Journal of Mathematical Psychology 72 (2016) 116–129

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

Bayes factors for state-trace analysis

Clintin P. Davis-Stober [a,*], Richard D. Morey [b], Matthew Gretton [c], Andrew Heathcote [c]

[a] *University of Missouri, United States*
[b] *Cardiff University, United Kingdom*
[c] *University of Tasmania and University of Newcastle, Australia*

HIGHLIGHTS

- We develop a new across-subjects Bayesian state-trace analysis.
- We present an improved computational method for Bayesian state-trace analysis.
- We apply our new methods to an existing data set.

ARTICLE INFO

Article history:
Available online 21 October 2015

Keywords:
Monotonicity
State-trace
Convex hulls
Order-constrained inference

ABSTRACT

State-trace methods have recently been advocated for exploring the latent dimensionality of psychological processes. These methods rely on assessing the monotonicity of a set of responses embedded within a state-space. Prince et al. (2012) proposed Bayes factors for state-trace analysis, allowing the assessment of the evidence for monotonicity within individuals. Under the assumption that the population is homogeneous, these Bayes factors can be combined across participants to produce a "group" Bayes factor comparing the monotone hypothesis to the non-monotone hypothesis. However, combining information across individuals without assuming homogeneity is problematic due to the nonparametric nature of state-trace analysis. We introduce group-level Bayes factors that can be used to assess the evidence that the population is homogeneous vs. heterogeneous, and demonstrate their utility using data from a visual change-detection task. Additionally, we describe new computational methods for rapidly computing individual-level Bayes factors.

1. State-trace analysis

In this paper we discuss Bayes factors that address the question of whether one psychological (i.e., latent or not directly observed) parameter, or more than one psychological parameter, mediates the interacting effects of two experimental manipulations. In particular, we develop Bayes factors to address the question of "latent dimensionality" using methods first proposed by Bamber (1979): state-trace analysis. On the assumption that there is a monotonic mapping from the latent variable(s) to the manifest (i.e., directly observable) dependent variable(s), state-trace analysis can identify a one dimensional system, that is, a system producing the observed behavior by variation of a single parameter. The signature of such systems is that they produce a monotonic relationship between two different dependent variables, or between the same dependent variable measured under two different conditions. The Bayes factors that we develop to address the question of monotonicity, and hence the dimensionality, of binary dependent variables, are based on the work of Klugkist, Laudy, and Hoijtink (2005) as applied to state-trace analysis by Prince, Brown, and Heathcote (2012). The aim of Prince et al.'s approach is to preserve the essentially non-parametric nature of state-trace analysis by making only fairly minimal statistical assumptions, the main one being that the dependent variable has a binomial distribution.

Prince, Hawkins, Love, and Heathcote's (2012) approach primarily focused on separate analyses of each participant's data, as they showed that state-trace analysis could be invalid when applied to data averaged over participants. In particular, they provided examples where the average of two monotonic relationships is non-monotonic. For inference at the group level, they suggested taking the product of individual participant's Bayes factors, but acknowledged the weakness in two necessary underlying assumptions: (1) that participants are either all of one type (e.g., monotonic) or another (e.g., non-monotonic) and (2) that the participants are entirely unrelated. At first sight, hierarchical

[*] Corresponding author.
E-mail address: cstober2@gmail.com (C.P. Davis-Stober).

http://dx.doi.org/10.1016/j.jmp.2015.08.004
0022-2496/© 2015 Elsevier Inc. All rights reserved.

modeling seems an obvious remedy to this weakness, but as we discuss in the second part of this paper it is far from obvious how to correctly specify such models without making strong parametric assumptions. The main problem is finding the correct space on which to hierachicalize; choosing the wrong space leads to the same difficulties as directly averaging the data.

Our contribution is twofold. We first summarize and extend Prince, Brown, & Heathcote's (2012) approach, and develop a new method for computing their Bayes factors that is much faster than Klugkist et al.'s (2005) methods as implemented by Prince, Hawkins et al. (2012). This methodology is based upon Laplace's approximation (Stigler, 1986) and efficient numerical algorithms (Genz, 1992). Second, we present an alternative group-level Bayes factor approach that tests whether all of the participants in a group have the same dimensionality. This method builds upon recent advances in testing mixture models of individual decision-making (e.g., Regenwetter, Dana, & Davis-Stober, 2011). We demonstrate how this group-level Bayes factor partially alleviates the conceptual difficulties identified by Prince et al. when averaging data across individuals. We then show how this new "aggregated Bayes factor" test, when applied in conjunction with the group-level and individual tests of Prince et al., can provide a comprehensive state-trace analysis.

Throughout the paper we use as an example data from Sense, Morey, Prince, Heathcote, and Morey (in preparation) examining the encoding dimensions that underpin short-term memory. Sense et al. sought to assess the evidence that participants make use of both a visual short-term memory store and an auditory short-term memory store when performing a visual change detection task with arrays of colored squares. That is, participants were shown set of squares at an array of different positions, then shortly after they were shown another set and had to indicate if the second set is the same as the first or whether it has changed. If visual short-term memory is limited, then it may sometimes be advantageous to verbally code visual stimuli (e.g., Murray, 1965; Schooler & Engstler-Schooler, 1990). If verbal recoding occurs, then tasks designed to study visual short-term memory do not solely index visual capacity.

Suppose a researcher attempted to manipulate the involvement of a hypothesized latent auditory dimension in two ways, by (1) presenting the stimuli in the set either sequentially (allowing time for verbal recoding) or simultaneously (making verbal recoding harder because less time is available), and (2) by requiring concurrent articulation during study (making verbal recoding harder) or allowing silent study (making verbal recoding easier). Fig. 1 illustrates data from this design using three different visual array set sizes (number of items to be remembered). Typically, a researcher would use dissociations between the effects of such manipulations – usually operationalized by an ANOVA interaction test – to infer a role for more than one latent variable (e.g., Shallice, 1988). For example, the articulation manipulation may have a lesser impact with simultaneous presentation (where there is little verbal recoding anyway) than with sequential presentation (where articulation reduces the boost in performance from verbal recoding).

Many authors (e.g., Bogartz, 1976; Dunn & Kirsner, 1988; Henson, 2006; Loftus, 1978) have observed that this dissociation logic is flawed, as interactions in the manifest variables can be caused, or masked, by scale effects in dependent variables (e.g., floor or ceiling effects). For example, a ceiling effect could mask a super-additive interaction at the latent level, making it appear as an additive effect in accuracy. This could be the case for set size 2 in Fig. 1 if the silent advantage was larger for simultaneous than sequential displays at the latent level but performance nearer ceiling in the simultaneous condition attenuated that advantage in accuracy relative to the less accurate sequential condition. Alternately, a ceiling effect might create a sub-additive interaction in accuracy even when additivity holds at the latent level.

Although researchers can sometimes take experimental precautions to avoid ceiling and floor effects associated with an ogival mapping between latent and manifest levels, Prince, Brown, and Heathcote (2012) pointed out that little can be done to guard against complicated (but still monotonic) mappings involving more than one region of maximum sensitivity (i.e., rate of change of accuracy per unit change at the latent level). For example, suppose sensitivity was maximal at 65% and 85% accuracy, assuming the usual floor at 50% and ceiling at 100%. Differences in the less sensitive 75% midrange region would be attenuated relative to differences around 85%, even though the latter difference is nearer to ceiling. When ANOVA-interaction testing is carried out by classical methods a further problem occurs; it is impossible to infer a lack of interaction – and hence evidence for the simpler, one-dimensional model – leading to an inherent bias towards ever more complex models.

State-trace analysis provides a solution to these problems no matter how complex the mapping between latent and manifest variables, as long as it is monotonic. Developments of state-trace analysis since Bamber's (1979) seminal work have focused on two areas. Work by Loftus and colleagues built on the insights of Loftus (1978) about the key role of monotonicity in understanding the limits of what can be inferred about latent variables from observed interactions (e.g., Busey, Tunnicliff, Loftus, & Loftus, 2000; Loftus & Irwin, 1998; Loftus, Oberg, & Dillon, 2004). Work by Dunn and colleagues focused on the implications of both Bamber (1979) and Loftus (1978) for dissociation logic, concluding that neither single nor double dissociations provide a necessary or sufficient basis for making inferences about latent dimensionality (Dunn, 2008; Dunn & Kirsner, 1988, 2003; Newell, Dunn, & Kalish, 2010). Newell and Dunn (2008) provided an overview of these developments and outlined areas where further progress is needed. Prominent among these areas was the need for statistical methods of quantifying evidence about monotonicity.

Prince, Brown, and Heathcote (2012) advocated the Bayesian approach to state-trace analysis, which we also pursue here, for two reasons. First, they argued that appropriate statistical methods should make minimal parametric assumptions, so that the fundamentally non-parametric nature of state-trace analysis is not compromised. This maintains one of the most useful aspects of state-trace analysis; it is an ideal complement to parametric modeling because it provides guidance not dependent on specific modeling assumptions about the number of parameters that should be allowed to vary as a function of experimental manipulations (for an example of this interplay see Heathcote, Bora, & Freeman, 2010; Heathcote, Freeman, Etherington, Tonkin, & Bora, 2009). A second key advantage of state-trace analysis is that it helps identify situations where complicated patterns of observed results, such as the potentially different pattern of interactions as a function of set size illustrated in Fig. 1, can be explained by a simpler underlying latent structure. This advantage relies on being able to quantify evidence for a null hypothesis – in the context of state-trace analysis, the monotonic hypothesis indicating a one-dimensional system – so null-hypothesis statistical testing is unsuitable.

Bayesian methods do require the additional assumption of a prior, but as we illustrate further below, this also provides an opportunity to take account statistically of pre-existing experimental and/or theoretical knowledge that is, in any case, essential in designing a successful state-trace experiment. In the next section, we describe Prince, Hawkins et al.'s (2012) approach to the design and analysis of state-trace experiments. We then present a computationally faster and more efficient method of computing their Bayes factors. In the following section we present a new Bayes-factor approach designed to evaluate evidence for monotonicity at the group level. We conclude with a summary and discussion.

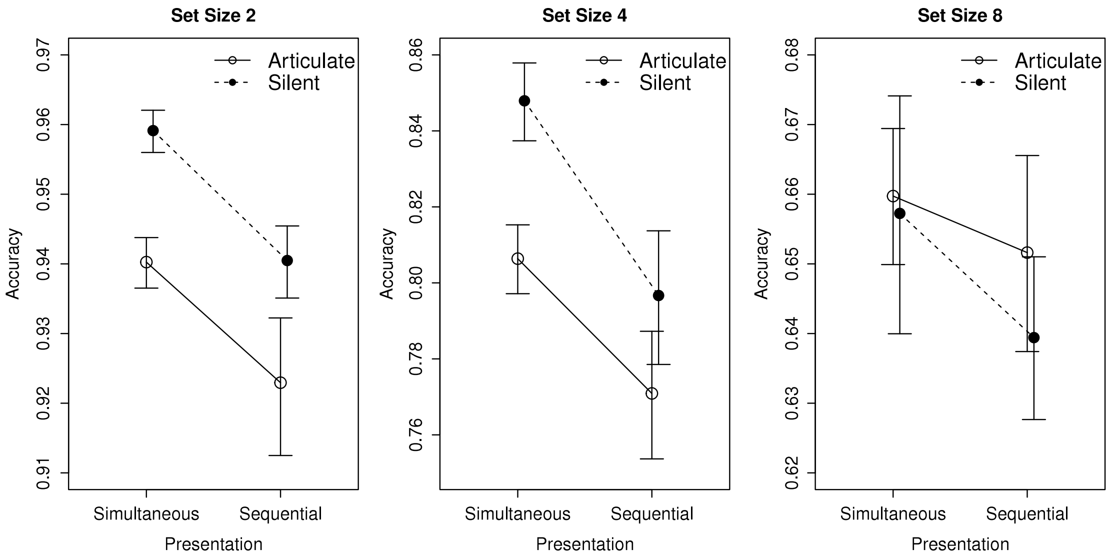

Fig. 1. Accuracy averaged over 15 participants from Sense et al. (in preparation). Accuracy for each participant was calculated assuming a uniform prior (i.e., $(M+1)/(N+2)$, where M is the number of correct responses and N the total number of responses, Rouder & Lu, 2005) were transformed to a probit scale before averaging. Within-subject standard error bars were calculated as described in Morey (2008).

2. Analysis of state-trace data

State-trace analysis is accomplished by assessing the monotonicity of a plot of one set of dependent-variable values against another. In our example data set, we plotted accuracy for the sequential condition against accuracy for the simultaneous condition (see Fig. 2). We follow Prince, Brown, and Heathcote (2012) in naming the independent variable whose levels constitute the axes of the state-trace plot the "state" factor[1] (with levels $s \in \{1 \dots S\}$ where $S = 2$), as it defines a state-space within which the behavior of the system under study is quantified. State-trace analysis focuses on the interaction between the state factor and a second independent variable, which we follow Prince et al. in calling the dimension factor (e.g., the articulate vs. silent), with levels $d \in \{1 \dots D\}$. State and dimension roles can be interchangeable, but there are sometimes advantages for one assignment over the other (see Prince et al. for an extended discussion).

When the dimension factor has only two levels ($D = 2$) the resulting state-trace plot has only two points, and so cannot reveal a non-monotonic relationship. When $D > 2$ the state-trace plot is potentially diagnostic of dimensionality; alternately, a third independent variable can be introduced. In the case of Sense et al. (in preparation), the third variable was a manipulation of the number of studied items. Prince, Brown, and Heathcote (2012) called this third manipulation the "trace" factor (with levels $t \in \{1 \dots T\}$), as it traces out a trajectory in the state space within each level of the dimension factor. For a trace factor with T levels (e.g., $T = 3$ in our example data) the T corresponding points within each dimension level (e.g., silent or articulate conditions in our example data) are joined by lines called "data traces". Fig. 2 illustrates state-trace plots for two of Sense et al. (in preparation) 15 participants, each of who performed between 2000 and 2500 trials over 4–5 sessions. Given each participant performed over 150 trials in each of the 12 experimental conditions, individual accuracy estimates are fairly precise as indicated by the credible intervals in Fig. 2.

Monotonicity among the points in the state space is equivalent to them having the same order on each axis. For a data set constituted of observations, $P_{t,d,s}$, the state-trace is a plot of the ordered pairs $(P_{t,d,1}, P_{t,d,2})$. In Fig. 2, the x-axis plots the sequential ($s = 1$) level of the state factor and the y-axis the simultaneous ($s = 2$) level of the state factor. The first level of the dimension factor is the articulate condition ($d = 1$, solid line) and the second the silent condition ($d = 2$, dashed line). The points on each line correspond to the levels of the trace factor; sets-size 2 ($t = 2$), 4 ($t = 4$) and 8 ($t = 8$). Monotonicity holds under equivalence of the joint order $rank(P_{.,.,1}) \equiv rank(P_{.,.,2})$, where $rank()$ is a function that returns the rank of the elements of the vector $P_{.,.,s}$, where the "." indices range over $t \in \{1 \dots T\}$ and $d \in \{1 \dots D\}$ respectively. Monotonicity indicates that only one latent variable determines performance, so the analysis of latent dimensionality is based on the probability of sets of joint orders that constitute different ways in which monotonicity can be realized.

Fig. 2 provides an example where monotonicity appears to be violated and an example where it appears to hold. In order to refer to points on the graph we will use the subscripts $t \in \{2, 4, 8\}$ to refer to set sizes, $d \in \{a, s\}$ to refer to the articulate and silent conditions and $s \in \{sim, seq\}$ to refer to the simultaneous and sequential conditions. For participant "a" in the top left panel of Fig. 2 monotonicity appears to be violated. For simultaneous displays, accuracy is always highest in the articulate condition for all set sizes ($P_{.,a,sim} > P_{.,s,sim}$). For sequential displays, in contrast, accuracy is higher in the silent than articulate condition for set-sizes 2, ($P_{2,s,seq} > P_{2,a,seq}$), and 8, ($P_{8,s,seq} > P_{8,a,seq}$), but is reversed for set-size 4, ($P_{4,s,seq} < P_{4,a,seq}$). This violation of monotonicity suggests that more than one latent variable may determine performance in participant "a". For participant "o" in the bottom left panel of 2 monotonicity appears to hold as accuracy is higher for the silent than articulate condition at each display size in both types of displays, i.e., ($P_{.,s,.} > P_{.,a,.}$). Hence, for participant "o" one latent variable may be sufficient to explain performance.

In general, there are $Q = (D \times T)!$ possible orders and so Q^2 joint orders. Assuming a prior in which all orderings are equally likely – for instance, if the priors on all points are identical and conditionally independent – the prior probability of any joint ordering is $1/Q^2$. A "model" can be thought of as a set of possible orderings; for instance, the monotonic model is the set of all joint orderings that are the same on each axis. If K is the number of joint

[1] A state factor can also be constituted of two different dependent variables. For example, Busey et al. (2000) examined state-trace plots where one axis was the accuracy of recognition decisions and the other axis confidence ratings for those decisions.

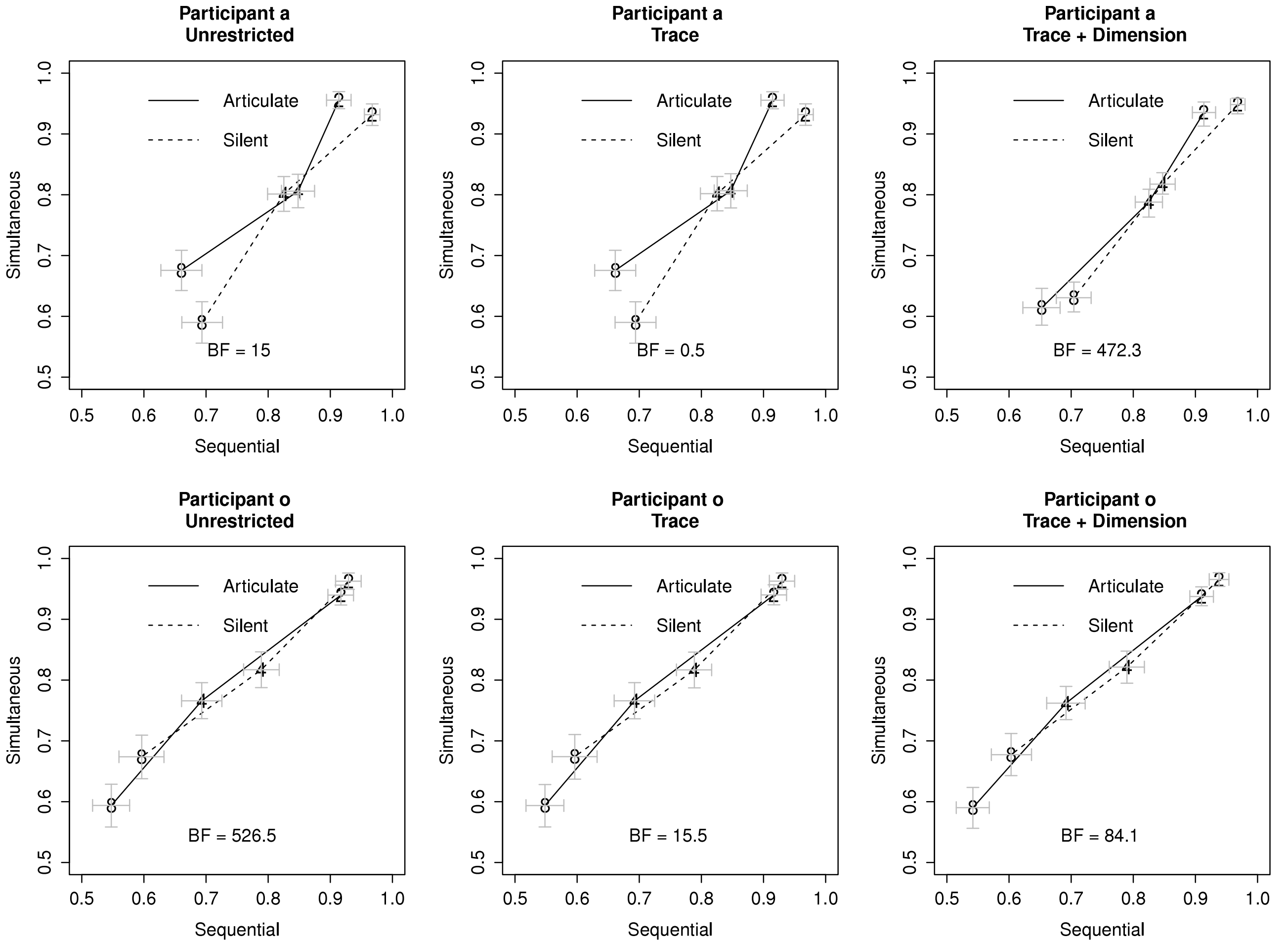

Fig. 2. State-trace plots of accuracy for two participants from Sense et al. (in preparation). The points labeled 2, 4 and 8 mark the mean accuracy for 10^4 posterior samples from the corresponding set-size condition (note that numbers always decrease from left to right, as the trace factor has a monotonic effect). Bars around the point indicate "standard" credible intervals (i.e., the 0.159 to 0.841 posterior quantiles). Sampling was performed on the probit scale then samples were transformed back to accuracy before averaging, and in all cases only samples with greater than chance accuracy were kept. For the left hand column no other restrictions on samples were enforced. For the middle column only samples following the trace-model order (i.e., increasing accuracy with decreasing set size) were retained. For the right hand column the dimension order constraint (i.e., greater accuracy for silent than articulate) was also imposed. Bayes factors (BF, with the monotonic model in the numerator and to non-monotonic model in the denominator, so $BF > 1$ supports the monotonic model) calculated under the corresponding restriction are displayed in each panel.

orderings that constitute a model then the prior probability π of that model is K/Q^2. In our example data set, there are $6! = 720$ monotonic orderings, and so the prior probability of the monotonic model is $\pi_M = 6!/(6!)^2 = 1/6! = 0.0014$. The prior probability of the complementary non-monotonic model is $\pi_{NM} = 1-0.0014 = 0.9986$.

The posterior probabilities, denoted $\pi^{(y)}$, can be computed determining the proportion of the posterior distribution that is consistent with each ordering. Unlike prior probabilities under the assumption that all orders are equally likely, these posterior probabilities cannot easily be obtained analytically. Simple Monte Carlo methods can be used, but in some cases they are impractical. Below we return to the issue of how posterior probabilities can be calculated, and suggest an improvement on past methods, but first we consider how inference about latent dimensionality can be performed once they have been obtained, by computing Bayes factors.

The Bayes factor between two models is simply the ratio of the posterior odds between the models to the corresponding prior odds; that is, the Bayes factor is the degree to which the data have changed one's preference between the two models. For the monotonic vs. non-monotonic models, for example:

$$BF_{M/NM} = \frac{\pi_M^{(y)}}{\pi_{NM}^{(y)}} \Bigg/ \frac{\pi_M}{\pi_{NM}}.$$

In Fig. 3, we plot $BF_{M/NM}$ (triangles) for each participant in our example data set. Clearly, for every participant there is strong evidence favoring monotonicity over non-monotonicity; even in the worst case, in order for the data not to change a belief in the non-monotonic, the prior preference for the non-monotonic model would have to be more than 25 times greater than the preference for the monotonic model.

However, $BF_{M/NM}$ is problematic when interest focuses on monotonicity associated with the interaction of state and dimension factors. This is because it conflates monotonicity associated with the interaction with the monotonic effect of the trace factor. The trace factor is often chosen because of prior knowledge that it has a monotonic effect at all levels of the state and dimension factors (e.g., accuracy decreases with increasing set size), and so it can be reliably used to sweep out the behavior of the system under study. Indeed, as shown in Fig. 3, a ratio of Bayes factors for the monotonic trace model (i.e., a model where all data traces are monotonic) to its complement (i.e., the "non-trace" model, a model where at least one data trace is non-monotonic), $BF_{T/NT}$, shows (as stars) that the example data very strongly support this prior belief in relation to set size.

Prince, Brown, and Heathcote (2012) proposed that a more appropriate approach is to use a Bayes factor that assumes the trace model is true, $BF_{(M/NM)|T}$. This Bayes factor is computed by considering only orders that conform to the trace model. This is critical for a "fair" test of the question of interest, a non-monotonic hypothesis with respect to the state and dimension factors; a theorist might maintain that this non-monotonic model holds, but simultaneously believe that the trace model should not be violated. In the case of the example data, no theorist would believe that performance would *increase* when the number of to-be-remembered items increases. If the non-monotonic model includes the non-trace models, however, the Bayes factor for the non-monotonic model will be unfairly penalized for being a complex model that no one would endorse.

Prince, Brown, and Heathcote (2012) also pointed out that when the trace model is assumed to be true two particular monotonic orders are not diagnostic for the question of interest. These orders occur when, for both levels of the state factor, values from one level of the dimension factor all fall below or all fall above values from the other level of the dimension factor. Graphically, these two non-diagnostic orders correspond to data traces that do not overlap on either axis. Non-overlapping data traces are non-diagnostic because they must always result in monotonicity, even when extrapolation of the true data traces would result in a non-monotonic plot. That is, the trace manipulation has not been sufficiently influential, and so we are effectively back to the two-point case where it is not possible to violate monotonicity.

In summary, Prince, Brown, & Heathcote's (2012) method requires three partitions of trace-model orders, into non-monotonic (NM), monotonic non-overlapping (NO) and monotonic-overlapping (MO) sets. Prior probabilities for each set can be obtained analytically (see Prince, Brown, & Heathcote, 2012, for details); in our example data set they are $\pi_{NM} = 7.33 \times 10^{-4}$, $\pi_{NO} = 3.85 \times 10^{-6}$ and $\pi_{MO} = 3.47 \times 10^{-5}$. Once corresponding posterior probabilities are estimated, Bayes factors can be obtained and evidence for monotonic vs. non-monotonic models conditional on the trace model can be computed.

Unfortunately, simple Monte Carlo methods are usually impractical for computing the posterior probabilities necessary to obtain $BF_{(M/NM)|T}$. To overcome this problem Prince, Brown, and Heathcote (2012) used an order-constrained Gibbs sampler (Gelfand, Smith, & Lee, 1992). Prince, Hawkins et al. (2012) provided an R package (StateTrace) that implemented this approach and managed sampling to automatically obtain estimates with a specified accuracy. Even so, this approach can be time consuming. Instead, we suggest the use of Laplace's method (Stigler, 1986), which enables posterior probability estimates to be obtained through numerical integration.

Laplace's method assumes a multivariate normal approximation to the posterior. Orderings of parameters correspond to integrals over regions of the multivariate normal, which can be computed using the algorithm outlined by Genz (1992). Rather than directly using the observed accuracies, it is beneficial to apply a probit (i.e., inverse cumulative normal) transform in order to yield better results where accuracies are close to floor or ceiling, with the multivariate normal approximation applied to the posteriors on the probit space. We have found that, except for very small sample sizes, the combination of the multivariate normal approximation on the probit space and Genz's algorithm for integration yields estimates in a tiny fraction of the time needed for sampling, and that these estimates correspond very well to the estimates obtained by sampling. R code to implement this analysis for the example data can be obtained is available in an online supplement (see Appendix B). This code is written so that it can be easily modified for other designs with any number of dimension-factor and trace-factor levels.

As Fig. 3 shows (open circles), $BF_{(M/NM)|T}$ is substantially reduced relative to its unconditional counterpart, although in most cases $BF_{(M/NM)|T}$ is greater than one, indicating the data support monotonicity. The horizontal dotted lines in the figure indicate conventional values are sometimes used to classify "positive" (>3) and "strong" (>20) and "very strong" (>150) evidence (Raftery, 1995). Participants in Fig. 3 are sorted in order of increasing $BF_{(M/NM)|T}$, and Fig. 2 provides state-trace plots for the participants with the weakest ("a") and strongest ("o") evidence for monotonicity according to $BF_{(M/NM)|T}$. The middle column of Fig. 2 plots estimates under the assumption of the trace-model prior, whereas the left hand column plots estimates under the assumption that all orders are equally likely. The two sets of estimates are very similar as the data conform quite closely to the trace model. Stronger support of monotonicity for participant "o" is evident in the data points being able to be joined by an always increasing line. Weaker support of monotonicity for participant "a" is evident in reversals in the increasing trend for set sizes of 8 and 2.

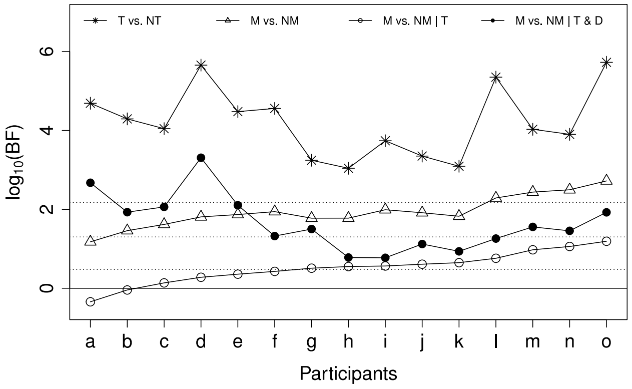

Fig. 3. Bayes factors calculated assuming greater than chance accuracy, for: (a) the Trace (T) model (a model under which points consistently increase or consistently decrease with the trace factor) vs. the Non-trace (NT) model (T vs. NT), (b) Monotonic (M) vs. Non-Monotonic (NM) model with no other order restrictions (M vs. NM), (c) the Monotonic vs. Non-Monotonic model with trace (i.e., accuracy decreases as set size increases) order restriction (M vs. $NM|T$) and (d) the Monotonic vs. Non-Monotonic model with the trace and dimension (i.e., greater accuracy in silent than articulate) order restrictions (M vs. NM | T & D). Horizontal dotted lines are placed at $\log_{10}(3)$, $\log_{10}(20)$ and $\log_{10}(150)$ corresponding to Raftery's (1995) boundaries between positive, strong and very strong evidence conventions.

Prince, Brown, & Heathcote's (2012) approach can be extended by adding further restrictions based on prior beliefs and knowledge. We explored imposing two constraints. First, true performance should always be above chance. We believe this is a reasonable restriction. The only way to obtain true performance below chance is if a participant flips their responses (i.e., consistently responds "same" when they intend "change"). Since performance is largely above chance, we see no reason to believe this occurred. This restriction does not substantially change the Bayes factors in this application, but it is possible that in other applications it might, and the reasonableness of the restriction would have to be considered.

A second restriction we considered was on the dimension factor: that true accuracy must be better in a silent condition than in the corresponding articulate condition. We believe this is another reasonable restriction, based on the fact that it is almost always found that dual-task performance is worse than single-task performance (Guérard & Tremblay, 2008; Jones, Farrand, Stuart, & Morris, 1995; Meiser & Klauer, 1999). In the case of the specific hypothesis to be tested in the example data, this restriction should be appealing to both those who might advocate a role for verbal recoding and those that do not. Advocates of verbal recoding would argue that articulation should *selectively* hurt performance in one condition, not help. Advocates of both theoretical positions might also suggest a slight dual task cost.

Indeed, Fig. 1 shows that there appears to be a dual task cost, on average, for our example data, although Fig. 2 shows that this was not the case for every participant. For example, participant "a" shows a number of reversals (e.g., for set size 8 in the sequential condition and set size 2 in the simultaneous condition). Imposing what we will call the "dimension" restriction corresponds to treating these reversals as measurement error. The right hand column of Fig. 2 shows estimates when both the trace and dimension restrictions are assumed, and Fig. 3 shows that the corresponding Bayes factors for monotonicity, $BF_{(M/NM)|T\&D}$ (plotted as solid circles) are consistently increased relative to Bayes factors with only the trace restriction $BF_{(M/NM)|T}$. We would argue that $BF_{(M/NM)|T\&D}$ provides the fairest assessment of the evidence for one underlying dimension provided by this data, due to the fact that the posited dual-task cost would be expected under either theoretical position being compared.

At the individual participant level, $BF_{(M/NM)|T\&D}$ provides a range of strengths of evidence, changing prior beliefs by factors ranging from about 6 to about 2000. That said, the evidence is very consistent, favoring monotonicity for every participant. Prince, Brown, and Heathcote (2012) suggested using the product of individual Bayes factors, which they called group Bayes factors (GBF) to quantify evidence at the group level. For the example data, $\log_{10}(GBF_{(M/NM)|T}) = 7.7$, and $\log_{10}(GBF_{(M/NM)|T\&D}) = 24.7$, both indicating very strong evidence in favor of all members of the group being monotonic relative to all members of the group being non-monotonic.

The GBF rests on two assumptions: first, that the information from one participant does not inform us about what we would believe about another participant, apart from the Bayes factor. Second, it assumes that participants are homogeneous, with all participants being either monotone or non-monotone. This assumption of group homogeneity seems reasonable in light of the consistency of the individual participant results in Fig. 3. However, in a simulation study, Hawkins, Prince, Brown, and Heathcote (2010) found increasing problems as sample size per participant decreases, with the GBF varying widely even when the group is homogeneous. Further, it is undesirable to decide which Bayes factors to compute on the basis of the data itself. It seems reasonable, a priori, to test whether the participants are homogeneous or heterogeneous, and so it would be desirable to compute a Bayes factor to test the homogeneity assumption before placing confidence in the strong result indicated by this the group Bayes factor.

3. Aggregated group-level Bayes factors for state-trace analysis

We present a new method for computing a group-level Bayes factor that evaluates whether all individuals satisfy monotonicity vs. the hypothesis that at least one individual does not. This methodology utilizes data averaged across all individuals in the group. Let J be the total number of participants in the experiment. Let $Z = T \times D \times S$ be the total number of experimental conditions. For the Sense et al. (in preparation) data we have three levels of the trace factor, two levels of the dimension factor and two levels of the state factor, yielding $Z = 3 \times 2 \times 2 = 12$ conditions. Let $\boldsymbol{M}$ be the vector of size $Z \times 1$ defined as $\boldsymbol{M}_i = \sum_{l=1}^{J} m_i^l, i \in \{1, 2, \ldots, Z\}$, where m_i^l is the l^{th} participant's number of correct responses in experimental condition i. Likewise, let $\boldsymbol{N}$ be the vector of size $Z \times 1$ defined as $\boldsymbol{N}_i = \sum_{l=1}^{J} n_i^l, i \in \{1, 2, \ldots, Z\}$, where n_i^l is the l^{th} participant's number of completed trials in experimental condition i. Let $\boldsymbol{P}$ be the $Z \times 1$ vector defined as $\boldsymbol{P}_i = \frac{\boldsymbol{M}_i}{\boldsymbol{N}_i}, i \in \{1, 2, \ldots, Z\}$, i.e., the average proportion correct for the Z-many experimental conditions.

Similar to Prince, Brown, and Heathcote (2012), we model responses within each experimental condition as a binomial random variable. What is different for our analysis is that each binomial random variable corresponds to the *average* response, at the group level, for the corresponding experimental condition. Thus, each probability of success parameter corresponds to the average probability of a correct response. Similar to Prince et al., we also assume that all responses across experimental conditions are independent. These assumptions give the following joint likelihood over all experimental conditions:

$$L(\boldsymbol{\theta}|\boldsymbol{M}) = \prod_{i=1}^{Z} \theta_i^{N_i}(1-\theta_i)^{N_i - M_i}, \quad \theta_i \in (0, 1), \ \forall i \in \{1, 2, \ldots, Z\}. \tag{1}$$

Let $\boldsymbol{\theta} = \theta_i, i = \{1, 2, \ldots, Z\}$, and let $\Theta = (0, 1)^Z$. Note that for each experimental condition i, $\boldsymbol{P}_i$ is the maximum likelihood estimate for θ_i.

The $\boldsymbol{M}$, $\boldsymbol{N}$, and $\boldsymbol{P}$ vectors are simply weighted sums (averages) of all participant responses. One attraction for analyzing these averaged data is the potential for increased statistical fidelity due to the greatly increased number of responses, yielding more stable results. However, as demonstrated by Prince, Brown, and Heathcote (2012), the monotonicity relationships for state trace data averaged over participants need not be representative of the monotonicity relationships for any individual participant. This problem is not unique to state-trace analysis. A classic illustration from decision theory is the averaging of utility functions across participants. The shape of the resulting average utility function need not resemble the utility function of even a single individual (e.g., Estes, 1956; Luce, 1999). Indeed, we can consider this 'averaging problem' as the classic Condorcet voting paradox in a different guise (Condorcet, 1785).

Hierarchical modeling is one recommended solution to the averaging problem (Morey, Pratte, & Rouder, 2008; Pratte, Rouder, & Morey, 2010; Rouder & Lu, 2005). In a hierarchical model, parameters of individuals are assumed to arise from parent populations of parameters. This allows the sharing of information across participants, without assuming that all participants have the same parameters, which is the implicit assumption when averages are used. Among the good properties of hierarchical models is "shrinkage", where individual participants' estimates are pulled towards a common population mean. Hierarchical shrinkage yields more efficient estimates than simply fitting each individual separately, due to the use of information from the population. Shrinkage would be especially useful with state-trace methods, because it can reduce the amount of data required to obtain clear inference from state-trace experiments.

Hierarchical modeling might thus appear to be helpful in the analysis of state-trace data. However, the non-parametric nature of state-trace analysis makes it difficult to build a hierarchical state-trace model. Consider the problem of defining population distributions on individuals' parameters. If we simply create hierarchical populations of probability parameters, then each probability will be pulled towards the observed mean probability by hierarchical shrinkage. This represents the averaging problem all over again: if the parameters are not shrunk in the correct parameter space, then there is no guarantee that the resulting shrunken estimates belong to the same family of models as the un-shrunken estimates. The averaging problem is, therefore, a special case of a more general problem that also affects hierarchical models.

A simple example is instructive. Consider a equal-variance signal detection model and a participant with $d' = 2$ under two conditions where response bias is manipulated and $\beta = 0.4$, and 1.4. These yield true hit rates of 0.95 and 0.73, and true false alarm rates of 0.34 and 0.08 respectively. For simplicity, suppose our estimates of these hit and false alarm rates were equal to their true values. Plotting the z-transform of the hit rates against the z-transform of the false alarm rates will yield points on a line with a slope of 1, as required by the equal-variance signal detection model. If we build a hierarchical model on the probabilities that shrink these estimates towards the mean estimate – say, with 50% shrinkage – we obtain hit rates of 0.85 and 0.77, and false alarm rates of 0.18, and 0.11. The z transform of the hit and false alarm rates no longer lie on a line with slope 1: that is, the pooled values are outside of the space predicted by the model. Pooling on a model's parameter space, on the other hand, can never – by definition – produce estimates that are outside of the space of the model.

For hierarchical state-trace to work, a parameterized model constraining paths through the state-space would have to be posited, so that shrinkage would occur in the parameter space of that model. Essentially, one would have to parametrically model the psychological process itself. Building a hierarchical state-trace model would require the very information that using state-trace is supposed to free us from needing. We thus need a method for using the information across participants that does not require making parametric assumptions about the psychological process if we are to retain the nonparametric flavor of state-trace analysis.

To solve the averaging problem in this context, we propose to evaluate the averaged data, $\boldsymbol{P}$, against a specific class of model. Rather than testing whether individual-level data support a collection of ordering conditions that correspond to a particular state-trace model, as in Prince et al., we propose to test whether the aggregated group-level data support the *convex hull* of those ordering conditions. More formally, suppose we have w-many orders of interest, e.g., all monotonic overlapping orders. Let S_k denote the region of Θ in which the parameter values, $\theta_i, i \in \{1, 2, \ldots, Z\}$, satisfy the k^{th} order, $k \in \{1, 2, \ldots, w\}$. In contrast to Prince, Brown, and Heathcote (2012), rather than evaluating whether data support the model defined by $\cup_{k=1}^{w} S_k$, we evaluate whether the aggregated group-level data support the model $\mathcal{S}_t = conv\{\cup_{k=1}^{w} S_k\}$, where $conv\{X\}$ denotes the convex hull of the set X.

The *convex hull* of a set X is defined as the smallest convex set containing X. The convex hull operation has a natural interpretation in terms of probability distributions. Let X be a finite set of values in $\mathbb{R}^Z$. Then the convex hull of X can be written as the following set:

$$conv\{X\} = \left\{ \sum_{j=1}^{|X|} a_j \boldsymbol{x}_j | \boldsymbol{x}_j \in X, a_j \geq 0, \ \forall j \text{ and } \sum_{j=1}^{|X|} a_j = 1 \right\}, \qquad (2)$$

where $|X|$ denotes the cardinality of X. As seen in (2), the convex hull over a set of values in $\mathbb{R}^Z$ can be equated with the set of all probability distributions over the elements of X, as the convex hull is defined via all nonnegative weights, a_j, that sum to one. This property of the convex hull operator holds for the general case where X is an infinite set (see Boyd & Vandenberghe, 2004).

Applied to our state-trace question, if the model defined by restricting Θ to the convex hull over a set of specified orders on the parameters $\theta_i, i \in \{1, 2, \ldots, Z\}$, is supported by the aggregated group-level data, $\boldsymbol{P}$, then these data be described as conforming to a *mixture distribution* where each individual in the group makes responses corresponding to one of the allowable orderings. In other words, $\boldsymbol{P}$ can be described as a weighted sum such that each a_j weight corresponds to the probability of selecting an individual who chooses according to the corresponding ordering relationship. See also Regenwetter et al. (2014) for a discussion of how convex hull models can be interpreted using averaged data.

Conversely, if the aggregated group-level data, $\boldsymbol{P}$, are not well described by this convex hull model then these data *cannot* be described via a probability distribution over individuals who all produce responses that satisfy one of the viable orders. This logically follows because the convex hull is defined as the set of all such distributions. Thus, at least a subset of the individuals did not make responses according to one of the specified orders. In other words, at least one individual (potentially more) violates the ordering conditions.

As a simple illustration, consider the case of two allowable orderings on the $\theta_i, i \in \{1, 2, \ldots, Z\}$, parameters in Eq. (1). Purely for illustrative and visual purposes, we will assume $Z = 3$, i.e., just three experimental conditions. In the left-hand side of Fig. 4, we consider the parameter space Θ restricted to two allowable orders, $\theta_1 \leq \theta_2 \leq \theta_3$ and $\theta_3 \leq \theta_2 \leq \theta_1$ (red-shaded region). Suppose we had data from two participants where each subject completed one hundred responses for each of the three conditions. Assume one subject made responses conforming to the $\theta_1 \leq \theta_2 \leq \theta_3$ ordering and gave the following proportion of correct responses: $\frac{0}{100}$, $\frac{70}{100}$, and $\frac{72}{100}$ for conditions 1, 2, and 3 respectively. Suppose the other subject made responses in accordance with the $\theta_3 \leq \theta_2 \leq \theta_1$ ordering and gave the following proportion of correct responses: $\frac{64}{100}$, $\frac{62}{100}$, and $\frac{0}{100}$ for conditions 1, 2, and 3 respectively. Momentarily putting aside the question of statistical inference, each subject appears to have produced data that conform to one of the two specified orders. However, averaging their data gives $\boldsymbol{P} = (\frac{64}{200}, \frac{132}{200}, \frac{72}{200}) = (0.32, 0.66, 0.36)$. This value of $\boldsymbol{P}$ is plotted as 'Data1' on the left-hand side of Fig. 4. Clearly, this group-level data does *not* satisfy either order as it is not contained within the red-shaded region on the left-hand figure, a classic example of the 'averaging problem' described by Prince, Brown, and Heathcote (2012). On the right-hand side of Fig. 4, we consider the model formed by restricting Θ to the convex hull of these two ordering conditions (green shaded region). Here, our hypothetical data generated by averaging these two participants is perfectly satisfied by the model. Indeed, this is true more generally. If all individuals in a state trace experiment make responses that conform to a viable order then the aggregated group-level data, $\boldsymbol{P}$, must also conform to the convex hull over the set of all viable orders.

Consider now a second set of data, with the same experimental design as above with two participants. Suppose now that the aggregated group-level data is equal to $\boldsymbol{P} = (\frac{160}{200}, \frac{60}{200}, \frac{160}{200}) = (0.8, 0.3, 0.8)$. This value is plotted in Fig. 4 as "Data2". As can be seen on the right-hand side of Fig. 4, Data2 violates the convex hull model (green-shaded figure). This implies that this particular set of aggregated data *cannot* have arisen from two participants who both satisfy one of the two viable orders. Depending upon the outcome of a formal statistical analysis, a researcher could conclude from this value of $\boldsymbol{P}$ that at least one of the participants made responses that violate both viable orders.

Finally, is important to note in Fig. 4 that the convex hull of the two viable orders (right-hand figure) occupies a larger portion of the parameter space than their union (left-hand figure). For some combinations of $S_k, k \in \{1, 2, \ldots, w\}$, the union can be identical to the convex hull, but this does not hold true in general. The union is necessarily contained within the convex hull, i.e., $\cup_{k=1}^{w} S_k \subseteq \mathcal{S}_t$. Our test remains viable whether or not the union equals the convex hull of the union. We next describe how to carry out a formal statistical test of the convex hull model via a Bayes factor.

3.1. Statistical analysis

Similar to Prince, Brown, and Heathcote (2012) we employ the order-constrained Bayes factor methodology of Klugkist et al. (2005) to evaluate our convex hull models. Let U_1 be defined as the 'encompassing model' formed by placing no a priori restrictions on the θ_i values. Then the Bayes factor for $\mathcal{S}_t$ and U_1, denoted ABF_{t1} (aggregated Bayes factor), is defined as the ratio of the two marginal likelihoods,

$$ABF_{t1} = \frac{p(\boldsymbol{M}|\mathcal{S}_t)}{p(\boldsymbol{M}|U_1)} = \frac{\int L(\boldsymbol{M}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathcal{S}_t)d\boldsymbol{\theta}}{\int L(\boldsymbol{M}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|U_1)d\boldsymbol{\theta}} \qquad (3)$$

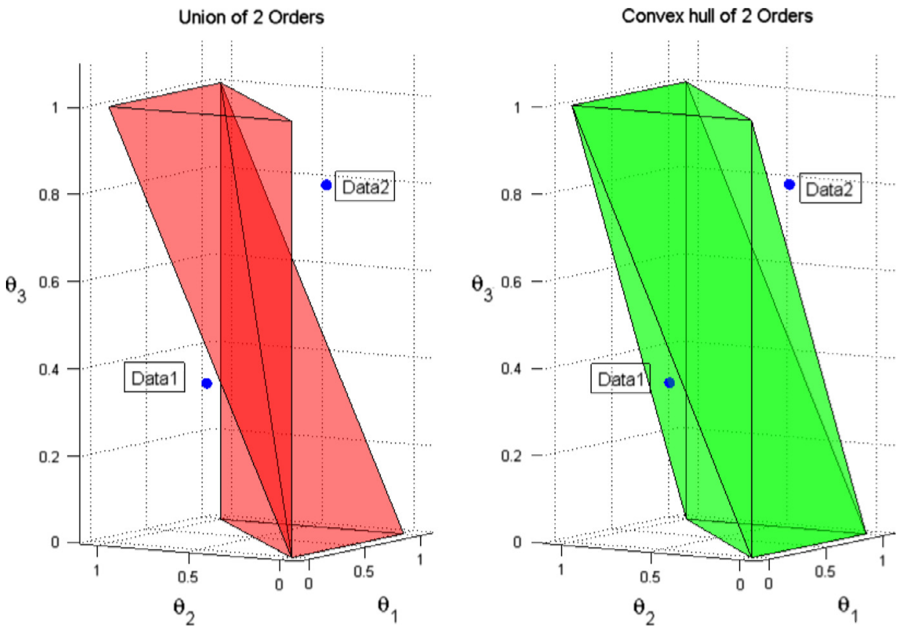

Fig. 4. The left-hand figure plots the union of the two allowable orderings on the θ_i, $i \in \{1, 2, 3\}$ parameters. The right-hand figure plots the convex hull of these two orderings. Data1 and Data2 represent $\boldsymbol{P}$ values that satisfy, respectively violate, the convex hull model. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

where $\pi(\boldsymbol{\theta}|\mathcal{S}_t)$ is the prior distribution of $\boldsymbol{\theta}$ under model $\mathcal{S}_t$, which is defined to be uniform on the support of $\mathcal{S}_t$. The ABF_{t1} is defined with respect to the encompassing model and evaluates the strength of evidence, in terms of the likelihood of generating the observed data, of our mixture model against the encompassing model. As described more fully in Klugkist et al. (2005), Eq. (3) can be further simplified. The Bayes factor under this inequality-constrained framework can be described as the ratio of two proportions: the proportion of the encompassing prior in agreement with the constraints of $\mathcal{S}_t$ and the proportion of the encompassing posterior distribution in agreement with the constraints of $\mathcal{S}_t$. This simplification gives,

$$ABF_{t1} = \frac{c_t}{d_t}, \tag{4}$$

where $\frac{1}{c_t}$ is the proportion of the encompassing prior in agreement with the constraints of $\mathcal{S}_t$ and $\frac{1}{d_t}$ is the proportion of the encompassing posterior distribution in agreement with the constraints of $\mathcal{S}_t$. The proportion, $\frac{1}{c_t}$, for a mixture model can be easily obtained via uniform Monte Carlo sampling. Under our uniform priors, this value is simply the volume occupied by $\mathcal{S}_t$ in Θ. We calculate the $\frac{1}{d_t}$ terms using standard Monte Carlo sampling methods, i.e., random draws from the conjugate beta posterior formed by a uniform prior over each θ_i in Eq. (1).

To speed computation of $\frac{1}{d_t}$ and $\frac{1}{c_t}$ it is helpful to obtain a *facet-defining* description of S_t. Geometrically, $\mathcal{S}_t$ can be described as the convex hull of all extremal points corresponding to all of the pre-specified order restrictions on the $\boldsymbol{\theta}$ values. Since the number of orderings is finite, $\mathcal{S}_t$ will take the form of a convex polytope. A convex polytope is compactly described by its collection of facet-defining inequalities, i.e., the shortest possible list of linear inequalities on the θ_i, $i \in \{1, 2, \ldots, Z\}$, parameters such that the solution set is the convex polytope of interest. For the Sense et al. (in preparation) re-analysis, we used the open-source software, PORTA (Christof, Löbel, & Stoer, 1997) to calculate the system of facet-defining inequalities for the $\mathcal{S}_t$ models we consider. This system of facet defining inequalities is provided in the Appendix A. To calculate either $\frac{1}{d_t}$ or $\frac{1}{c_t}$, we need only check if the sampled $\boldsymbol{\theta}$ values from the posterior (prior, respectively) satisfy each and every linear inequality in this list. If a sampled $\boldsymbol{\theta}$ value satisfies all inequalities then it lies within $\mathcal{S}_t$, otherwise not. In supplementary materials (see Appendix B) we provide R functions using the rPorta package (Nunkesser, Straatmann, Wenzel, Christof, & Loebel, 2009) to calculate the ABF_{t1} for arbitrary state-trace designs.

3.2. Re-analysis of Sense et al. data

We now demonstrate, step-by-step, how to carry out a series of convex hull tests on aggregated state trace data. For the Sense et al. (in preparation) data, Table 1 presents all twenty possible condition orders conforming to the trace constraint. We will exclude the two non-overlapping conditions in our analysis (Orders 1 and 20). For each of the twelve experimental conditions, we assign a triple, (t, d, s), referring to the trace, dimension, and state value of that experimental condition, e.g., $(8, a, seq)$ refers to the condition of set size 8, articulate condition under the sequential manipulation — see Table 1. Let $P(Correct|(t, d, s))$ denote the average probability of a correct response to condition (t, d, s). To construct our model, let the θ_i assignments be as follows:

$$\begin{aligned}
\theta_1 &= P(Correct|(8, a, sim)), & \theta_2 &= P(Correct|(8, s, sim)),\\
\theta_3 &= P(Correct|(4, a, sim)), & \theta_4 &= P(Correct|(2, a, sim)),\\
\theta_5 &= P(Correct|(4, s, sim)), & \theta_6 &= P(Correct|(2, s, sim)),\\
\theta_7 &= P(Correct|(8, a, seq)), & \theta_8 &= P(Correct|(8, s, seq)),\\
\theta_9 &= P(Correct|(4, a, seq)), & \theta_{10} &= P(Correct|(2, a, seq)),\\
\theta_{11} &= P(Correct|(4, s, seq)), & \theta_{12} &= P(Correct|(2, s, seq)).
\end{aligned}$$

Thus, for Order 4 in Table 1, we would need to jointly impose the following two orders on the θ_i values,

$$\begin{aligned}
&\theta_1 < \theta_2 < \theta_3 < \theta_4 < \theta_5 < \theta_6 \quad \text{and}\\
&\theta_7 < \theta_8 < \theta_9 < \theta_{10} < \theta_{11} < \theta_{12}.
\end{aligned}$$

Likewise, we consider the restrictions placed by the remaining 17 joint orders on the θ_i values. Each of the 18 joint orders defines a convex subset of the parameter space. We proceed by finding the facet-defining inequalities that correspond to the convex hull

Table 1
The set of 20 condition orders (accuracy increasing from left to right) conforming to the trace constraint (as set size increases accuracy decreases). Sets size is indicated by a digit (2,4 or 8); a = articulate condition, s = silent condition. The two non-overlapping orders are indicated by NO in the subset column. The 5 orders conforming to the dimension constraint (accuracy in articulate less than silent) are indicated by D in the subset column.

Number	1	2	3	4	5	6	Subset
1	8a	4a	2a	8s	4s	2s	D & NO
2	8a	4a	8s	2a	4s	2s	D
3	8a	4a	8s	4s	2a	2s	D
4	8a	8s	4a	2a	4s	2s	D
5	8a	8s	4a	4s	2a	2s	D
6	8a	4a	8s	4s	2s	2a	
7	8a	8s	4a	4s	2s	2a	
8	8a	8s	4s	4a	2a	2s	
9	8a	8s	4s	4a	2s	2a	
10	8a	8s	4s	2s	4a	2a	
11	8s	8a	4a	2a	4s	2s	
12	8s	8a	4a	4s	2a	2s	
13	8s	8a	4a	4s	2s	2a	
14	8s	8a	4s	4a	2a	2s	
15	8s	8a	4s	4a	2s	2a	
16	8s	8a	4s	2s	4a	2a	
17	8s	4s	8a	4a	2a	2s	
18	8s	4s	8a	4a	2s	2a	
19	8s	4s	8a	2s	4a	2a	
20	8s	4s	2s	8a	4a	2a	NO

over the 18 orders. Using the PORTA software, we obtain 34 linear inequalities on the θ_i, $i \in \{1, 2, \ldots, 12\}$ parameters. These are the constraints that will be employed in calculating the ABF_{t1} test. See Appendix A for a complete list of the inequalities.

Next, we use this system of inequalities, along with samples from the prior and posterior distributions to obtain the values of $\frac{1}{c_t}$ and $\frac{1}{d_t}$ respectively. For these data, we obtained a Bayes factor, ABF_{t1}, of 1666.7 in favor of S_t, defined as the convex hull over Orders 2–18 from Table 1. This is not surprising given what we already know regarding the individual-level analyses. Thus, we conclude that these averaged data support monotonicity. All code and data used in these examples are available as an online supplement (see Appendix B).

Next, we consider the model formed by the convex hull of all orders conforming to the dimension constraint, omitting non-overlapping orders.[2] For the Sense et al. data, these orders correspond to Orders 2–5 in Table 1. We calculated the facet-defining inequalities for this convex hull (see Appendix A) and carried out the order-constrained statistical methodology described above. We obtained a Bayes factor of 2088.6 in favor of this convex hull model. This is somewhat stronger evidence than the first model we considered. We conclude that the aggregated data strongly support this model.

3.3. *Testing convex hulls and the averaging problem*

Evaluating the convex hull of a set of viable relations allows for an interpretable test on aggregated data, but does not completely eliminate the averaging problem. If the convex hull model is rejected on a set of aggregated group-level data, we can be confident that at least one individual in the group makes responses that violate the viable orders. However, if the convex hull model is supported via the ABF_{t1} test we cannot be confident that all members satisfy monotonicity. In other words, it is possible to average responses from individuals who violate a set of viable orders and obtain an aggregated value of $\boldsymbol{P}$ that satisfies the convex hull model.

This problem is analogous to a recent discussion concerning the random preference transitivity model of Regenwetter, Dana, and Davis-Stober (2011). This is a model of individual-level preference that is defined as the convex hull of all transitive preference relations. The model allows an individual to change his or her preferences at each sampled time, with the restriction that all preferences satisfy transitivity (Regenwetter, Dana, & Davis-Stober, 2011). Similar to our group-level application, if an individual's responses violate this model of transitivity the inference is on solid footing as, by convexity, the aggregated data could *not* have been generated by solely transitive preferences (up to sampling error). On the other hand, if an individual's responses are well-described by this model we cannot be certain that this person only made choices consistent with transitive preferences at all time points. Due to the averaging problem, it is possible for aggregated responses conforming to intransitive preferences to satisfy transitivity when aggregated across responses (Birnbaum, 2011). Regenwetter, Dana, Davis-Stober, and Guo (2011) argued that this problem is partially mitigated to the extent that the convex hull of a set of viable preferences is restrictive, i.e., occupies a small proportion of the parameter space.

For this individual-level model of transitivity, one could get around this problem by examining the individual trials that the participant completed, but, as noted by Regenwetter, Dana et al. (2011), it is not always clear how to parse individual, trial-level data, unless the experiment was designed to follow a particular "blocking" structure (Birnbaum, 2011). Said differently, there is often not a unique way to decompose an aggregated within-subject data vector into trial-level data vectors of equal length, — see Regenwetter et al. for a full discussion of this problem.

The averaging problem is similar for our proposed aggregated Bayes factor test, but with a few important differences. As we demonstrated, if the aggregated-across-subjects data violate the convex hull of a set of viable orders then, by convexity, at least one individual does not satisfy the specified monotonicity relationship. Should the model be well-fit, as in our analysis of the Sense et al. (in preparation) data, it is still possible, due to the averaging problem, for these aggregated data to have been generated by individuals that do not satisfy the specified monotonicity relationship. In this way, testing the convex hull of a set of viable orders for aggregated group level data provides a one-directional test. If the model is not well-fit, we arrive at the strong conclusion that the group is not homogeneous with respect to the monotonicity relationship being tested. If the model is supported we cannot unequivocally state that all individuals satisfy monotonicity.

However, similar to the models of Regenwetter, Dana, and Davis-Stober (2011) and Davis-Stober (2012), these convex hull models are often extraordinarily parsimonious. For example, the convex hull of the 18 orders described above occupies only 0.06% of Θ. Furthermore, in contrast to the within-subjects application of Regenwetter et al., the across-subjects application described here has an advantage in that the aggregated data are naturally decomposed into responses from individual participants. In other words, should we find support for the convex hull of a set of viable orders, we need only examine the individual participant responses to guard against this instantiation of the averaging problem. Returning to the Sense et al. data, we found strong support for the convex hull model, thus, we need to examine individual responses, which we know to be well-described by monotonicity, see Figs. 1 and 2. Thus, ABF_{t1} is interpretable for these data as indicating that all participants are monotonic.

[2] For these data, we do not consider the convex hull model formed by the two non-overlapping orders. For reasons described in Prince, Brown, and Heathcote (2012), these orders are not diagnostic for determining monotonicity in state-trace analyses. Testing non-overlapping orders is useful primarily as a diagnostic technique for evaluating the efficacy of the experiment, i.e., rejecting the non-overlapping orders indicates that the experiment was effective at producing condition overlap. For these reasons, we do not consider a convex hull model for these orders. The Sense et al. data, at the individual level, indicates substantial overlap.

3.4. Sensitivity of the convex hull test

This line of inquiry raises the question of how sensitive the ABF_{t1} test is at detecting groups with individuals that violate the specified ordering conditions. The ability of the ABF_{t1} test to "detect" a group with individuals that violate the specified ordering conditions depends upon multiple factors. First, if a minority of individuals violate monotonicity, the sensitivity of the test strongly depends upon how severe the violations are. If data from only a few individuals violate the model, and not severely, then the aggregated group-level data may not violate the convex hull model strongly, or at all, yielding a Bayes factor in favor of the monotonic model. Second, the sensitivity of the ABF_{t1} test also depends upon how heterogeneous the responses of all the group members are. In general, the more heterogeneous the responses of the group members the less sensitive the ABF_{t1} test is. For example, suppose we have a group of eighteen individuals who each produce responses that conform to one of the Orders 2–18 in Table 1, with each individual having responses that conform to a distinct order. If we add individuals to this group whose responses do not conform to any of these orders, all else equal, we will need to add many more such violating individuals to reject the monotonic convex hull model than if the original 18 participants made responses conforming to the same order.

To examine the sensitivity of the ABF_{t1} test at detecting violating individuals, we carried out the following simulation study. We modeled our simulation conditions after the general experimental design of the Sense et al. (in preparation) experiment. We simulated a total of fifteen individuals in each group, with each individual responding to 100 trials per condition, with twelve total experimental conditions. For the ABF_{t1} test, we considered the convex hull model defined over Orders 2–18 from Table 1. The goal of this simulation study is to evaluate the sensitivity of the ABF_{t1} test under different levels of heterogeneity of responses. All individual-level responses were generated from independent binomial random variables, one per condition, with the binomial parameters systematically chosen to model differing levels of heterogeneity in the group.

Our simulation study was carried out under two conditions: "high" and "low" levels of group heterogeneity. In both conditions we systematically varied the proportion of individuals who violated monotonicity to those who satisfied it. We considered 15 possible configurations per condition: 1 individual violating with 14 satisfying, 2 violating to 13 satisfying, and so on, with, at the extreme, 15 violating monotonicity and 0 satisfying. We begin by describing the high heterogeneity condition. For each individual assigned to satisfy monotonicity, we randomly sampled binomial parameter values across all experimental conditions by uniformly sampling from the convex hull model over Orders 2–18. Once these parameter values were sampled, data was generated according to the sampled parameter values with 100 responses per condition per subject. For monotonicity violating individuals, we randomly sampled the binomial parameter values from the remainder of the space that does not conform to the convex hull model and generated data accordingly. This represents a worst case scenario for the convex hull test, as the uniform sampling ensures a highly heterogeneous mix of individuals that satisfy the convex hull model, as well as those that violate it. In the low heterogeneity condition, we proceeded as in the previously described condition with the exception that all simulated individuals assigned to satisfy monotonicity produce data generated according to the same randomly selected order. The violating individuals were simulated randomly as in the high heterogeneity condition. Across all conditions we applied the ABF_{t1} test to the aggregated, group level data for each simulated data set.

The results of this simulation are displayed in Fig. 5. The x-axis displays the number of simulated violating individuals in the group, with the remainder simulated to conform to the convex hull model according to either the minimum or maximum heterogeneity methods. The y-axis displays the proportion of the simulated data sets, per condition, that yielded a Bayes factor smaller than 1, indicating support for the encompassing model. The values on the y-axis were obtained for each condition and ratio-membership level by repeating the following process 4000 times: independently sampling response probabilities that conform/violate the convex hull model in the specified ratio according to either the high or low heterogeneity condition, simulating data responses, and carrying out the $ABGF_{t1}$ test. Thus, the values on the y-axis are interpreted as the proportion of simulated data sets that yielded at least weak support for the encompassing model. For example, under the low heterogeneity condition, 7.3% of simulated data sets generated by groups with a single violating individual would be extreme enough to provide any evidence against the convex hull model by the ABF_{t1} test.

As Fig. 5 shows, the ABF_{t1} test is more sensitive at detecting monotonicity violating individuals when there is less heterogeneity present. Under the low heterogeneity condition, if there is at least one-third (5 subjects) monotonicity violating individuals in the group, there is an excellent chance of obtaining evidence against the convex hull model via ABF_{t1}, and for only one-fifth (3 subjects) there is a greater then even chance. As seen in Fig. 5, the high heterogeneity condition causes the convex hull test to be much less sensitive. Under these conditions, it is difficult to obtain evidence against the convex hull model until the number of monotonicity violating individuals is in the group majority. Both of these conditions represent "extremes" and actual data will tend to fall somewhere in between.

3.5. Relationship between ABF_{t1}, GBF and individual tests

For the Sense et al. (in preparation) data, the evidence is overwhelmingly in support of monotonicity. Thus, ABF_{t1}, GBF, and the individual-level tests are all in agreement. One could, however, construct alternative scenarios where the various tests could systematically disagree, as they each evaluate different hypotheses and, in the case of ABF_{t1} and GBF, are defined on different data structures. For example, it is possible for GBF to not strongly support either all monotonicity or all non-monotonicity in the group, but for ABF_{t1} to strongly support at least one individual violating monotonicity.

The group-level tests become especially important when the number of trials at the individual level is very small, e.g., a study could have a large number of participants but few responses per participant. In this case one might expect most individual-level tests to be equivocal. However, the GBF could still strongly favor monotonicity in two ways. First, weak preferences expressed by all individual tests could be very consistent. In this case the ABF_{t1} is likely to also support homogeneity. Second, the GBF could be dominated by a few individual tests displaying much stronger preferences than the remainder. In this case the ABF_{t1} can help differentiate between the case where the remaining individuals are also consistent with monotonicity, in which case the ABF_{t1} would tend to support homogeneity, and the case where some or all are inconsistent, in which case the ABF_{t1} would tend to support heterogeneity.

4. Discussion

We have presented a series of Bayes factor techniques for determining whether the observed points in a state-trace plot (Bamber, 1979) of binary dependent variables are monotonic, with the aim of deciding if one latent psychological variable, or more, mediates the interacting effects of two manipulations (experimental "state" and "dimension" factors). We summarized previous

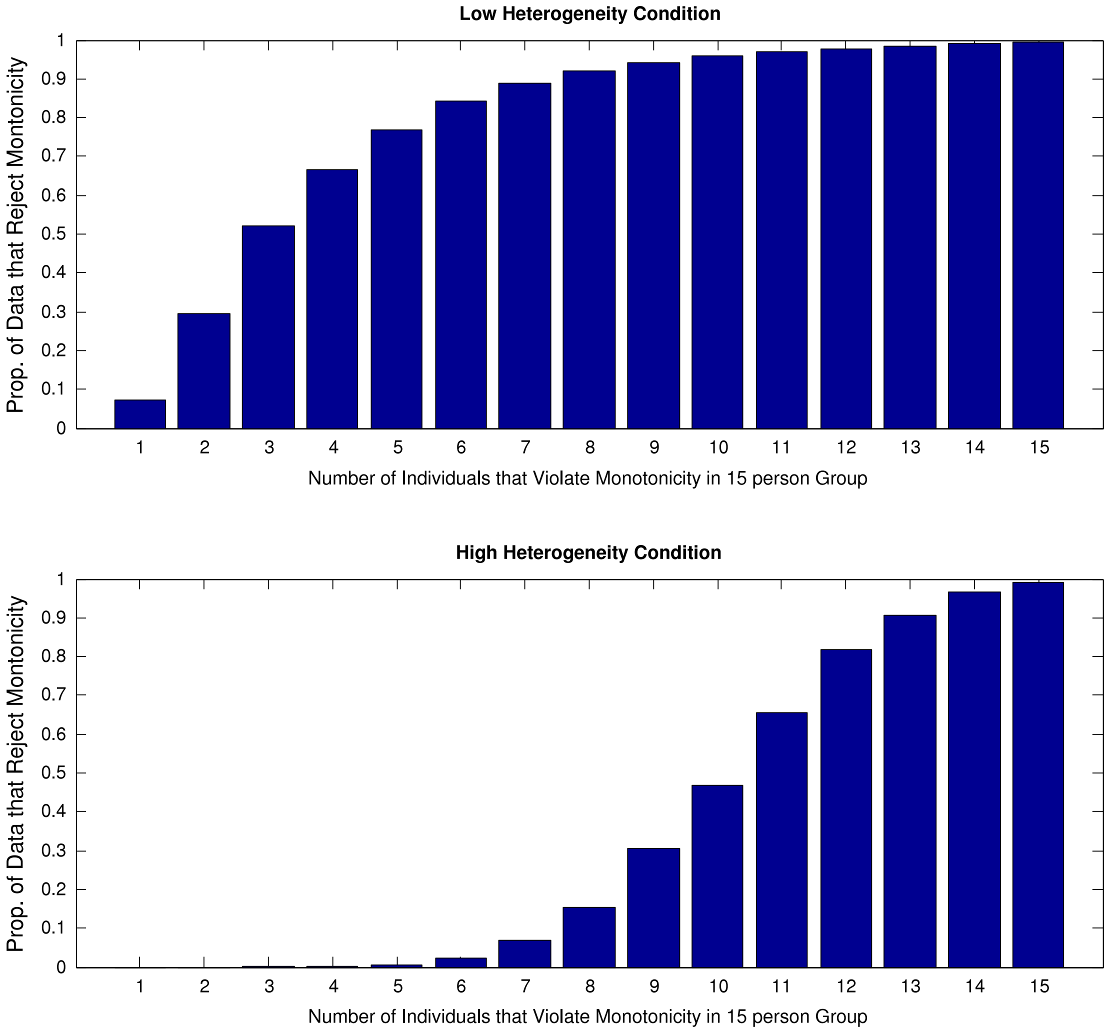

Fig. 5. The x-axis displays the number of simulated violating individuals in a group of 15 members. The y-axis displays the proportion of the simulated data sets, per condition, that yielded a Bayes factor smaller than 1, indicating support for the encompassing model.

Bayesian methods for evaluating these questions at the individual level, extended them by showing how a priori plausible restrictions can be added to produce more focused tests, and developed a new computational methods that greatly speed their application. We then presented a new group-level, across-participant state-trace analysis. This new "aggregated Bayes factor" (ABF_{t1}) is applied to data averaged over participants. It can provide unambiguous (up to sampling error) evidence for heterogeneity among participants in their monotonicity. This allows it to be used to guard against inappropriate use of Prince, Brown, & Heathcote's (2012) method of combining individual-level Bayes factors, the "group Bayes factor" (GBF), which assumes homogeneity. However, it cannot unambiguously support homogeneity, so when homogeneity is supported individual-level tests need to be examined. Hence, we recommend that ABF_{t1}, GBF and individual-level tests all be calculated in concert with one another in order to provide the most solid basis for state-trace inference.

There is an additional caveat with respect to the ability of the individual tests to support the null hypotheses. Note that state-trace analysis as typically presented relies on testing whether a finite set of observed points have the same ordering across multiple dependent variables. The null hypothesis is that these points are monotonic. However, the theoretical hypothesis of interest is typically not about the limited set of points observed, but rather all about points that might have been observed; the null hypothesis is that all points that might be observed in a state-trace plot follow a single, monotone function. Even if this null hypothesis is violated, there still might be finite subsets of points which are monotone. The null hypothesis in the state-trace analysis is in a sense, therefore, a different null hypothesis from the one of theoretical interest, and moving from support for the null that *these* points are ordered to the null that *all* points are ordered involves an implicit smoothness assumption.

It would be desirable to explicitly test the more general and theoretically interesting null hypothesis, but this does not seem possible within a nonparametric state-trace analysis. Testing hypotheses about points on curves that could have been observed but were not involved in making assumption about the curve, and this is precisely the sort of assumption that state-trace analysis attempts to avoid. The cost of this nonparametric orientation is that, although evidence for the alternative hypothesis of nonmonotonicity is straightforward to interpret, evidence for the null is not. For potential users of state-trace analysis that desire a test of the more theoretically interesting null hypothesis, semi-parametric hierarchical modeling may provide a way forward, but at the cost of stronger assumptions. Future development in this area would be welcome, giving researchers more analytical choices.

In a recent article, Ashby (2014) argues state-trace analysis is conceptually unable to uncover the number of underlying cognitive 'systems' required to model a given cognitive task. In reply, Dunn, Kalish, and Newell (2014) point out that no statistical procedure can make inferences about the number of cognitive systems, because the concept of a system is not appropriately defined. They also assert that the logic of state trace analysis is sound, as it does allow for rigorous inference about the number of latent variables required to model a particular behavior. Examples generated by Ashby (2014) appear to contradict this assertion because they violate the fundamental assumption made by state-trace analysis, that the mapping between latent and manifest variables is monotonic. Although this assumption is often true in cognitive models, the examples provided by Ashby (2014), some of which concern latent variables related to decision criteria, underline the fact that it is not always true. As long as care is taken on this point, we agree with Dunn et al. that state-trace analysis can provide valid inferences about the number of parameters needed to model a particular behavior. Hence, we believe it provides valuable and relatively assumption-free guidance about the application of Occam's razor in theory testing and development.

Because it is important to develop parsimonious theories, we believe Bayesian methods – which do not have the bias against simplicity inherit in frequentist null-hypothesis tests – are an ideal means of inference about state-trace plots. Some researchers may feel cautious about using Bayesian methods because they require specification of a prior. In contrast, we see the ability to systematically incorporate prior knowledge into inference as a distinct advantage. Prince, Brown, and Heathcote (2012) provided one demonstration of the utility of this advantage, by incorporating a "trace model" prior (i.e., the assumption that a third manipulation, the trace factor, has a monotonic effect) into their inference. As we demonstrated here, this avoids spurious inflation of the evidence supporting a one-dimensional explanation of the interaction between state and dimension factors by conflating it with evidence supporting the trace model. We then extended Prince et al.'s methods, showing how to incorporate additional prior assumptions, including above-chance performance and a restriction on the ordering produced by the dimension factor. Our example analysis demonstrated that assuming these extra priors provided a clearer answer to a psychological question, whether change detection involved verbal recoding. We argued that both proponents and opponents of verbal recoding are likely to agree on these priors, just as they are likely to agree on the trace prior, because they are orthogonal to their point of disagreement. Hence, it is both uncontroversial and advantageous to incorporate all of these points of prior agreement in order to produce more focused inferences about the issue at hand.

One reason that researchers could feel caution about specifying priors is that they may not have strong grounds for a quantitative specification (e.g., an effect of a particular magnitude or set of magnitudes). Fortunately, this is not an issue with respect to the ordinal priors relevant to state-trace analysis. In supplementary material (see Appendix B) we provide general-purpose R functions that enable users to impose arbitrary prior orders on some or all of the levels of both the trace and dimension factors. This software complements and extends Prince Prince, Hawkins, et al.'s (2012) StateTrace package. It also incorporates the Laplace approximation, enabling accurate Bayes factors to be calculated in seconds rather than overnight. Together, this functionality provides a comprehensive and fairly general purpose[3] approach to state-trace inference with binary dependent variables.

Being able to specify more nuanced ordinal priors also addresses another potential objection to Prince, Brown, & Heathcote's (2012) approach to inference, that it relies on a uniform prior assumption under which all orders are equally likely. We believe that in some cases this sort of prior is desirable. For example, in larger designs there may still be a number of ways that the full set of points in a state-trace plot can be ordered even after trace and dimension assumptions have been imposed. These depend on the quantitative effect sizes associated with the magnitudes of differences between factor levels. Unless researchers have prior quantitative empirical results or theoretical models that generate quantitative priors it seems to us quite reasonable to assume all orders are equally likely a priori. However, wherever prior constraint can be placed on orders in a way that is orthogonal to the research question addressed by state-trace data – as was the case in our example analysis – we strongly encourage researchers to make use of them. In this way they are afforded two of the major advantages of Bayesian analysis, systematic incorporation of prior knowledge and even-handed inference that can provide evidence for simpler models (i.e., that can accept a null model) as well as evidence for more complex models.

A weakness of state-trace analysis is that it is difficult to apply to averaged data. This is because averaging distorts ordinal relationships, a fact that has been appreciated at least since the articulation of the classic Condorcet voting paradox (Condorcet, 1785). Prince, Brown, and Heathcote (2012) noted that averaging is particularly problematic for state-trace analysis, and so advocated individual-level analysis and group-level analysis by taking the product of individual-participant Bayes factors (i.e. the *GBF*). However, their approach to group-level analysis makes the strong assumption that groups are homogeneous, with either all participants being monotonic or all participants being non-monotonic. The aggregated Bayes factor based on averaged data that we developed here provides a test of the assumption that all participants are monotonic. This test is limited, because it can (up to sampling error) reject homogeneity of monotonicity but not affirm it, at least not in a strong sense (Regenwetter, Dana et al., 2011). However, we argued that an apparent alternative solution, hierarchical modeling, is not viable without accepting an alternative limitation, sacrificing the largely assumption-free nature of state-trace analysis. Hence, we believe that our approach, testing averaged data using the properties of convex polytopes, combined with Prince, Brown, & Heathcote's (2012) group Bayes factor represents a useful approach.

The methods that we used to develop the test of averaged data could also be extended to test homogeneity of non-monotonicity. That is, one could also consider the model formed by the convex hull of the set of all non-monotonic relations on the $\boldsymbol{\theta}$ values. Such a test, if rejected, could provide evidence for whether at least one individual supports monotonicity. As with the previous analysis, we would proceed by using PORTA to calculate the facet-defining inequalities for the convex hull of all such non-monotonic relations. We would then apply the order-restricted Bayes methodology described above. This route would provide another test of heterogeneity, this time from the perspective of all individuals satisfying non-monotonic relationships. As before, should this model be satisfied, due to the averaging problem, this test could not confirm homogeneous non-monotonicity across individuals.

Tests of membership in convex hulls might conceivably be carried out by the same fast Laplace methods that we used to set unions of orders. Because only computation need only be carried out for the group rather than each participant sampling methods are not particularly burdensome, and we have not yet explored this possibility. However, this issue may become more pressing when testing homogeneity of non-monotonicity as the set of all non-monotonic orders can be much larger than the set

[3] One practical restriction that remains concerns the number of points in the state-trace plot (n). Our computational methods first enumerate the $n!$ permutations of these points, then cull them to obtain the subset following the trace-model order. This method works in a reasonable time for designs with up to 10 points (i.e., a 5 trace level by 2 dimension level design, which has 3628,800 permutations), and the number of integrations required also remains manageable (e.g., there are 252 trace-model orders for the 5 × 2 design and 1680 for a 3 × 3 design). However, a more efficient method that directly generates the trace-model subset will be required for larger designs.

of all monotonic orders. This issue aside, one might consider a further extension, building on the first, by computing a Bayes factor directly comparing the convex hull of all viable monotonic orders to that of the convex hull of all non-monotonic relationships. This Bayes factor pits the same models against each other as the *GBF*, but based average rather than individual data, and so makes the same assumption of group homogeneity. More broadly, we note that all mixture models formulated by taking the convex hull of specified orders will necessarily be nested within the linear ordering polytope — as the linear ordering polytope is defined as the convex hull over all linear orders. Future work could explore connections between the polytopes considered in our state-trace analysis and the well-studied facet structure of the linear ordering polytope (e.g., Fishburn, 1992; Suck, 1992).

Acknowledgments

Heathcote and Morey were supported by ARC Discovery Project DP110100234 and Heathcote by DP120102907. Davis-Stober was supported by National Science Foundation grant (SES-1459866, PI: Davis-Stober). We would like to thank Dr. E-J. Wagenmakers and two referees for their helpful comments and suggestions. Any errors or inconsistencies are the sole responsibility of the authors.

Appendix A

Below is PORTA output for the facet-defining description of the monotonicity mixture model over Orders 2–19 from Table 1. Each x_i variable in the below inequalities refers to θ_i in our model. For example, row (2) corresponds to the linear inequality, $-\theta_2 \leq 0$.

```
DIM = 12
VALID
1 1 1 1 1 1 1 1 1 1 1 1
INEQUALITIES_SECTION
(  1) -x1                                     <= 0
(  2)    -x2                                  <= 0
(  3)                   -x7                   <= 0
(  4)                      -x8                <= 0
(  5)                                +x11-x12 <= 0
(  6)                         +x9-x10         <= 0
(  7)                      +x8   -x10         <= 0
(  8)                      +x8       -x11     <= 0
(  9)                   +x7   -x9             <= 0
( 10)                   +x7              -x12 <= 0
( 11)             +x5-x6                      <= 0
( 12)       +x3-x4                            <= 0
( 13)    +x2   -x4                            <= 0
( 14)    +x2      -x5                         <= 0
( 15) +x1   -x3                               <= 0
( 16) +x1            -x6                      <= 0
( 17)                                    +x12 <= 1
( 18)                            +x10         <= 1
( 19)                +x6                      <= 1
( 20)          +x4                            <= 1
( 21) -x1         +x5   +x7          -x11     <= 1
( 22) -x1+x2            +x7-x8                <= 1
( 23)    -x2+x3            +x8-x9             <= 1
( 24)       -x3      +x6      +x9        -x12 <= 1
( 25)       -x3   +x5         +x9    -x11     <= 1
( 26)          -x4   +x6         +x10    -x12 <= 1
( 27)          -x4+x5            +x10-x11     <= 1
( 28)          +x4-x5            -x10+x11     <= 1
( 29)          +x4   -x6         -x10    +x12 <= 1
( 30)       +x3   -x5         -x9    +x11     <= 1
( 31)       +x3      -x6      -x9        +x12 <= 1
( 32)    +x2-x3            -x8+x9             <= 1
( 33) +x1-x2            -x7+x8                <= 1
( 34) +x1         -x5   -x7          +x11     <= 1
END
```

Below is PORTA output for the facet-defining description of the monotonicity mixture model over Orders 2–5 from Table 1. Each x_i variable in the below inequalities refers to θ_i in our model.

```
DIM = 12
VALID
0 0 0 0 0 0 1 1 1 1 1 1
INEQUALITIES_SECTION
(  1) -x1                                     <= 0
(  2)                   -x7                   <= 0
(  3)                                +x11-x12 <= 0
(  4)                            +x10    -x12 <= 0
(  5)                         +x9-x10         <= 0
(  6)                         +x9    -x11     <= 0
(  7)                      +x8   -x10         <= 0
(  8)                      +x8       -x11     <= 0
(  9)                   +x7-x8                <= 0
( 10)                   +x7   -x9             <= 0
( 11)             +x5-x6                      <= 0
( 12)          +x4   -x6                      <= 0
( 13)       +x3-x4                            <= 0
( 14)       +x3   -x5                         <= 0
( 15)    +x2   -x4                            <= 0
( 16)    +x2      -x5                         <= 0
( 17) +x1-x2                                  <= 0
( 18) +x1   -x3                               <= 0
( 19)                                    +x12 <= 1
( 20)                +x6                      <= 1
( 21)    -x2+x3            +x8-x9             <= 1
( 22)          -x4+x5            +x10-x11     <= 1
( 23)          +x4-x5            -x10+x11     <= 1
( 24)    +x2-x3            -x8+x9             <= 1
END
```

Appendix B. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.jmp.2015.08.004.

References

Ashby, F. G. (2014). Is state-trace analysis an appropriate tool for assessing the number of cognitive systems? *Psychonomic Bulletin & Review*, *21*(4), 935–946.

Bamber, D. (1979). State-trace analysis: A method of testing simple theories of causation. *Journal of Mathematical Psychology*, *19*(2), 137–181.

Birnbaum, M. H. (2011). Testing mixture models of transitive preference: Comment on Regenwetter, Dana, and Davis-Stober. *Psychological Review*, *118*(1), 675–683.

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

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press.

Busey, T. A., Tunnicliff, J., Loftus, G. R., & Loftus, E. F. (2000). Accounts of the confidence-accuracy relation in recognition memory. *Psychonomic Bulletin & Review*, *7*(1), 26–48.

Christof, T., Löbel, A., & Stoer, M. (1997). Porta-polyhedron representation transformation algorithm. Software package, available for download at http://www.zib.de/Optimization/Software/Porta.

Condorcet, M. (1785). Essai sur l'application de l'analyse à la probabilité de décisions rendues à la pluralité de voix, Imprimerie royal, Paris.

Davis-Stober, C. P. (2012). A lexicographic semiorder polytope and probabilistic representations of choice. *Journal of Mathematical Psychology*, *56*(2), 86–94.

Dunn, J. C. (2008). The dimensionality of the remember-know task: A state-trace analysis. *Psychological Review*, *115*(2), 426–446.

Dunn, J. C., Kalish, M. L., & Newell, B. R. (2014). State-trace analysis can be an appropriate tool for assessing the number of cognitive systems: A reply to Ashby. *Psychonomic Bulletin & Review*, *21*(4), 947–954.

Dunn, J. C., & Kirsner, K. (1988). Discovering functionally independent mental processes: the principle of reversed association. *Psychological Review*, *95*(1), 91.

Dunn, J. C., & Kirsner, K. (2003). What can we infer from double dissociations? *Cortex*, *39*(1), 1–7.

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin*, *53*(2), 134–140.

Fishburn, P. C. (1992). Induced binary probabilities and the linear ordering polytope: A status report. *Mathematical Social Sciences*, *23*(1), 67–80.

Gelfand, A. E., Smith, A. F. M., & Lee, T. M. (1992). Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *Journal of the American Statistical Association*, *87*(418), 523–532.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, *1*(2), 141–149.

Guérard, K., & Tremblay, S. (2008). Revisiting evidence for modularity and functional equivalence across verbal and spatial domains in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(3), 556–569.

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

Heathcote, A., Bora, B., & Freeman, E. (2010). Recollection and confidence in two-alternative forced choice episodic recognition. *Journal of Memory and Language*, *62*(2), 183–203.

Heathcote, A., Freeman, E., Etherington, J., Tonkin, J., & Bora, B. (2009). A dissociation between similarity effects in episodic face recognition. *Psychonomic Bulletin & Review*, *16*(5), 824–831.

Henson, R. (2006). Forward inference using functional neuroimaging: dissociations versus associations. *Trends in Cognitive Sciences*, *10*(2), 64–69.

Jones, D., Farrand, P., Stuart, G., & Morris, N. (1995). Functional equivalence of verbal and spatial information in serial short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(4), 1008–1018.

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: A Bayesian approach. *Psychological Methods*, *10*(4), 477–493.

Loftus, G. R. (1978). On interpretation of interactions. *Memory & Cognition*, *6*(3), 312–319.

Loftus, G. R., & Irwin, D. E. (1998). On the relations among different measures of visible and informational persistence. *Cognitive Psychology*, *35*, 135–199.

Loftus, G. R., Oberg, M. A., & Dillon, A. M. (2004). Linear theory, dimensional theory, and the face-inversion effect. *Psychological Review*, *111*(4), 835–865.

Luce, R. D. (1999). *Utility of gains and losses: Measurement-theoretical and experimental approaches*. Psychology Press.

Meiser, T., & Klauer, K. C. (1999). Working memory and changing-state hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(5), 1272–1299.

Morey, R. D. (2005). Confidence intervals from normalized data: A correction to Cousineau. *Tutorial in Quantitative Methods for Psychology*, *4*, 61–64.

Morey, R. D., Pratte, M. S., & Rouder, J. N. (2008). Problematic effects of aggregation in z roc analysis and a hierarchical modeling solution. *Journal of Mathematical Psychology*, *52*(6), 376–388.

Murray, D. J. (1965). Vocalization-at-presentation and immediate recall, with varying presentation rates. *Quarterly Journal of Experimental Psychology*, *17*, 47–56.

Newell, B. R., & Dunn, J. C. (2008). Dimensions in data: testing psychological models using state-trace analysis. *Trends in Cognitive Sciences*, *12*(8), 285–290.

Newell, B. R., Dunn, J. C., & Kalish, M. (2010). The dimensionality of perceptual category learning: A state-trace analysis. *Memory & Cognition*, *38*(5), 563–581.

Nunkesser, R., Straatmann, S., Wenzel, S., Christof, T., & Loebel, A. (2009). rporta: R/porta interface. Retrieved from http://CRAN.R-project.org/package=rPorta (R package version 0.1-93).

Pratte, M. S., Rouder, J. N., & Morey, R. D. (2010). Separating mnemonic process from partici- pant and item effects in the assessment of roc asymmetries. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(1), 224–232.

Prince, M., Brown, S., & Heathcote, A. (2012). The design and analysis of state-trace experiments. *Psychological Methods*, *17*(1), 78–99.

Prince, M., Hawkins, G., Love, J., & Heathcote, A. (2012). An R package for state-trace analysis. *Behavior Research Methods*, *44*(3), 644–655.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–164.

Regenwetter, M., Dana, J., & Davis-Stober, C. P. (2011). Transitivity of preferences. *Psychological Review*, *118*(1), 42–56.

Regenwetter, M., Dana, J., Davis-Stober, C. P., & Guo, Y. (2011). Parsimonious testing of transitive or intransitive preferences: Reply to birnbaum. *Psychological Review*, *118*(1), 684–688.

Regenwetter, M., Davis-Stober, C. P., Lim, S.-H., Guo, Y., Popova, A., Zwilling, C., et al. (2014). Qtest: Quantitative testing of theories of binary choice. *Decision*, *1*(1), 2–34.

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, *12*(4), 573–604.

Schooler, J. W., & Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: some things are better left unsaid. *Cognitive Psychology*, *22*, 36–71.

Sense, F., Morey, C.C., Prince, M., Heathcote, A., & Morey, R.D. (n.d.). Opportunity for ver- balization does not improve visual change detection performance: A state-trace analysis (in preparation).

Shallice, T. (1988). *From neuropsychology to mental structure*. Cambridge University Press.

Stigler, S. M. (1986). Laplace's 1774 memoir on inverse probability. *Statistical Science*, *1*, 359–363.

Suck, R. (1992). Geometric and combinatorial properties of the polytope of binary choice probabilities. *Mathematical Social Sciences*, *23*(1), 81–102.