Designing State-Trace Experiments to Assess the Number of Latent Psychological Variables Underlying Binary Choices

Guy Hawkins (Guy.Hawkins@newcastle.edu.au)
Melissa Prince (Melissa.Prince@newcastle.edu.au)
Scott Brown (Scott.Brown@newcastle.edu.au)
Andrew Heathcote (Andrew.Heathcote@newcastle.edu.au)
School of Psychology, University of Newcastle
University Drive, Callaghan, 2308, NSW Australia

Abstract

State-trace analysis is a non-parametric method that can identify the number of latent variables (dimensionality) required to explain the effect of two or more experimental factors on performance. Heathcote, Brown, and Prince (submitted) recently proposed a Bayes Factor method for estimating the evidence favoring one or more than one latent variable in a state-trace experiment, known as Bayesian Ordinal Analysis of State-Traces (BOAST). We report results from a series of simulations indicating that for larger sample sizes BOAST performs well in identifying dimensionality for single and multiple latent variable models. A method of group analysis convenient for smaller sample sizes is presented with mixed results across experimental designs. We use the simulation results to provide guidance on designing state-trace experiments to maximize the probability of correct classification of dimensionality.

Keywords: State-trace analysis; Bayesian analysis; Bayes Factor; Encompassing prior method; Simulation.

State-Trace Analysis

State-trace analysis (Bamber, 1979), also known as dimensional analysis (Loftus, Oberg, & Dillon, 2004), is a method for determining whether a single latent variable is capable of explaining the joint effect of two experimental factors. Dimensionality is traditionally assessed by testing for an interaction between the two factors. However, interactions can be scale dependent (e.g., distorted by floor or ceiling effects) when response variables are bounded (e.g., accuracy data, see Dunn & Kirsner, 1988; Loftus, 1978). State-trace analysis overcomes this problem by assessing the ordinal relationships between the effects of experimental factors. One factor is comprised of a set of indicator variables, and is referred to as the *state factor*. A second factor is a variable thought to differentially influence performance over levels of the state factor, and is referred to as the *dimension factor*.

State-trace analysis is most easily described by an example. For this purpose we use the *disproportionate face inversion effect* (DFIE), the finding in perceptual and recognition memory studies that stimulus inversion has a more deleterious effect on faces than other mono-oriented stimuli (Valentine, 1988; Yin, 1969). This result has been taken to suggest that faces are encoded along a 'configural' dimension that is not available to mono-oriented non-face stimuli (e.g., houses; Maurer, Le Grand, & Mondloch, 2002). In this example, stimulus type (faces or houses) is the state factor and stimulus orientation (upright or inverted) is the dimension factor, as inversion is thought to differentially affect memory for faces and houses.

State-trace analysis results are shown in a state-trace plot: a scatterplot of the co-variation of performance for the levels of the state factor (e.g., faces or houses). Memory accuracy results for the two levels of the state factor form the two axes of the state-trace plot. Each point on the plot represents a pair of measurements, with a pair of (x, y) coordinates for each level of the dimension factor (e.g., upright and inverted). For our example there would be two coordinate pairs on the state-trace plot, one for upright stimuli and one for inverted stimuli. To infer dimensionality a third variable, referred to as a *trace factor*, is added to the state-trace design. The trace factor sweeps out a set of coordinate pairs for each level of the dimension factor. Levels of the trace factor within each level of the dimension factor are usually connected with a line in the state-trace plot, with each line referred to as a *data trace*.

Latent dimensionality is identified by assessing whether all of the data points in the state-trace plot fall on a single monotonic (i.e., always increasing or always decreasing) function, indicating evidence for a single latent variable. Monotonicity holds if all of the x axis values in a state-trace plot have the same order as the y axis values. Although a monotonic plot is necessary to infer a single latent variable, it is not sufficient: monotonicity cannot be diagnostic unless the data traces overlap on at least one axis. Hence, an assessment of whether data traces overlap is essential to a proper assessment of dimensionality. Similarly, it is important to establish that the trace factor does not itself affect dimensionality, so that results in favor of more than one dimension can be unambiguously attributed to the effect of the dimension factor. This can be checked by determining if the trace factor has a monotonic effect within each level of the dimension factor.

A Bayesian Approach to State-Trace Analysis

Given an observed state-trace plot, where the effects of the underlying latent variable(s) are perturbed by measurement error, how can we determine whether a monotonic curve best describes the data? A number of statistical methods for assessing departures from monotonicity have been suggested (see Loftus et al., 2004; Newell & Dunn, 2008). Recently Heathcote et al. (submitted) proposed a Bayes Factor approach to state-trace model selection, known as Bayesian Ordinal Analysis of State-Traces (BOAST), based on Klugkist, Laudy, and Hoijtink's (2005) encompassing prior method. The encompassing prior method uses Bayes Factors to select

among models defined by inequalities. The advantage of this approach is that it automatically accounts for differences in flexibility amongst models, which is a key issue in state-trace analysis as a one-dimensional model is far less flexible than a multi-dimensional model.

BOAST assumes binomially distributed data (e.g., a binary two-alternative forced choice response used to measure recognition accuracy in the DFIE example), with state-trace models being defined by sets of inequality constraints on binomial probability parameters. For example, we define a 'trace' model, which instantiates the assumption that the trace factor does not change dimensionality, by specifying that the trace factor has a monotonic effect on performance within each level of the dimension factor. This specification implies that, for a trace factor with three levels and an overall increasing effect on accuracy, that accuracy is smaller for the first level of the trace factor than the second level, and smaller for the second level than the third. The trace model is, therefore, an order constrained special case of an 'encompassing' model that places no restrictions on the order of parameters.

When model M_i is an order constrained version of an encompassing model M_k, Bayes Factors can be estimated from prior and posterior samples from the encompassing model (Klugkist, Kato, & Hoijtink, 2005). The proportion of prior ($\hat{\pi}$) and posterior ($\hat{\Pi}$) samples that adhere to the order constraints of the more restricted model M_i are used to estimate a Bayes Factor from the ratio of the two sample counts,

$$BF_{ik} \approx \frac{\hat{\Pi}}{\hat{\pi}}. \tag{1}$$

This Bayes Factor indicates the strength of evidence in favor of M_i over M_k. Intuitively this is the case because it is the ratio of the probability that the model fits the data before the data are observed, which is proportional to the complexity of the model (e.g., the maximally complex encompassing model will always fit any data pattern), to the actual fit of the model to the data. If this ratio is greater than one it indicates that the model fits better than chance.

A set of such Bayes Factors, assuming the same encompassing model, can be used to compare a set of order-restricted models by calculating each models posterior model probability, $p(M_i|D)$, given observed data D. The quantity $p(M_i|D)$ is the probability that model M_i is the 'true' (data generating) model, on the assumption that one model in the set is the true model. Model selection based on $p(M_i|D)$ can also be justified on other grounds, even when the set does not contain the true model (e.g., it selects the model that is most likely to minimize a measure of error in predicting new data), so we refer to it simply as a method of selecting the 'best' model. For a set of models M_i, $1\ldots m$ that are assumed to have a probability p_i of being the best model prior to observing the data, the posterior model probability for M_i is:

$$p(M_i|D) = \frac{p_i \times BF_{ik}}{\sum_{j=1}^{m} p_j \times BF_{jk}} \tag{2}$$

for any $k = 1\ldots m$ which includes i. Throughout we assume each model is equally likely to be the best model before observing the data.

Our aim here is to assess, via simulation, how often BOAST analysis selects the correct number of latent variables, either one or more than one. We begin by simulating an individual participant analysis. We then examine a method of aggregating participant results to select the best characterization of dimensionality for a group of participants.

Simulations

Figure 1 shows state-trace data consistent with a single latent variable model (1D) and a two latent variable model (2D). In both cases the trace factor has a clear monotonic effect on performance; that is, as the level of the trace factor increases so too does the dependent variable. The two models also both exhibit moderate and equal data trace overlap. These two patterns were used to generate simulated data (by using their coordinates to specify binomial probability parameters) and we will refer to them as the 1D and 2D models.
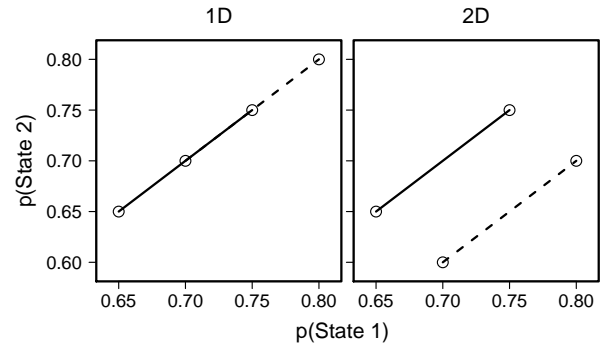

Figure 1: The two models on which simulations were based. p(State 1) and p(State 2) refer to the proportion of correct responses for the first and second level of the state factor, respectively. The two lines on each plot represent data traces, one for each level of the dimension factor. The solid lines are identical for both models, and the dashed line for the 2D model is the same as the dashed line for the 1D model but transposed downward by 0.1.

We next elaborated the 1D and 2D models with 2 trace levels shown in Figure 1, which we call the $T2$ designs, by creating variants with three and four trace levels, $T3$ and $T4$ designs respectively. In $T2$ designs the two levels of the trace factor provided data for the end points of the data traces. For $T3$ and $T4$ designs the additional levels were evenly spaced between the end points of each data trace. One purpose of these simulations was to provide guidance on experimental design in terms of the trade-off between number of trials contributing to the estimates of each point in the state-trace plot and the number of levels in the trace factor. For a fixed sample size (number of trials) there is a trade-off between these two factors, with more trace levels resulting in fewer trials

per point. For each model and each T we explored 6 total trial numbers (n) with total n conserved across each T at 192, 384, 768, 1536, 3072 and 6144. For example, a model with $n = 192$ had 24 observations per coordinate of each point for $T2$, 16 observations for $T3$, and 12 observations for $T4$. In total we performed 36 simulations (2 models $\times$ 3 trace levels $\times$ 6 sample sizes). For each simulation 1000 Monte Carlo replicates were sampled from binomial distributions with parameters determined by the design and model. Sufficient posterior samples were obtained so that posterior proportions of monotonic samples had 90% credible intervals less than 0.025; prior proportions were determined analytically assuming a uniform prior (see Heathcote et al., submitted, for details).

BOAST Results

For each simulation we estimated Bayes Factors to test four mutually exclusive models, which we refer to as the non-trace (NT), no-overlap (NO), unidimensional (UD) and multidimensional (MD) models. Together these models account for all possible orders (i.e., together they constitute the encompassing model). Posterior model probabilities were calculated for each Monte Carlo replicate for each model by dividing each Bayes Factor by the sum of all four Bayes Factors (i.e., Equation 2), which we refer to as p(NT), p(NO), p(UD) and p(MD), respectively. Figure 2 illustrates results in terms of the proportion of comparisons selecting one of the four models (i.e., where the models posterior probability was greatest amongst the set of four models). Figure 2 can be interpreted by comparing the height of corresponding points across the panels in each row. In particular, the 'highest' point indicates which of the four models is most often supported.

The Trace Model

An important first check in any state-trace analysis is to determine whether the trace model is supported. For example, we described study duration as a possible trace factor. In this case the trace model indicates that accuracy increased as study durations became longer for both levels of the state and dimension factors. In contrast, support for the non-trace model indicates that the order dictated by the trace factor was violated. Even when the trace model is the data generating model, measurement noise can cause violations of the trace model (i.e., support for the non-trace model) to arise more frequently when differences between levels of the study duration factor produce only small changes in accuracy. Support for the non-trace model clouds any conclusions about underlying dimensionality of the state factor since the effects of the dimension and trace factors are confounded, and can suggest that the experimental design needs to be improved by using more widely spaced trace factor levels.

The non-trace model results are shown in the left column of Figure 2. The figure demonstrates a number of key points. As expected, evidence for the trace model is similar across both 1D and 2D simulations, since the trace factor should

have a consistent effect irrespective of underlying dimensionality. Secondly, as total sample size increases the lines always approach zero, indicating consistent selection of the trace model. That is, BOAST recovers the trace model with increasing reliability as measurement error decreases due to an increase in sample size. Finally, the probability of selecting the non-trace model approached zero with lower total trials for $T2$ compared to $T3$ and $T4$. As seen in Figure 2, selection is approximately zero for $T2$ at $n = 768$, whereas this increased to $n = 1536$ for $T3$ and $T4$ in the 1D model, and to $n = 3072$ for $T4$ in the 2D model. Thus, for smaller n, the trace model had a greater chance of being supported in $T2$ designs compared to $T3$ and $T4$ designs. This occurs because the combination of a smaller sample size (and hence greater measurement noise) and closer spacing between results for adjacent trace levels as T increases makes a violation of monotonicity within a data trace more likely.

The No-Overlap Model

When the trace model holds it implies that one of the three remaining models best describes the data, as they are each trace models. A monotonic state-trace plot is a special case of the trace model where all data points have the same ordering for both levels of the state factor. A non-overlapping monotonic plot is a case where data traces for both levels of the dimension factor do not cross over at any point along either axis of the state-trace plot. In this case, monotonicity is not diagnostic of dimensionality, as both one-dimensional and multi-dimensional data generating models produce monotonic state-trace plots when there is a failure of data trace overlap. Hence, an important second check in a state-trace analysis is to determine whether the no-overlap model holds.

Results for the no-overlap model differed between the 1D and 2D data generating models. The 1D simulations results generally give some support for the no-overlap model, which is perhaps not surprising given the the 1D model produces monotonic data. Of more concern is the fact that this support was inconsistent as a function of sample size, n, for $T4$ and to a lesser degree for $T3$. That is, support for the no-overlap model initially increased with n, but then decreased, from $n = 1536$ for the $T4$ design and from $n = 768$ for the $T3$ design. In contrast, the no-overlap model consistently received little support across all T and n in the 2D simulations. Overall, these results suggest that when there is in fact trace overlap in a one-dimensional data generating model, the no-overlap model is more often rejected in designs with fewer trace levels.

The Unidimensional and Multidimensional Models

For both data generating models the unidimensional and multidimensional posterior model probabilities provided support for the true model dimensionality. For the 1D case support for the unidimensional model (middle right column of Figure 2) increased with sample size, but the level of support was smaller for larger T. For the 2D case support for the multidimensional model (right column of Figure 2) also in-
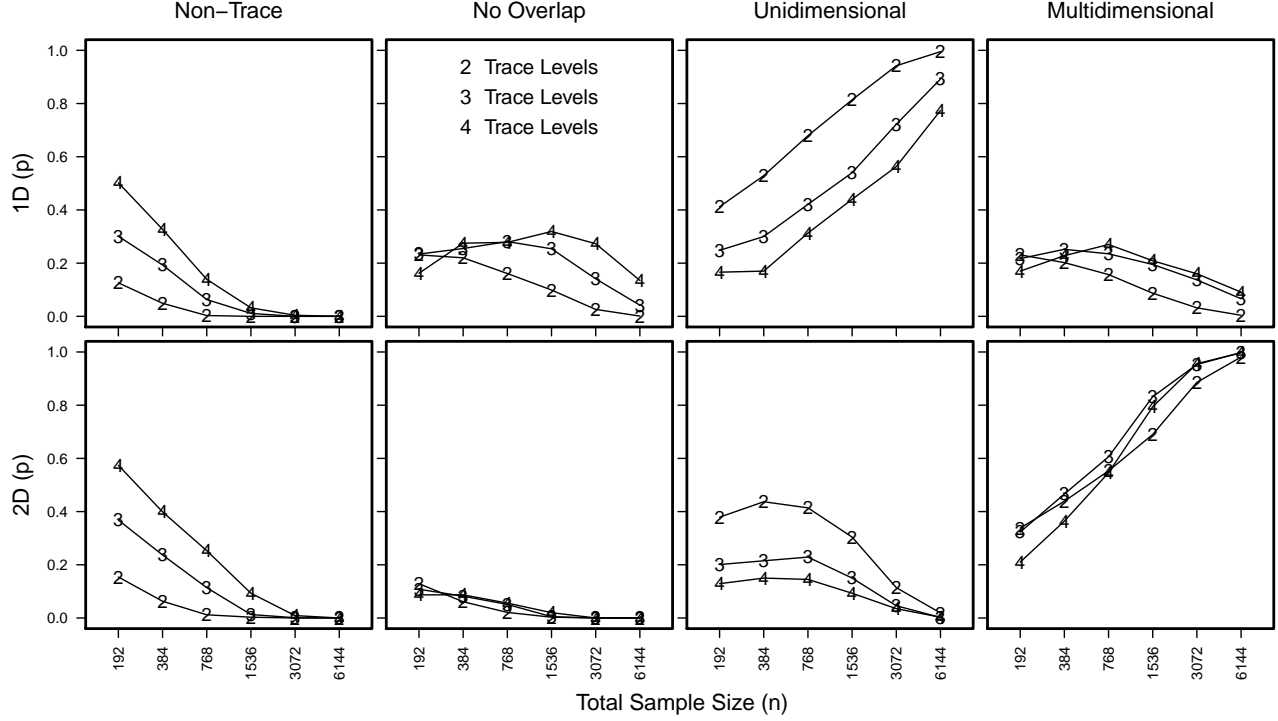
Figure 2: Model selection results for both data generating models, type of comparison, number of trace levels, T, and number of 'trials', n. Columns correspond to each of the mutually exclusive models being tested and rows to the type of the data generating model. On each plot the x axis represents the six levels of n and the y axis represents the proportion of simulations in which posterior model probability favored the model specified for each column. The lines group designs with the same T.

creased with sample size. In contrast to the 1D case, the level of support was similar for all T, although it was slightly less for the $T4$ design for smaller n (likely reflecting the larger level of support for the non-trace model) and slightly less for the $T2$ design for the second and third largest value of n, with all T designs perfectly selecting the true model for the largest sample size. Across the 1D and 2D data generating models support for the wrong dimensionality was generally low and decreased with sample size, although there was some inconsistency for the three smallest sample sizes.

Overall, the results of the simulation study indicate that accurate results for all comparisons can only be guaranteed for quite large sample sizes. This indicates that analysis of individual participant data may not produce clear results in applications where it is not possible to measure performance on a large number of trials for each individual. In such situations it would be desirable to have a method of combining results over participants in a way that improves correct identification at the group level. In the next section we extend the analysis of our simulation results to assess the performance of one such method suggested by Heathcote et al. (submitted), the group Bayes Factor.

Group Bayes Factors

A Bayes Factor for a group of participants, assuming each participant contributes independent evidence, can be obtained by taking the product of each participants Bayes Factor. Hence, a group Bayes Factor for a model M_i is given by $GBF_i = \prod_{j=1}^{N} BF_{ij}$, where N is the number of subjects. Group Bayes Factors can then be combined to obtain a posterior model probability for model M_i at the group level. Again we assume each model is equally likely to be the best model before observing the data, and so:

$$p(M_i|D) = \frac{GBF_i}{\sum_{k=1}^{m} GBF_k} \qquad (3)$$

for a set of $k = 1 \ldots m$ models that includes model i.

We examined the utility of group Bayes Factors using the simulations from the previous section. For each simulation we sampled with replacement (i.e., resampled) sets of individual Bayes Factors from the 1000 available. The sets were of sizes (N) 8, 16 and 32, representing experiments with different numbers of participants. These N's cross with total trials n in a balanced manner. For example, a set of $N = 32$ with $n = 192$ trials provides results from a total of 6144 trials, equivalent to the set $N = 16$ with $n = 384$ trials, and $N = 8$ with $n = 768$ trials. The resampling procedure was repeated 500 times for each possible grouping: two data generating models (1D, 2D), with three trace levels ($T2$, $T3$, $T4$), three total trial sizes ($n = 192$, 384, 768), and three participant sample sizes ($N = 8$, 16, 32), for each of the four comparisons (non-trace, no-overlap, unidimensional, multidimensional), a
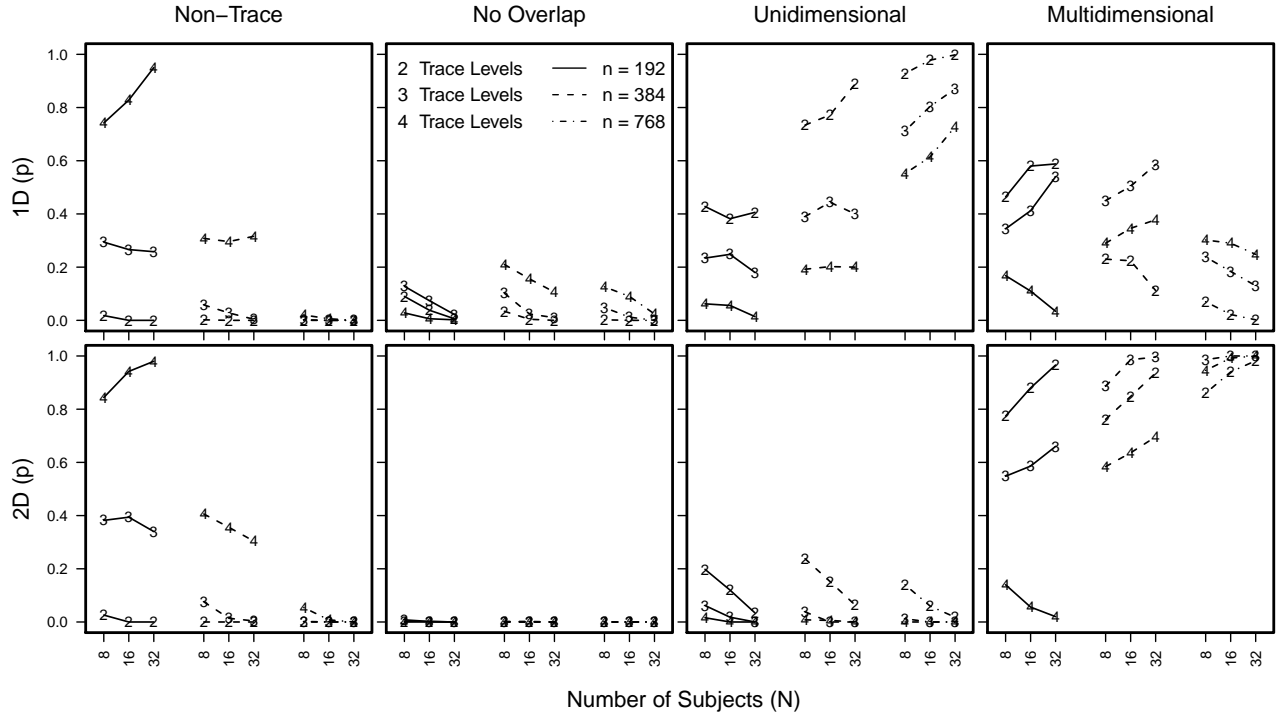
Figure 3: Group level results for the 216 comparisons. The rows and columns represent the same data generating models and comparisons as Figure 2. On each plot, the x axis represents the three levels of N that were resampled for each of $n = 192$, 384, 768, and the y axis represents the proportion of cases in which the posterior model probability at the group level favored the model specified for each column.

total of 216 combinations ($3^3 \times 2$ models $\times$ 4 comparisons). For each of the 500 repetitions of the 216 combinations we estimated group Bayes Factors, and then calculated the proportion of comparisons selecting one of the four models (i.e., where the models posterior probability was greatest amongst the set of four models), with results shown in Figure 3.

For the no-overlap model the group Bayes Factors results were much the same as for the individual analysis, except that the inconsistent effect of sample size for the individual analysis of the 1D data generating model disappeared in the group analysis. For the trace model performance was excellent when $n = 768$ but the wrong (non-trace) model received increasing support when there were fewer observations per participant for all but the $T2$ design. These problems with the trace model caused corresponding failures to identify the correct dimensionality for lower values of n, whereas for $n = 768$ performance in identifying dimensionality was similar to that of the largest samples sizes in the individual analysis. In particular, the 2D data generating model was almost perfectly identified, but with higher T designs being slightly better, whereas performance in classifying the 1D data generating model was very good for $T2$ designs but decreased markedly for the $T3$ and $T4$ designs.

Conclusions

We aimed to investigate the ability of BOAST analysis to identify latent dimensionality. The results of individual par-

ticipant data indicated that large sample sizes produced strong support for the correct outcome for both 1D and 2D data generating models across designs with two, three and four levels in the trace factor. Classification for the 1D data generating model was most reliable in designs with two trace levels, whereas the opposite tendency was evident for the 2D data generating model; dimensionality assessment was more accurate with larger numbers of trace levels. Overall these results indicate that a design with three trace levels provides the best compromise for accurate diagnosis of both single and multiple latent variable data generating models.

We also explored a group analysis procedure that is advantageous where it is practically difficult to obtain a large number of responses from each individual participant, such as in cases where the number of available stimuli is limited, but where larger numbers of participants are available. Generally, this method was found to be very effective in identifying the 2D data generating model. However, our results indicate that it should be used with caution as it could be biased against detecting cases in which only one latent variable is present in certain experimental designs. When each participant contributed a smaller number of responses (192 or 384) results could be inaccurate even for the largest number of participants (32). For 768 observations per participant performance was more accurate and improved with group size for the 1D

data generating model. In contrast to the individual participant results, the group level analyses indicate that designs with two levels in the trace factor produce the best compromise of most accurate classification across number of trials per participant and different numbers of participants for both 1D and 2D data generating models. However, these results should be used with some caution given the three and four trace level designs demonstrated a large proportion of cases supporting the non-trace model (possibly due to the small experimental effects of the trace factor in these larger trace level designs), which had strong consequences for the correct classification of dimensionality.

Our individual and group analyses indicate that the ideal number of trace levels in a state-trace experiment is dependent on the intended approach to data collection. If only a small number of trials per participant are obtainable it seems wiser to use a trace factor with few levels so as to maximise data per point, and then combine across participants with group Bayes Factors. In contrast, if many trials per participant can be obtained, correct classification of dimensionality is possible with a three level trace factor through individual participant analysis, which confers additional benefits such as the exploration of individual differences in performance.

In summary, these results indicate that the success of BOAST analysis, and likely any state-trace analysis method, depends strongly on the particular model producing the state-trace plot. This highlights a caveat on our group analysis, which assumes all participants have an identical underlying model (rather than just having the same dimensionality but possibly different magnitudes of the effects of experimental factors). As well as being unrealistic, this assumption likely magnifies the effects of a particular data pattern. In ongoing research we will simulate groups of participants that vary in the effects of experimental factors (while maintaining a consistent dimensionality) in order to check the generality of the group analysis results reported here.

References

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

Dunn, J. C., & Kirsner, K. (1988). Discovering functionally independent mental processes: The principle of reversed association. *Psychological Review*, *95*, 91–101.

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

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.

Loftus, G. R. (1978). On interpretation of interactions. *Memory & Cognition*, *6*, 312–319.

Loftus, G. R., Oberg, M. A., & Dillon, A. M. (2004). Linear theory, dimensional theory, and the face–inversion effect. *Psychological Review*, *111*, 835–863.

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

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

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

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology*, *81*, 141–145.