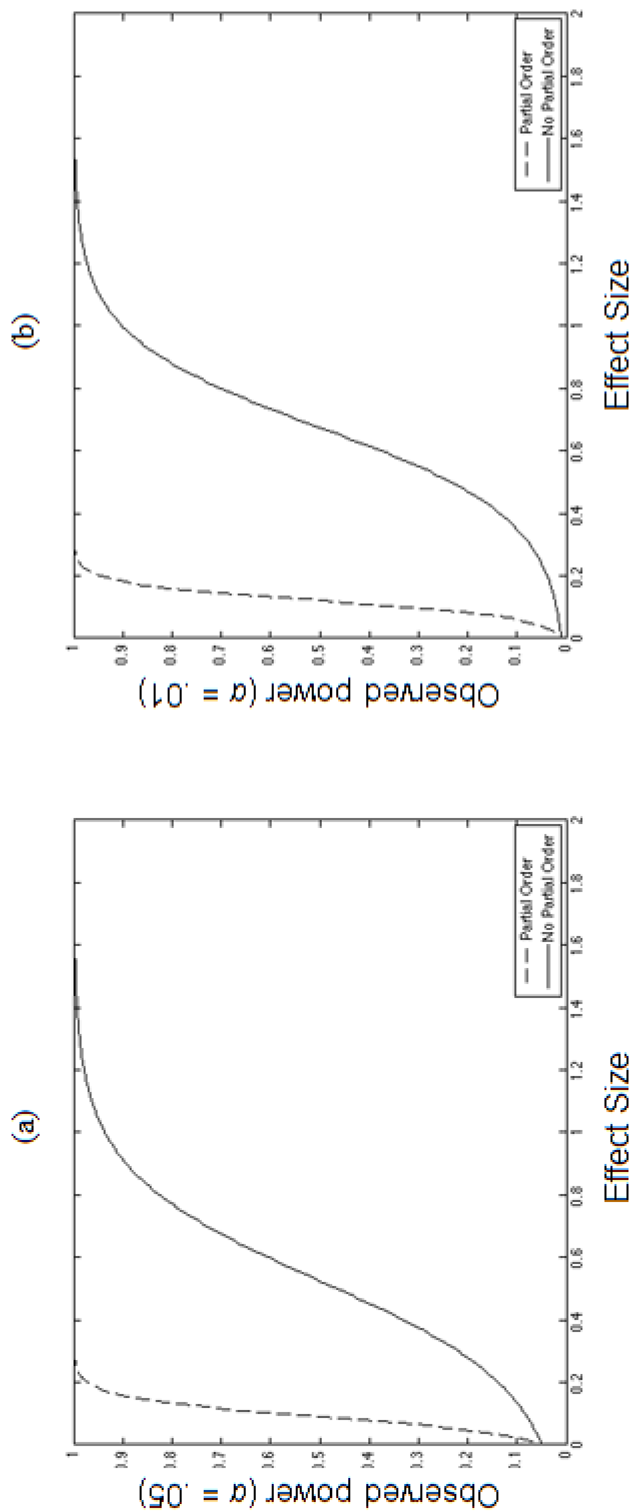

Control-learn
Control-transfer
Switch-learn
Switch-transfer
RB
II
0.7
0.75
0.8
0.85
0.9
0.95
1
0.6
0.65
0.7
0.75
0.8
0.85
0.9
0.95
1



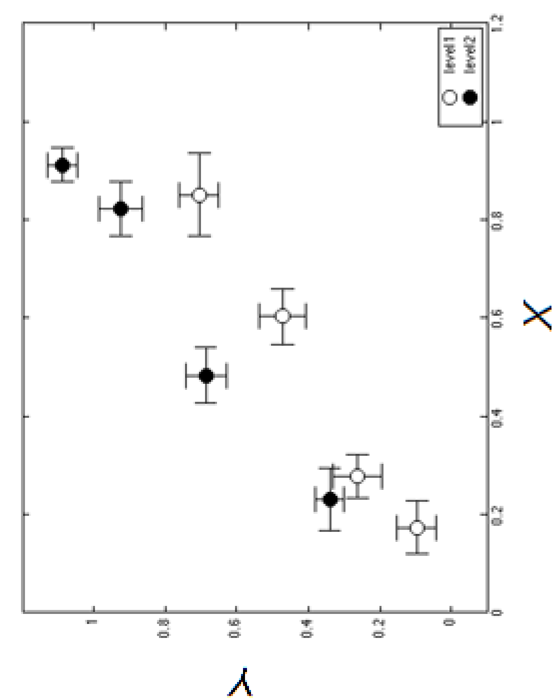
No partial order
Partial order
δ no partial order
δ partial order
Density
δ'
0
0.02
0.04
0.06
0.08
0.1
0.12
0
5
10
15
20
25
30


(a)
Observed power (α = .05)
Effect Size

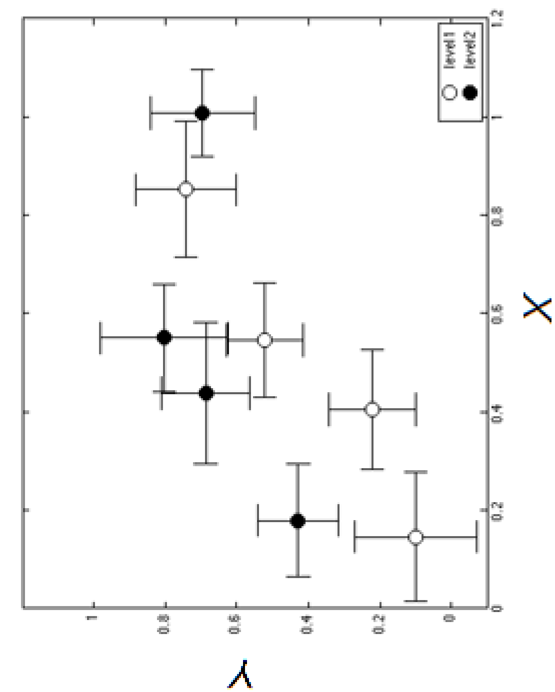

(b)
Observed power (α = .01)
Effect Size

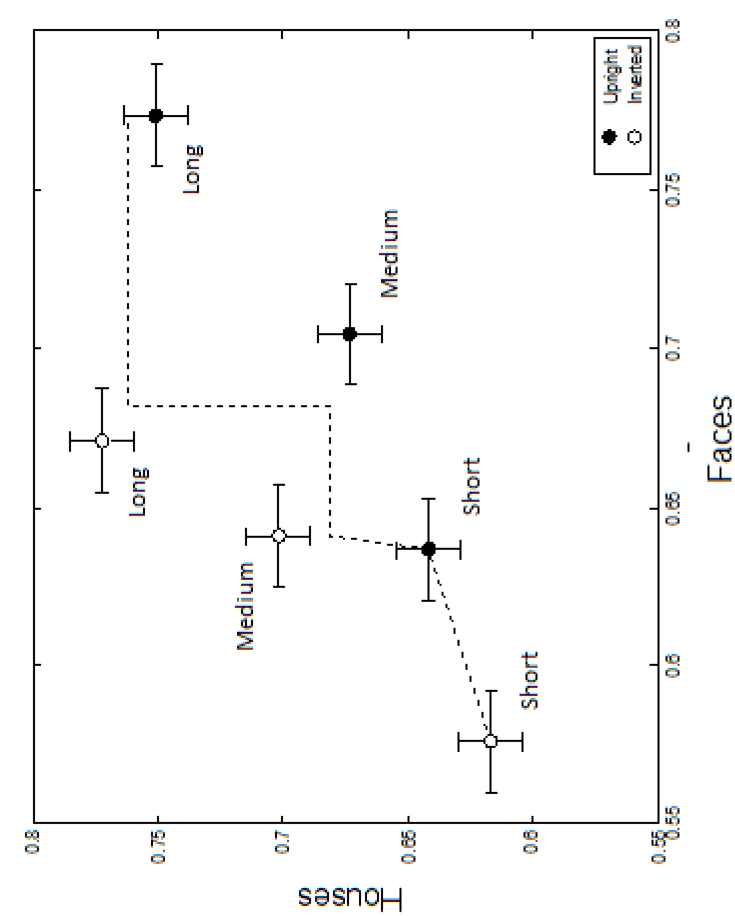

(a)
(b)
Observed power (α = .05)
Observed power (α = .01)
Effect Size
Partial Order
No Partial Order

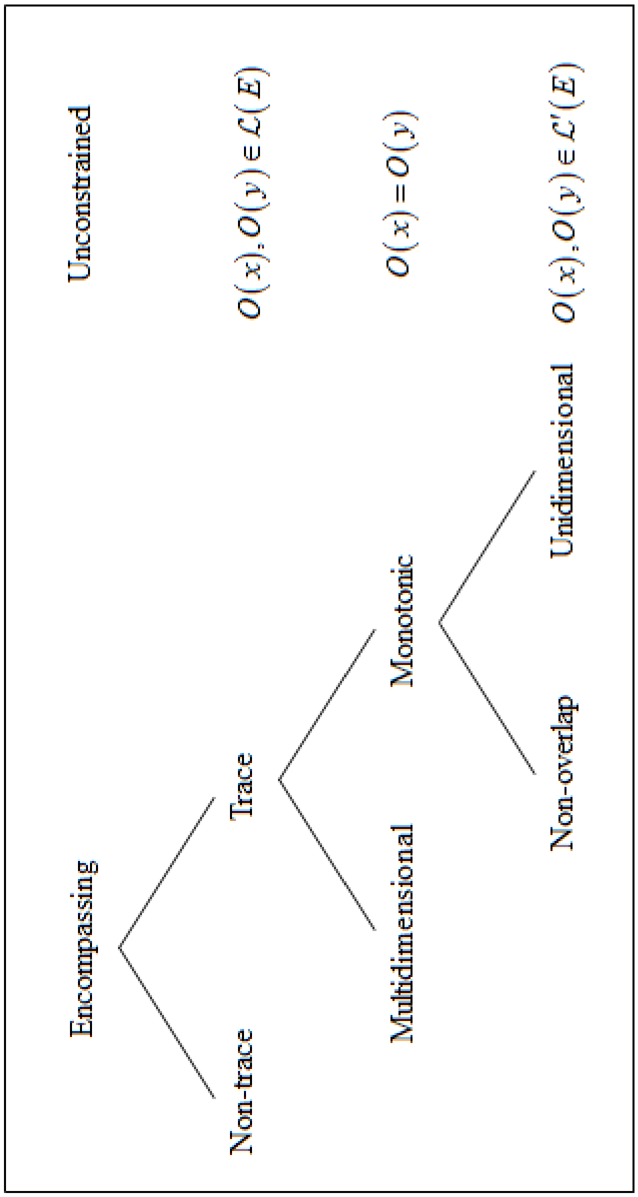
(a)


level1
level2
Y
X
0
0.2
0.4
0.6
0.8
1
1.2

(b)


level1
level2
Y
X
0
0.2
0.4
0.6
0.8
1
1.2


Faces
Houses
Upright
Inverted
Long
Medium
Short
Long
Medium
Short
0.55
0.6
0.65
0.7
0.75
0.8


Unconstrained
Encompassing
Non-trace
Trace
$O(x), O(y) \in \mathcal{L}(E)$
Multidimensional
Monotonic
$O(x) = O(y)$
Non-overlap
Unidimensional
$O(x), O(y) \in \mathcal{L}'(E)$