Journal of Mathematical Psychology 90 (2019) 118–131

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

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

A hierarchical Bayesian state trace analysis for assessing monotonicity while factoring out subject, item, and trial level dependencies

Patrick Sadil *, Rosemary A. Cowell, David E. Huber

Department of Psychological and Brain Sciences, University of Massachusetts, Amherst, MA 01003, USA

HIGHLIGHTS

- State-Trace Analyses (STA) test the latent dimensionality of cognitive processes.
- STA assess monotonicity across conditions between dependent measures.
- Current methods for STA assume independence between measures.
- Unmodeled dependence can bias state-trace analyses.
- We present a hierarchical model for STA that accounts for such dependencies.

ARTICLE INFO

Article history:
Received 16 May 2018
Received in revised form 8 October 2018
Available online 22 March 2019

ABSTRACT

State trace analyses assess the latent dimensionality of a cognitive process by asking whether the means of two dependent variables conform to a monotonic function across a set of conditions. Using an assumption of independence between the measures, recently proposed statistical tests address bivariate measurement error, allowing both frequentist and Bayesian analyses of monotonicity (e.g., Davis-Stober, Morey, Gretton, & Heathcote, 2016; Kalish, Dunn, Burdakov, & Sysoev, 2016). However, statistical inference can be biased by unacknowledged dependencies between measures, particularly when the data are insufficient to overwhelm an incorrect prior assumption of independence. To address this limitation, we developed a hierarchical Bayesian model that explicitly models the separate roles of subject, item, and trial-level dependencies between two measures. Assessment of monotonicity is then performed by fitting separate models that do or do not allow a non-monotonic relation between the condition effects (i.e., same versus different rank orders). The Widely Applicable Information Criterion (WAIC) and Pseudo Bayesian Model Averaging – both cross validation measures of model fit – are used for model comparison, providing an inferential conclusion regarding the dimensionality of the latent psychological space. We validated this new state trace analysis technique using model recovery simulation studies, which assumed different ground truths regarding monotonicity and the direction/magnitude of the subject- and trial-level dependence. We also provide an example application of this new technique to a visual object learning study that compared performance on a visual retrieval task (forced choice part recognition) versus a verbal retrieval task (cued recall).

State-trace analyses test the dimensionality of psychological spaces (Bamber, 1979; Dunn, 2008) by framing the situation in terms of manipulations (e.g., experimental study conditions) that affect dependent variables (e.g., accuracy in cued-recall versus recognition) via unobservable, latent variables. The goal is to probe the dimensionality of the latent space. For example, a study might use a range of different experimental study conditions and compare cued-recall versus recognition performance to assess how many latent psychological variables are required to explain performance. If there is just one latent psychological variable underlying behavior (e.g., memory strength), then the two dependent variables must be affected in the same qualitative manner (i.e., the same rank order across conditions for each dependent measure). Conversely, if there are two or more latent psychological variables (e.g., item-strength versus association-strength), different rank orders are possible. Terminology for the different components of a state-trace analysis can differ (compare Davis-Stober, Morey, Gretton, & Heathcote, 2016; with Dunn, 2008; Dunn & Kalish, 2018). We adopt the terminology of Dunn and Kalish (2018), referring to the *outcome space* as all possible combinations of values for the dependent variables and the *predicted* or *model state-trace* as the subset of the outcome space feasibly reached by a model that maps an independent factor (i.e., experimental conditions), via latent variables, onto the outcome space. The goal of a state-trace analysis is to identify the model that best accounts for the *observed* state-trace (i.e., the values of the dependent measures in an experiment,

* Correspondence to: Department of Psychological and Brain Sciences, University of Massachusetts, 135 Hicks Way, Amherst, MA 01003, USA.
E-mail address: psadil@umass.edu (P. Sadil).

https://doi.org/10.1016/j.jmp.2019.01.003
0022-2496/

across experimental manipulations) — where "best" is a measure of a model's ability to recapitulate the data at hand, penalized by its flexibility.

State-trace analysis provides a "scale-free" assessment of the dimensionality of the latent psychological space. In other words, the analysis enables assessment of whether one latent variable is sufficient to account for the data, or whether additional variables are required, with this assessment making only the minimal assumption that the dependent measures lie on a monotonic scale in terms of the underlying psychological variables. This is achieved by making only the minimal assumption that the latent variables map in a monotonic manner onto the observable outcomes. That is, a state-trace analysis provides a way to make inferences about the dimensionality of the latent space without committing to a specific model of the transformation from independent experimental factors to latent variables, nor latent variables to the predicted state-trace. Although a state-trace analysis can be performed to compare models with an arbitrarily-sized latent space, typically the models under comparison have one versus two latent dimensions. Comparisons between such "unidimensional" and "bidimensional" models will be the focus of this paper.

In the following explanation, an observed instance of the dependent variables defining the outcome space X and Y will be referred to as x and y, respectively. Likewise, a value of the latent variables A and B will be referred to as a or b, respectively. The transformation of the latent variables into outcomes will be written as functions, f, g, or h.

In a unidimensional model, outcomes are assumed to be separate monotonic functions of a single latent variable

$$x_{uni} = f_{uni}(a),$$
$$y_{uni} = g_{uni}(a). \quad (1)$$

In a bidimensional model, the outcomes depend on two latent variables. The resulting state-trace is defined by[1]

$$x_{bi} = f_{bi}(a, b),$$
$$y_{bi} = g_{bi}(a, b). \quad (2)$$

Despite the generality of the definition of each of these models, the unidimensional and bidimensional models will typically make different predictions about the monotonicity of the true state-trace. Specifically, the unidimensional model always predicts that the data plotted in the outcome space will exhibit a monotonic relationship provided that f the mapping from the latent variable to both dependent measures is monotonic. For instance, if an increase in memory strength from low to medium elicits higher recognition accuracy, the unidimensional model predicts that a further increase in memory strength from medium to high *cannot decrease* recognition accuracy. Thus, an observed non-monotonic state-trace result falsifies any unidimensional model. In the event that a unidimensional model is rejected, further statistical or rhetorical arguments may be required to advocate for a particular multidimensional model — that is, specification of the psychological constructs that correspond to the separate latent variables (Dunn, Kalish, & Newell, 2014).

Observing evidence for monotonicity in the data requires more careful inference considering that there are different ways that the bidimensional model can be conceptualized. In one approach, the model comparison asks whether the data are better described by a monotonic function or a non-monotonic function. In the approach we take here, the unidimensional model is a special case of the bidimensional model. A key idea of a state-trace analysis is that the unidimensional model implies order restrictions on the DV means, and our model comparison will take advantage of this by considering ordered-restricted models as nested within the set of all possible orderings. That is, the unidimensional model will be *nested* within the bidimensional model.

Given this conceptualization of the bidimensional model as encompassing the unidimensional model, parsimony may be invoked to give credence to the unidimensional model (Dunn, 2008). With this nested model assumption, the bidimensional model can produce monotonic state traces in circumstances when the experiment is not well calibrated to the system in question or the latent dimensions are highly correlated (Davis-Stober et al., 2016; Dunn et al., 2014; Prince, Brown, & Heathcote, 2012). In other words, although the bidimensional model has two different latent variables that can in principle map onto the chosen dependent variables in different ways, the particular manipulations in a reported dataset may happen to affect both latent variables in the same manner, producing a situation where there is no discernible difference between the latent variables. Providing an example of this, Jang, Lee, and Huber (2019) reanalyzed a prominent state-trace result in the metamemory literature that reported separate monotonic functions for manipulations of two independent variables, with each state trace performed at a different level of a third independent variable. Their reanalysis revealed a non-monotonic function when the state trace analysis included all three independent variables concurrently. Thus, the separate monotonic functions were found to reflect at least two latent variables that happened to affect the dependent measures in the same way.

To summarize this distinction, one approach asks of each dataset whether the function relating the two dependent measures is monotonic versus non-monotonic. In contrast, the nested modeling approach asks whether the data require more than one latent variable, but if not, this lends support for the unidimensional model. Critically, a monotonic function does not falsify the bidimensional model. To make this distinction clear, suppose that an analysis of an experiment favored a monotonic function but then a direct replication of the experiment favored a non-monotonic function. Such a pattern of results could either be viewed as two separate conflicting conclusions regarding monotonicity or, alternatively, this pattern of results could be viewed as favoring the bidimensional model considering that the bidimensional model can explain both findings. Extrapolating from this example, consider that each participant in an experiment is a replication. Thus, if 50% of participants clearly produce a monotonic function while the remaining 50% clearly produce a non-monotonic function, one approach would say the data are equivocal regarding monotonicity whereas the other approach would favor the bidimensional model for its ability to explain both groups of participants.

Although the uni- and bidimensional models will typically correspond to different state-trace results, dependent variables are, of course, greatly influenced by sources of noise. That is, although a state-trace analysis provides a qualitative comparison between models, its use requires a way of quantitatively comparing these models in light of measurement error (see Eq. (3)).

$$x \sim f_{uni}(a) \quad y \sim g_{uni}(a) \quad vs. \quad x \sim f_{bi}(a, b) \quad y \sim g_{bi}(a, b). \quad (3)$$

The property of Eq. (3) that sets it apart from Eqs. (1) and (2) is that (3) has introduced an unknown probabilistic error term resulting in a functional relationship between the latent variables and the dependent measures. Whereas f and g were deterministic functions in (1) and (2) (i.e., any input would always produce the same output), f and g signify probability distributions in (3) (i.e., a given input will produce structured but variable output). The inferential challenge lies in determining which model provides the

[1] For the bidimensional model, it is not necessary that both f_{bi} and g_{bi} depend on both a and b, but rather that at least one of these functions depends on both latent variables or that one function depends on one latent variable while the other function depends on the other latent variable.

best account of the data when f and g are unknown functions that include this probabilistic error.

Statistical techniques exist for adjudicating between the uni- and bidimensional models (Davis-Stober et al., 2016; Dunn & James, 2003; Dunn & Kirsner, 1988; Kalish, Dunn, Burdakov, & Sysoev, 2016; Loftus, Oberg, & Dillon, 2004; Pratte & Rouder, 2012; Prince et al., 2012), but we identified two limitations of the current methods, both relating to the role of unacknowledged dependencies in the data. First, most of these tests are unable to partition variance into both item and subject effects (see Pratte & Rouder, 2012 for an exception).[2] Second, these techniques assume independence between X and Y. Whereas the inability to account for item and subject variability only limits the power of a study to detect non-monotonicity, the assumed independence between X and Y could systematically bias the results.

Regarding the independence assumption, consider a situation in which participants decide to devote more resources to one task (e.g., a dependent measure of recognition) at the expense of the other task (e.g., a dependent measure of cued recall), with the choice of which task to favor differing for different participants. This results in a negative dependency between the two measures at the subject level. With sufficient data, violations of this independence assumption should not matter (i.e., the data will overwhelm the prior assumption of independence). However, with limited data per subject, the independence assumption allows subject effects to freely move as necessary to capture the randomly varying means for each subject (i.e., fitting noise owing to a limited sample). In contrast, a model that includes a term capturing the subject effect dependency will be appropriately constrained (aka, exhibit "shrinkage") in its estimate of subject effects, depending on the pattern of results for the other subjects. For instance, if subjects who perform well on recognition tend to also perform well on cued recall (i.e., a positive dependency), then a particular subject who did well on cued recall but not recognition may have their posterior distribution for recognition adjusted upwards in light of the general pattern across the other subjects (i.e., poor recognition performance for this subject is deemed to be sampling error).

Such dependencies can also occur at the item level (e.g., high frequency words are easier to recall but harder to recognize than low frequency words). Furthermore, if the experiment collects both measures in relation to the same item for the same subject (e.g., following item recognition, use of that same item as a cue for recall), such dependencies can also be identified at the trial level (e.g., a greater focus on one task than the other, with this focus varying from trial to trial). Thus, any of the random factors of an experiment (e.g., subject, item, and possibly trial) may exhibit a dependency pattern that could be negative (e.g., focus on one task at the expense of the other) or positive (e.g., some words are more memorable regardless of how they are tested). Next, we consider how these unacknowledged dependencies may affect the ability of a monotonic model to capture the data.

Consider a hypothetical experiment with real-valued dependent measures plotted along the X and Y axes for four conditions, as shown in Fig. 1. In the figure, the colored ellipses show the spread of data owing to variance and covariance, with green ellipses (left column) showing a positive dependence whereas red ellipses (right column) show a negative dependence. In both cases, the variance for each Dependent Variable (DV) is the same, corresponding to the range of the ellipse along the X and Y axes. The dependencies (i.e., the tilt of the ellipse) could reflect subject, item,

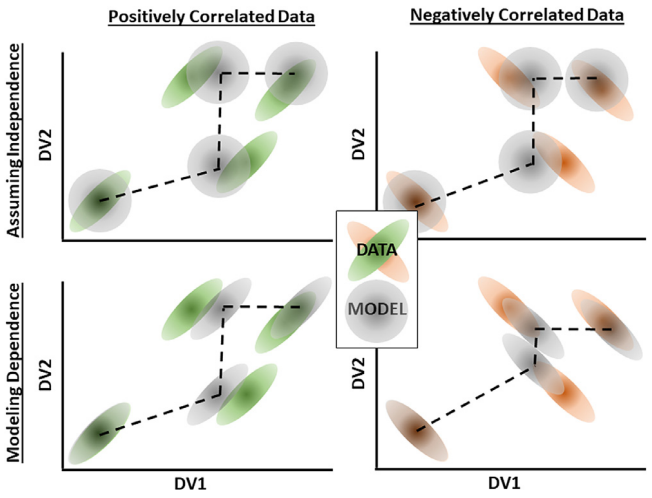

Fig. 1. Monotonicity assessment errors due to unacknowledged dependencies. Panels depict hypothetical datasets with differing patterns of correlation on two tasks (DV1, DV2), across four experimental conditions. Data from the experimental conditions are depicted with colored ellipses (green for a positive correlation and red for a negative correlation). The gray ellipses represent a hypothetical model of these data and are connected by an approximately best-fitting monotonic function in each case (dashed line). These functions were chosen by hand for illustrative purposes. The top row is a model that assumes independence between the two tasks whereas the bottom row is a model that estimates the correlation in the data. In these hypothetical examples, the standard deviations are assumed to be constant and known (all ellipses have equal standard deviation). Note that average performance per condition and variance for each task is identical in all situations. However, a model that only considers condition means and variances, but not covariances, neglects valuable information that could constrain model fitting. For instance, the bottom row shows that the best-fitting monotonic function fails to capture the middle two conditions in the case of positively correlated data but not negatively correlated data. . (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and/or trial level correlations in the data (the model we developed includes the combination of all three sources of dependency). For ease of exposition, consider that this hypothetical experiment only allowed variation for one of these three possible random factors: e.g., one subject tested for many trials on the same item; or one subject tested on many items, but with each item only tested once; or many subjects tested with the same item, with that item only tested once per subject. These three situations are mathematically identical (i.e., just one random factor underlying the different data points for each condition), demonstrating that this hypothetical example applies equally to subject, item, or trial dependencies.

A hypothetical best-fitting monotonic function is drawn with the dashed lines (i.e., the line segments are constrained to have positive slope), and the gray ellipses centered on each junction of the piecewise linear monotonic function capture the best-fitting condition means for each of the four conditions. The observed condition means are the same in all cases and are positioned as to map out a non-monotonic function. In this case, the question of interest is the power to reject the (incorrect) monotonic model (i.e., to what extent does the monotonic model misfit the data).

The top row shows application of a model that assumes independence whereas the bottom row depicts a model that captures the dependence, allowing the direction of the dependence to be determined by the observed data patterns. Judging by the overlap between the gray circles (assumed independence) and the colored ellipses in the top row it appears that the monotonic model performs similarly regardless of the dependency in the data, with perhaps a slightly worse fit (greater power to reject the monotonic model) in the case of negatively correlated data (the simulations reported below verified this greater power with negatively correlated data). Next, consider the bottom row, which

[2] The Kalish et al. (2016) technique could be modified by relaxing the block-diagonal constraint to estimate subject and item effects simultaneously, but this would still be under an assumption of independence between the dependent measures (i.e., the subject or item effects for one dependent measure would not be constrained by values for the other dependent measure).

shows application of a model that explicitly captures the observed dependency. For ease of exposition, the dependency as determined by the model is set equal to the dependency in the data, but in real applications, the model's estimate of dependency may differ from reality in an attempt to capture as much of the data as possible despite the monotonicity constraint. As seen in the bottom row, the model is clearly capturing more of the data when the observed data are negatively correlated (red) as compared to when they are positively correlated (green). More specifically, for the middle two conditions under positively correlated data, the positive dependency in the model results in a complete misfit of the corresponding data distributions. In contrast, for the middle two conditions under negatively correlated data, the misfitting condition means can nonetheless capture much of the data distribution.

In summary, the choice of a statistical model that assumes independence versus one that captures the observed dependence influences the statistical power to reject the monotonic model. More specifically, if the pattern across the conditions is non-monotonic, but with a general trend of increasing conditions means, there appears to be greater power to reject the monotonic model for positively correlated data *provided that the statistical model includes this positive correlation in its explanation of the data*. The complementary scenario is equally true: when the pattern of condition means exhibits a generally decreasing trend, it is easier to correctly reject the monotonic model for negatively correlated data. To be clear, we are not suggesting that dependencies will necessarily always produce the same bias for or against one of the models in question; the effect of dependencies will depend on the arrangement of the condition means. However, the example shown in Fig. 1 suggests that a failure to include the observed dependencies in the statistical model may lead to model selection errors (we provide quantitative evidence of these model selection errors in the model recovery section of this paper).

Our development of a state-trace analysis technique that addresses subject, item, and trial-level dependencies is motivated by a recently submitted experiment that examined two different accuracy measures (Sadil, Potter, Huber, & Cowell, submitted). Next, we describe this experiment considering that it serves as an example application of our technique, but we reiterate that our technique is generally applicable and could be readily adapted to real-valued dependent measures (e.g., d-prime rather than accuracy) and other combinations of random factors (e.g., subject effects but no item or trial effects). Our experiment was designed to test for the existence of part-to-part visual associations as one latent variable and part-to-whole associations as a second latent variable. More specifically, we asked whether people could learn and benefit from associations between a particular visual feature at one location of a visually presented object (e.g., the curve on the left side of a computer mouse) and a particular visual feature elsewhere on that same object (e.g., the cord emerging at the front of the mouse), even if they were never able to identify the whole object at the time of learning. The 'Binocular' learning condition presented the object to both eyes so that the visual details were apparent and the object was explicitly identified. The 'Word' learning condition presented only a word that named the object, but no visual details. The 'CFS' learning condition presented the object under Continuous Flash Suppression (Tsuchiya & Koch, 2005), such that visual details were provided, but the observer was not aware that anything was presented and thus they could not name the object. Additionally, learning of these associations was compared to a "Not Studied" baseline condition, which assessed the pre-experimental strength of these associations. The models in question were a unidimensional model, which assumed only one kind of learned information underlying performance, versus a bidimensional model, which held that at least two kinds of associations could be learned (i.e., part-to-part associations versus part-to-name associations).

After study, observers performed back-to-back memory tasks that were designed to be differentially sensitive to the two putative association types: (1) an intact-rearranged two alternative forced-choice (2AFC) between pairs of visual fragments; followed by (2) a cued-recall task to name the object given one of the visual fragments that appeared in the immediately preceding 2AFC test (i.e., this was a test of the same item, allowing assessment of trial level dependencies). Our experimental design tested multiple subjects and used the same set of items for all subjects (with a different random assignment of item to condition for different subjects) and so the data could potentially exhibit subject, item, and trial level dependencies. The results were first analyzed using the technique developed by Kalish et al. (2016), which addresses only subject effects under an assumption of independence between the dependent measures. This technique led to rejection of the unidimensional model. However, we were concerned that unacknowledged dependencies in the data (particularly at the trial level) may have biased this conclusion. Indeed, the results we report below demonstrate that with (simulated) negatively correlated data, the Kalish et al. technique more readily rejects the unidimensional model. This concern led us to develop a new state trace analysis technique. Because there was only one pair of observations at the level of a particular trial for a particular item and a particular subject, we adopted a hierarchical Bayesian modeling approach to allow the use of sparse data to identify dependencies at the subject, item, and trial-levels (Kruschke, 2014; Rouder & Lu, 2005). Ultimately, our new technique reached the same conclusion as the Kalish et al. technique, but through its development we came to appreciate the ways in which our technique might have produced difference conclusions.

1. Hierarchical bivariate probit model for state-trace analysis

The problem in question is a comparison between the following two models

$$(x, y) \sim h_{uni}\,(a) \quad vs. \quad (x, y) \sim h_{bi}\,(a, b) \tag{4}$$

which include the unknown bivariate distributions h_{uni} and h_{bi}. Considering that the data in question are correct/incorrect responses, we used a hierarchical-Bayesian bivariate probit to specify these bivariate distributions (Greene, 2017, pp. 807–810; Stan Development Team, 2017b). A graphical description of the hierarchy is presented in Fig. 2, in the style used by Kruschke (2014). The model assumes that performance on the two tasks is a thresholded process whereby a participant makes a correct response when the value of evidence, z (not depicted), is above some threshold (fixed to 0). Each pair of observations corresponds to a single trial, defined by a unique combination of a subject, s, encountering an item, i, that was studied in condition, c. For example, when both z values are above 0, this corresponds to a trial in which the observer is correct for both tasks. This joint outcome, (x, y), on a given trial is captured with following expression

$$\begin{pmatrix} x \\ y \end{pmatrix}_{c,s,i} \sim \psi^2 \left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}_{c,s,i}, \Omega_c \right) \tag{5}$$

in which the symbol ψ^2 represents the bivariate probit distribution with parameters $\mu \in \mathbb{R}$ and produces a pair of values that are either 0 or 1. The construction of μ_x and μ_y are given below, but for intuition they may be viewed as the means of a bivariate normal distribution whose samples, z, are 1 when they are over the threshold and 0 when they are below the threshold. Ω_c is a standardized covariance matrix (i.e., a correlation matrix), $\Omega_c = \begin{pmatrix} 1 & \rho_c \\ \rho_c & 1 \end{pmatrix}$, where each ρ_c is the correlation between measures in condition c. The variance is fixed at 1 for identifiability; more specifically, identifiability is a potential issue considering that model behavior

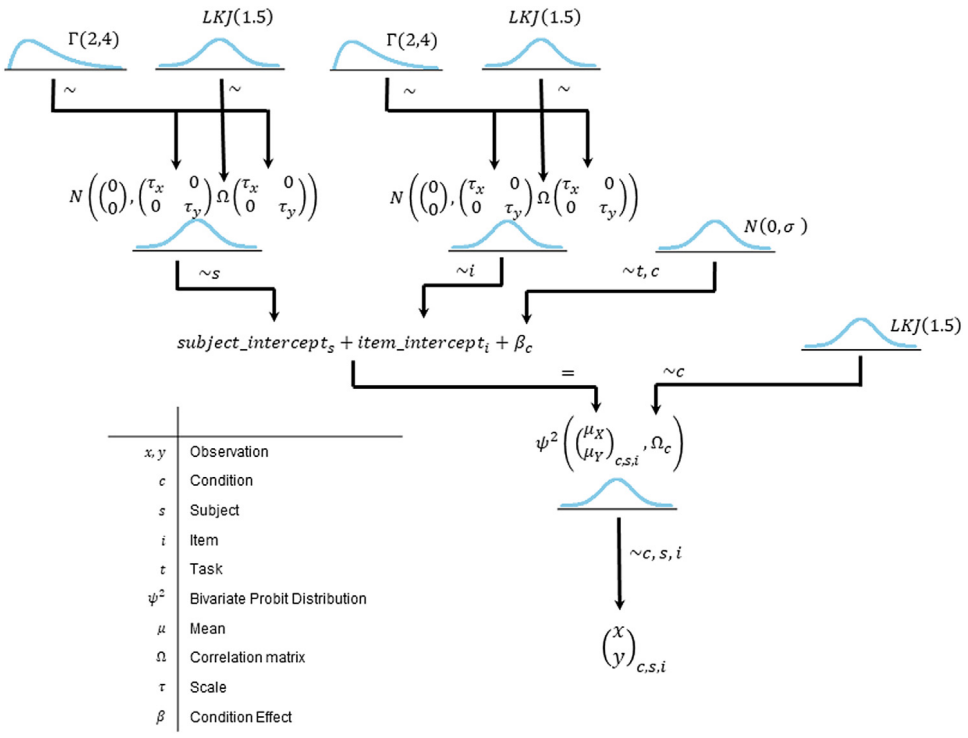

Fig. 2. Schematic of the Bayesian hierarchical bivariate probit model. The tilde symbols (∼) indicate that a parameter is sampled from a distribution whereas the equals symbols (=) indicate deterministic equations. Subscripts indicate different levels of a factor (e.g., items, subjects, or conditions). Distributions are labeled with either *N* for normal (either univariate or bivariate), *Γ* for gamma, or *LKJ* for Lewandowski et al. (2009), which is a distribution on correlation matrices. In the normal distributions, the first parameter signifies the mean, and the second parameter signifies the standard deviation. The two parameters of the gamma are shape and rate. The single *LKJ* parameter is a shape parameter (see footnote in main text for an explanation of the effect of this parameter). In this diagram, vectors are presented as columns. Subject- and Item-level correlations enter the model through the construction of the bivariate probit means. Trial-level correlations enter the model through the rightmost LKJ distribution. For illustrative purposes, the diagram shows centered parameterization of the subject- and item-effects, but in practice the model used a non-centered parameterization. Likewise, the details of the construction of β_c and the mixing proportions have been omitted. See text for further details.

would be the same if the deviations were τ_x and τ_y, rather than 1, provided that the μ_x parameter is rescaled by τ_x and the μ_y parameter is rescaled by τ_y. Dependencies between the two measures are modeled through a combination of the condition-dependent trial-level correlation, ρ_c, and the between-trial dependencies that occur in the construction of the μ parameters through subject and item effects.

In generative terms, this bivariate probit can be seen as a bivariate extension of an Equal Variance Signal Detection model in which the criteria for both decisions have been fixed to 0 (Macmillan & Creelman, 2005, pp. 126–144). However, it would be a mistake to view the use of a bivariate probit as a commitment to a particular parametric form of the latent dimensions underlying task performance. Instead, this probit is just a convenient method for linking a linear model to a pair of binary outcomes. Rather than thinking of the dimensions of the bivariate probit as the latent dimensions *A* and *B*, the dimensions of the bivariate probit can be thought of with the relatively atheoritic language of "information required to make a correct response."

Beyond trial-level correlations, and changes in these correlations with condition, our technique also includes subject effects and item effects and correlations for both subject and items effects through the construction of each μ (Kruschke, 2014, pp. 221–260; Rouder & Lu, 2005). The trial-level means (one for *X* and one for *Y*) of the bivariate probit are each constructed with the following equation

$$\mu_{c,i,s} = subject\_intercept_s + item\_intercept_i + \beta_c, \tag{6}$$

which says that each trial is the summation of a subject- (s), item- (i), and condition- (c) effects. Put another way, the means of the bivariate probit were modeled with an additive structure, with intercepts that varied by item, subject, and condition. Intuitively, each of these effects can be thought of as their respective propensity to elicit the correct response for a given trial.

We note that subject-level and item-level correlations were constrained to take on the same value for all conditions. This constraint was not imposed for theoretically important reasons, but rather because the model is difficult to identify if multiple forms of correlation freely vary with condition (e.g., if the data of just one condition exhibited a negative correlation, it would be difficult to determine if this negative correlation existed at the subject-level, the item-level, and/or the trial-level). To simplify the situation, we allowed the trial-level correlation to take on different values for different conditions, while the subject and item correlations were fixed across conditions. If instead we had applied our approach to an experiment with only one random factor (e.g., subject differences), we might have allowed for correlation differences between conditions for that factor (e.g., subjects could have a different correlation in one condition as compared to others).

The condition effects (the β_c parameter) were fixed effects, with the same value on all trials of a given condition across all subjects. These fixed condition effects are based on separate, independent, normal prior distributions (one for *X* and one for *Y*). These normal distribution priors had a mean of 0 and a variance equal to 0.25. In contrast, the subject and item effects were drawn from bivariate normal prior distributions, capturing dependencies between dependent measures (e.g., a subject that is good at task *X* is also good at task *Y*). Non-centered parameterizations of the bivariate normal priors for the item- and subject-effects were used to facilitate the exploration of the posterior distribution by decomposing that distribution into a mean (shown as 0 in Fig. 2), a standard deviation (scale) along dependent measure (τ), and a correlation matrix

(Ω). This decomposition decouples the sampling of random-effects from the sampling of hyperparameters at higher levels of the hierarchy, often resulting in increased sampling efficiency (Betancourt & Girolami, 2015; Stan Development Team, 2017b). All correlation matrices (including the trial level correlation) were given *LKJ* hyperpriors[3] (Lewandowski et al., 2009) with parameter value of 1.5, and all standard deviations were given gamma hyperpriors ($\Gamma(shape = 2, rate = 4)$). To model these subject- and item-effects as deviations from group-level performance, the means of the bivariate normal priors were fixed to 0 for each kind of effect (e.g., a subject effect of 0 corresponds to a subject whose average performance is equal to the group-level average). Fig. 2 shows the model as described so far, with a simple prior placed on the condition effects, but as described below, application of the model used a more complicated condition effect prior to capture the difference between the unidimensional and bidimensional models.

The net effect of these prior distributions moving down the hierarchy is similar to two independent probit regression models on each dependent measure, considering that the priors at each level are symmetric and centered at 0 (i.e., independence). Given that the mode of a Γ (2, 4) distribution equals 0.25, the most likely contribution of subject and item effects to the variance of the standard normal $\mu_{c,s,i}$ value is .25 + .25 = .5, with the normally distributed condition effect contributing an additional .25 variance. Thus, after passing a $\mu_{c,s,i}$ distribution with a mean of 0 and variance equal to .75 through the probit transformation, the net prior distribution over the accuracy scale is weakly informative, with a mode at .5, but with non-zero density at 0 and 1 (the effect of the order constraints through mixture modeling, as described below, push this composite prior somewhat closer to uniform over the 0 to 1 accuracy scale). We note that prior distributions that impose a smaller variance (larger density around 0.5 on the accuracy scale) or larger variance (larger density at 0 and 1 on the accuracy scale) produced comparable results (see below).

Eqs. (5) and (6) emphasize the data generating process, how the model describes the production of response on a given trial. The model determines the probability of the data according to the following expression, which uses the probability mass function of the bivariate probit (referred to with ψ^2) that results from the above equations

$$p\left(\begin{pmatrix}x\\y\end{pmatrix}_{c,s,i} | \begin{pmatrix}\mu_X\\\mu_Y\end{pmatrix}_{c,s,i}, \Omega_c\right) = \psi^2\left(\begin{pmatrix}x\\y\end{pmatrix}_{c,s,i} | \begin{pmatrix}\mu_X\\\mu_Y\end{pmatrix}_{c,s,i}, \Omega_c\right), \tag{7}$$

That is, the probability of an observation (i.e., a pair of accurate or inaccurate responses for each task) given a set of model parameters (left hand side of the above equation) is assigned by the probability mass function of the bivariate probit (right hand side of the above equation). Note that the probability mass function of the bivariate probit has a natural interpretation in terms of the cumulative density function of the bivariate normal distribution. To be explicit, if we call the probability density function of the bivariate normal ϕ^2, then the probability of the four possible trial outcomes (e.g., all combinations of accurate and inaccurate performance on the two tasks) can be assigned with the following expressions.

$$\begin{aligned}
p\left(\begin{pmatrix}0\\0\end{pmatrix} | \begin{pmatrix}\mu_X\\\mu_Y\end{pmatrix}, \Omega\right) &= \int_{x=-\infty}^{0}\int_{y=-\infty}^{0} \phi^2\left(\begin{pmatrix}\mu_X\\\mu_Y\end{pmatrix}, \Omega\right) dxdy\\
p\left(\begin{pmatrix}1\\0\end{pmatrix} | \begin{pmatrix}\mu_X\\\mu_Y\end{pmatrix}, \Omega\right) &= \int_{x=0}^{\infty}\int_{y=-\infty}^{0} \phi^2\left(\begin{pmatrix}\mu_X\\\mu_Y\end{pmatrix}, \Omega\right) dxdy\\
p\left(\begin{pmatrix}0\\1\end{pmatrix} | \begin{pmatrix}\mu_X\\\mu_Y\end{pmatrix}, \Omega\right) &= \int_{x=-\infty}^{0}\int_{y=0}^{\infty} \phi^2\left(\begin{pmatrix}\mu_X\\\mu_Y\end{pmatrix}, \Omega\right) dxdy\\
p\left(\begin{pmatrix}1\\1\end{pmatrix} | \begin{pmatrix}\mu_X\\\mu_Y\end{pmatrix}, \Omega\right) &= \int_{x=0}^{\infty}\int_{y=0}^{\infty} \phi^2\left(\begin{pmatrix}\mu_X\\\mu_Y\end{pmatrix}, \Omega\right) dxdy
\end{aligned} \tag{8}$$

The expressions in (8) imply that the probability of an outcome is given by the "amount" of the bivariate normal contained within the appropriate quadrant (analytic calculation given by Pan, 2017). The form of the model in (7) is useful for defining our inferential approach, but in contacting the data, we assume a mixture of multiple bivariate probits, similar to prior state-trace models (Davis-Stober et al., 2016; Prince et al., 2012).

Model comparison between the bidimensional and unidimensional models requires the imposition of order constraints (i.e., constraints on the rank order of the condition means for each of the two dependent variables), but the model described so far is unconstrained in its placement of the conditions means. To restrict the model to be strictly monotonic, as per the unidimensional model, order constraints were placed on the condition effects, β. For example, in our experiment with 4 conditions (labeled as 1, 2, 3, or 4), a monotonic state-trace could be realized by requiring that the condition parameters are in the following order: $\beta_1 \leq \beta_2 \leq \beta_3 \leq \beta_4$ in the construction of both μ_X and μ_Y. A non-monotonic state-trace could be realized, for instance, by requiring that the condition effects for the construction of μ_X have order $\beta_1 \leq \beta_2 \leq \beta_3 \leq \beta_4$ whereas the order of μ_Y is $\beta_1 \leq \beta_3 \leq \beta_2 \leq \beta_4$ (where the second and third β have been swapped). The full unidimensional and bidimensional models can then be realized as mixture models over all allowable orderings, where each component of the mixture is a copy of the unconstrained model in Fig. 2, with the imposition of a particular order constraint.

We next describe how, within one component of the mixture, a particular ordering on β is enforced, using the technique developed by Bürkner (2017) and Bürkner and Charpentier (preprint). Conceptually, this technique involves sampling three kinds of parameters. One parameter will be referred to as $\beta^{intercept}$, and another β^{raw}. $\beta^{intercept}$ corresponds to the lowest β and the sum of $\beta^{intercept}$ and β^{raw} corresponds to the highest β. The third kind of parameter is a vector whose elements are between 0 and 1, and sum to 1 (referred to as ζ). This third kind serves to place the remaining β values between the lowest and the highest values. That is, these three kinds of parameters are combined to create a vector of condition effects for all conditions that follow a particular order for the β_c in Eq. (6). After choosing the $\beta^{intercept}$ and β^{raw}, ζ models the normalized proportion of β^{raw} for the "interior" conditions (i.e., the conditions lying between the worst and best).[4]

$$\beta_c = \beta^{intercept} + \beta^{raw} \sum_{i=1}^{n\_conditions-1} \zeta_i \quad c \in 2, \ldots, n\_conditions - 1 \tag{9}$$

The ζ are constrained such that $\zeta_i \in [0, 1]$ for all i and $\sum_{i=1}^{n\_conditions-1} \zeta_i = 1$ (i.e., the ζ are the weights of a simplex). Any distribution capable of generating a simplex could be a suitable

[3] We used *LKJ* hyperprior on correlation matrices rather than the Wishart on covariance matrices owing to the non-centered parameterization of our model. That is, using the *LKJ* hyperpriors allowed for independent control over the population-level scale and correlation parameters. Moreover, given that the trial-level standard deviations were fixed to 1, for consistency this matrix was distributed as an *LKJ*. For the LKJ distribution, in this 2x2 case, a shape parameter equal to 1 implies uniform density across all correlations, a shape parameter between 0 and 1 implies a trough (i.e., values closer to 0 put higher density on both extremely low and extremely high correlations), and a shape parameter greater than 1 concentrates the density symmetrically around 0.

[4] Note that this scheme assumes the order is monotonically increasing and requires that $0 < \beta^{raw}$.

prior for ζ. We used a two-stage method.[5] First, *n_conditions* − 2 intermediate parameters, ζ^{raw}, are sampled from a normal distribution,

$$\zeta^{raw} \sim N\left(0, \sigma_\zeta\right). \tag{10}$$

The ζ^{raw} are then concatenated with a 0 and the resulting vector (still denoted as ζ^{raw} in the following equation) is transformed with the softmax function to produce the final ζ

$$\zeta = \frac{\exp\left(\zeta^{raw}\right)}{\sum_{i=1}^{n\_conditions-1} \exp\left(\zeta_i^{raw}\right)}. \tag{11}$$

The standard deviation hyperprior parameter, σ_ζ, determines the prior likelihood on the spread of the β between $\beta^{intercept}$ and β^{raw}. In the model recovery and initial applications of the technique to real data, σ_ζ was set to 3, which allows for substantial variability in how the interior points cluster between the lowest and highest condition effects. Values of 1 were tried and observed to produce similar results.[6] As stated above, an $N\left(0, \sigma\right)$ prior was placed on each of $\beta^{intercept}$ and β^{raw}, with the standard deviation σ set to 0.5.

The resulting β_c values, together with $\beta^{intercept}$, define a vector of condition effects that follow a predetermined order. To create either the unidimensional or bidimensional models, we used the Davis-Stober et al. (2016) mixture-model approach by combining the different allowable orders under each model.[7] By "mixture", we mean that each of the allowed orderings, j, is assigned a probability of being the true order of condition effects λ_j. The number of λ_j values to estimate is the $k!^m$ possible permutations of condition effect orderings within a model, where k is the number of conditions and $m = 1$ in the unidimensional case and $m = 2$ in the bidimensional case. The same two-stage approach that models the ζ was used to model λ.[8] The likelihood of a given trial, incorporating all *n_orders* that are allowed by a given model is then

$$p\left(\begin{pmatrix} x \\ y \end{pmatrix}_{c,s,i} | \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}_{c,s,i}, \Omega_c\right) = \sum_{j=1}^{n\_orders} \lambda_j \psi^2\left(\begin{pmatrix} x \\ y \end{pmatrix}_{c,s,i} | \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}_{j,c,s,i}, \Omega_c\right). \tag{12}$$

The difference between Eqs. (12) and (7) is the presence of the mixing parameters, λ_j. Eq. (12) implies that the probability of the observed data on a given trial is equal to the weighted sum of the probabilities assigned to bivariate probits whose means have been constructed with each of the allowable orderings in a model. The weights are equal to the estimated likelihood of that order. The components of the mixture are differentiated by how the μ parameters are constructed (i.e., the order constraints put on the condition effects) and the weights, λ_j, instantiate uncertainty about the true order of condition effects for a given model.

[5] The two-stage method was chosen as a prior for ζ rather than directly sampling from e.g., a Dirichlet for computational reasons. That is, initial attempts to sample ζ directly from a Dirichlet often resulted in autocorrelated chains, whereas an informal assessment of the two-stage method appeared to result in more efficient sampling. Note that in the section *Applications to real data*, the model reached the same conclusions regardless of whether the interior conditions are determined with the two-stage method or whether they are the same sampled directly from a Dirichlet prior with shape parameter of 1.

[6] Values above 5 are not advised, as that would pull substantial density away from 0. High values of ζ^{raw} tend to produce β_c that can cluster too strongly near the lowest and highest condition effects. However, this is only a soft recommendation because this parameter was explored only informally.

[7] Note that, unlike Davis-Sober et al., the model that results from our procedure is a union of the allowable orders, rather than a consideration of the implied convex hull.

[8] In all applications of the model in this paper, the same prior that was used for construction of ξ was used in construction of λ.

We note that – as discussed by Kalish et al. (2016) and Prince et al. (2012) – a researcher's prior knowledge about an experiment can reduce the number of orders under consideration by excluding implausible orders. Indeed, the conditions of an experiment whose state-trace will be analyzed are often chosen after careful planning and piloting so as to be sufficiently diagnostic of the models in question. That is, even though the effects of the different conditions are not known in advance of the experiment, certain orders often be safely be assumed to be implausible, and excluding them may be necessary for a fair test of the models in question (Davis-Stober et al., 2016; Prince et al., 2012). For example, in the experiment that led to the development of this technique, the arrangement of two of the four study conditions could easily be predicted before the start of the experiment. More specifically, performance for any condition involving some form of study (Binocular, CFS, Word) should be no worse than the No study condition. Similarly, we can safely assume that performance following Binocular study should be no worse than after CFS or Word study, considering that binocular study provides both the visual details and the object's name. Thus, for the unidimensional model, rather than considering the $4! = 24$ possible orderings, we need only consider $2! = 2$ plausible orders (two arrangements of the Word and CFS conditions, pinning No study in the lowest position and Binocular in the highest position). For the bidimensional model, rather than considering $4!^2 = 576$ possible orderings, we need only consider $2!^2 = 4$ plausible orders (two arrangements of the Word and CFS conditions for each dependent measure, again pinning No study in the lowest position and Binocular in the highest position). This drastically reduces the computational burden when applying the technique. Critically, limitations on possible orders should be determined without reference to actual results.

In summary, dependencies between the two dependent measures are modeled via correlations in population-level subject and item effects, as well as trial-level correlations between the dependent measures that differ based on the condition. This setup reflects the presence of three related but conceptually distinct kinds of correlational structure that may be present in a particular dataset. First, correlations may be present between the two measures across subjects. High, positive correlations in subject effects might correspond to a large role of motivation, whereby certain participants tend to perform well or poorly in both tasks. Second, correlations may be present between the two measures across stimuli (e.g., visual objects that are easy to name from a part are also easy to recognize from a pair of parts). Finally, after controlling for subject and item effects, dependencies may still remain between the outcome variables, owing to, e.g., within-trial tradeoffs (more effort on one task than the other) or between-trial fluctuations (more effort on some trials than others). Modeling all three kinds of dependency simultaneously and with high precision is most efficiently achieved with a hierarchical model (Kruschke, 2014; Rouder & Lu, 2005).

Inference about the latent dimensionality of the cognitive space can proceed through model comparison. Both the unidimensional and bidimensional models are fit (where the two models are differentiated based on the orders of condition effects under consideration). Then, any suitable criterion for picking the best-fitting model can be used.

In the present paper, we chose the Widely Applicable Information Criterion for model comparison (WAIC; Watanabe, 2010). The WAIC is a measure of the predictive accuracy of a model. Performance of the models under comparison is measured in terms of a difference in the expected value of their log of predictive distribution. That is, a model which is chosen by WAIC can be expected to provide better predictions of future data. The significance of this difference can be assessed by comparing an estimate of the standard error of this difference to some pre-determined threshold. In

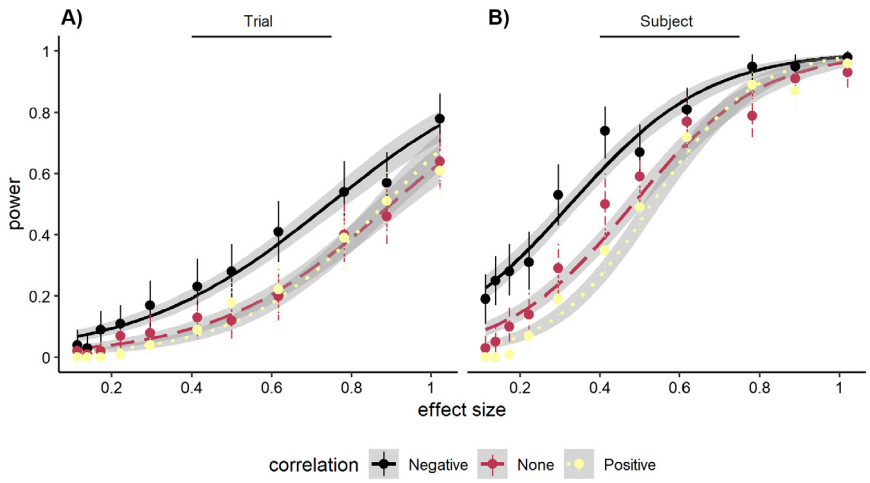

Fig. 3. Power to reject a monotonic function when (A) the data included correlations at the trial level or (B) the subject level. Panels depict power curves for datasets analyzed using the Coupled Monotonic Regression (CMR) technique of Kalish et al. (2016). The model was applied to simulated data that were generated either with trial-level correlation (A) or subject-level correlation (B). Correlations were set to be +.9 (positive), 0 (none), or −.9 (negative). The symbols show the results of each simulated dataset and the lines show a logistic regression. Note that the CMR technique assumes independence between the dependent measures and yet the results from applying CMR are clearly affected by correlations in the data. See text for additional details.

the model comparisons presented here, we chose the frequently-adopted threshold of two standard errors. For further details on the calculation of the WAIC and estimation of the standard errors, we refer readers to Vehtari, Gelman, and Gabry (2017).

2. Model recovery via simulation

Any new statistical analysis technique needs to be validated through experimental and/or simulation studies. In this section, we show that the model can recover ground-truth from simulated data. That is, when the data are generated from a non-monotonic function, the technique picks the bidimensional model, and, conversely, when the data are generated from a monotonic function, the technique picks the simpler, unidimensional model. Our simulations also explored the technique's robustness to different settings for the subject-level and trial-level correlations. Before presenting these model recovery simulations, we provide quantitative evidence of the claim that correlations can bias the results of a State-Trace Analysis.

Fig. 1 suggests that a state-trace analysis that captures the correlation ought to have greater power to reject a monotonic model when the data are positively correlated, for a situation where the general trend across the conditions is a positive increase. Fig. 1 also suggests that when using a model that incorrectly assumes independence, there is little difference between positively and negatively correlated data, although this is an empirical question that we address next (Fig. 1 was generated by hand to outline the situation conceptually). To explore how a model that incorrectly assumes independent data might be affected by correlations in the data, we conducted a power analysis by applying the Kalish et al. (2016) technique to simulated data generated with positive or negative correlations. We simulated 200 non-monotonic datasets with 20 trials from four conditions with means of (1, 2, 3, 4) and (1, 3, 2, 4) along the two dependent measures. Kalish et al. defined a measure of effect size for non-monotonicity based on the arrangement of condition means and the variance of each condition, so in this simulation we manipulated effect size by altering the variance of the data. The generated data did not include variability due to subject or item effects. We then applied the continuous version of the Coupled Monotonic Regression technique of Kalish et al. (2016) to each of these datasets and tallied which ones were rejected at an alpha level of 0.01. In all simulated datasets, the correct result is a rejection of monotonicity, so the proportion of rejections can serve as an estimate for power. A similar power analysis was provided by the authors of the Coupled Monotonic Regression technique, but the simulations here were repeated with correlations of −0.9, 0, and 0.9.

Results of this demonstration are presented in Fig. 3A. Despite assuming independence between the dependent measures, the power of the technique nevertheless depends on the correlation present in the data. In particular, this technique exhibits highest power for negatively correlated data and lowest power for positively correlated data. Notably, this conclusion is the opposite to that suggested by the bottom row of Fig. 1, which indicates that a statistical model that explicitly models the dependency in the data will instead have greater power to reject the monotonic model with positively correlated data (below we confirmed this hypothesis when applying our technique to positively versus negatively correlated data). More generally, this is a simple demonstration that correlations in the data matter, in terms of assessing monotonicity.

Next, we investigated whether the biasing effect of correlation in the data is unique to the trial-level, or whether it can also occur owing to other sources; namely, correlations between how well a subject does on one task versus how well they do on the other task. To assess the impact of subject-level correlation, we repeated the above simulation but varied the correlation present at the level of subjects (and set the trial-level correlation to 0). We again generated 200 datasets at varying levels of effect sizes (altering both trial-level and subject-level variability). The results are plotted in 3B, which shows the same pattern as in 3A.

These simulations support the claim that unacknowledged correlations in the data can bias a state-trace analysis when using a model that assumes independence, with this occurring regardless of the source of the correlation (i.e., for both trial and subject correlations). Thus, it is important that any statistical model address these correlations. Next, we report model recovery simulation results with our proposed technique to assess its ability to recover ground truth with different correlations in the data.

Three pairs of datasets were simulated with different trial-level correlations (Fig. 4), set to be 0, 0.5, and −0.9. Each pair consisted of two simulated datasets from the model described in the previous section: one with an ordering that described a monotonic function, thereby compatible with both models (although this would be consistent with a larger proportion of the allowable orderings under the unidimensional model), and one with an ordering that

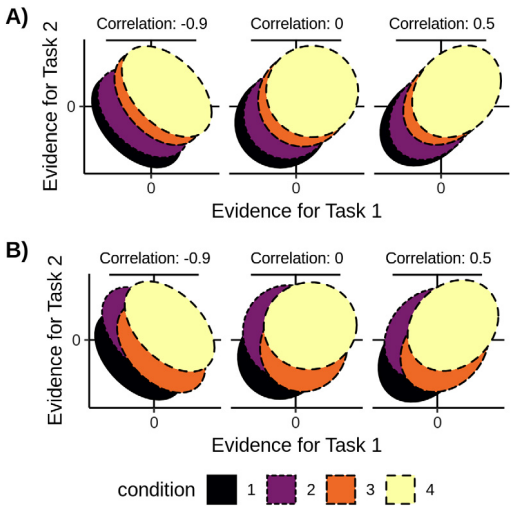

Fig. 4. Simulated data for model recovery. (A) Simulated monotonic and (B) non-monotonic datasets, at each of 3 correlations. Evidence values above 0 produce correct responses for the task corresponding to that evidence dimension. Each dataset included 8000 observations. The ellipses indicate the 95% highest density interval of fitting a bivariate normal distribution to these simulated datasets. The proposed technique recovered the true model in all cases.

could only be explained by a bidimensional model. Each dataset included 50 subjects presented with 40 items in each of 4 conditions (i.e., 8000 observations in total). The remaining parameters for the simulated datasets were specified in the unconstrained space (i.e., as probit values). The condition effects in the monotonic model were $((-1,-1),(-0.5,-0.5),(0.5,0.5),(1,1))$, and for the non-monotonic model they were $((-1,-1),(-0.5,0.5),(0.5,-0.5),(1,1))$. The variability of subject and item effects was set to 0.5 probits (i.e., subject and item effects were sampled from a bivariate normal distribution with no correlation and standard deviation of 0.5 along both dimensions). These parameters produced data from each condition that were highly overlapping (Fig. 4).

As discussed in the last section, not every possible order may be reasonable for a given experiment. Therefore, in these simulations we consider only a subset of the possible orders for each model, namely, only those orders with the lowest and highest condition effects appropriately placed. Again, we emphasize that this decision of which orders to consider should be made during the planning stage of an experiment, before the data have been collected.[9]

All models were fit using the No-U-Turn Sampling algorithm with the R interface to the Stan language (Stan Development Team, 2017a, 2017b). The WAIC was calculated via the loo package (Vehtari, Gabry, Yao, & Gelman, 2018; Vehtari et al., 2017). Six chains were run (starting from random parameter initializations), with 1000 samples of warmup and an additional 1000 samples from the posterior (6000 posterior samples, overall). To assess convergence of the model to the posterior distribution, the *split* $-\hat{R}$ was calculated and chains were monitored for divergences (Betancourt, 2017; Betancourt & Girolami, 2015; Gelman, 2014, Chapter 11; Gelman & Rubin, 1992). The *split* $-\hat{R}$ involves first splitting each chain in half and then comparing the variance within each (split) chain to the variance across each (split) chain for each parameter individually. Chains that sample from disparate regions of the posterior often have a variance ratio much greater than 1. In these model recovery simulations, all *split* $-\hat{R}$ values were below 1.1. Aside from lack of convergence between the chains, another potential problem is a chain that fails to adequately sample from the posterior distribution. Posteriors that are challenging to adequately sample from can be identified as such when constructing chains with Hamiltonian Monte Carlo (or variants such as the No-U-Turn Sampler) by checking for divergences in the simulations that define each iteration of the chain (Betancourt, 2017; Duane, Kennedy, Pendleton, & Roweth, 1987; Neal, 1996). There were no divergences.

Results: In all six comparisons, the correct model was chosen by more than 3 standard errors. This is encouraging, as it suggests that the technique can pick the correct model given sufficient power. To test whether this technique also handles subject correlations, the above analysis was repeated three times, using subject correlations of 0, −0.5, and 0.5. Again, the correct model was chosen by more than 3 standard errors in all cases.

Beyond model comparison, the bottom row of Fig. 1 suggests that a monotonic model that captures correlations in the data should better fit negatively correlated data than positively correlated data, provided that the data have non-monotonically arranged means that are positively increasing (notably, this is the opposite of the results when applying CMR; see Fig. 3). To assess this, we compared the WAIC of the unidimensional model to non-monotonic data for the situation of no trial-level correlations, but subject-level correlations set to −.5, 0, or +.5. As expected, the WAIC for the unidimensional model applied to negatively correlated data indicated better predictive performance as compared to the situation with positively correlated data, with a difference in WAIC scores between negative and positively correlated data of approximately 1 standard error. Thus, as compared to CMR, assumptions of independence versus allowing dependencies produce seemingly opposite results in terms of goodness of fit of the monotonic model, as a function of correlations in the data. Somewhat surprisingly, the largest change across the different levels of correlation in the data was that, for uncorrelated data, the WAIC value for the unidimensional model was nearly 40 standard errors worse than for the correlated data. However, on further reflection, this is sensible, demonstrating that the model can capitalize on correlations when they exist: as the correlation approaches 1 (either positive or negative), the outcomes become more highly constrained, resulting in better generalization (i.e., better WAIC), but with uncorrelated data, a wider range of mean values work reasonably well, resulting in worse generalization. This again underscores that correlations in the data ought to affect the inferences made in a state trace analysis.

One advantage of the proposed technique is that, in addition to mitigating bias due to correlation between the dependent measures, that correlation can be estimated. For example, the most posterior distributions of the trial-level correlation parameters straddled the values that were used to generate the data (Fig. 5). While there were a few failures where the posteriors for each correlation parameter differed between conditions or missed the true value, these failures tended to occur when the model was not matched to the data. For example, the unidimensional model exhibits spurious differences between the correlations across condition means when the data were nonmonotonic and the true correlation was 0.5 (Fig. 5, bottom right).

3. State-trace analysis applied to real data

To assess the method using real data, we applied the proposed technique to a recently collected dataset (Figure 6; Sadil

[9] One adjustment was made in the model recovery simulations, relative to the presentation of the model in the previous section. In these simulations, the $\beta^{intercept}$ parameter was constrained in the fitting routine to always be less than the β^{raw}, for each order. This contrasts with the description of the model in the previous section, and in the application to real data. In practice, if there is a baseline condition with a probit value below 0 (as there was during model recovery), this constraint will not have a substantial effect on the posterior parameter estimates.

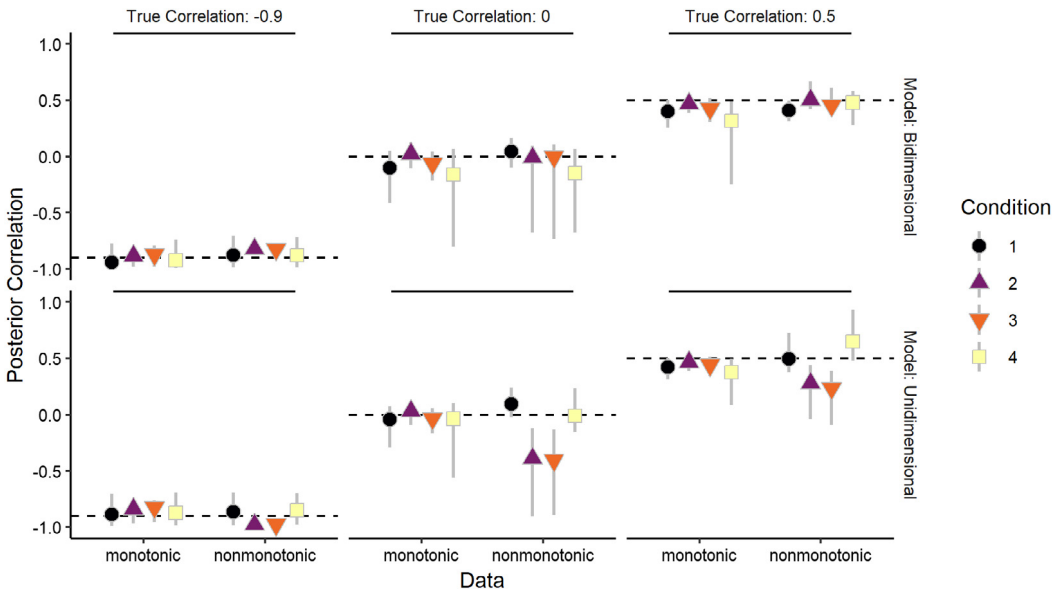

Fig. 5. Recovery of simulated trial-level correlation. Posterior distributions (median and 95% highest density interval) of each of the four trial-level correlation parameters (one for each condition) across the simulated datasets. The horizontal axis is grouped by whether the data were generated with a monotonic or non-monotonic function, and the top and bottom rows show the posteriors after fitting with the bidimensional and unidimensional models, respectively (the unidimensional model is nested within the bidimensional model). Most posterior distributions straddle the value that was used to generate the data (dashed line), but the largest misses occur when the data were not matched to the model.

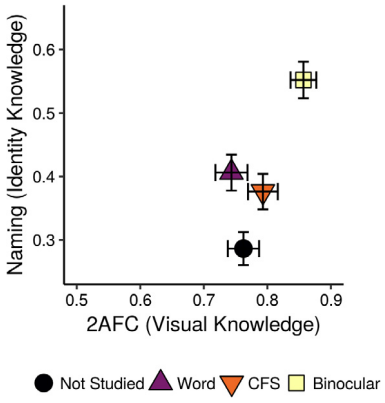

Fig. 6. Real data for analysis by proposed technique. Average performance on the two tasks in the four conditions. Error bars indicate within-subject Confidence Intervals, including correction by Morey (2008). 48 participants contributed 32 trials in each of the 4 conditions. For further details, see Sadil et al. (submitted).

et al., submitted). In contrast to the exclusively monotonic or non-monotonic effects used during model recovery, this application to real data revealed identifiability problems that can arise when data are not clearly monotonic or non-monotonic. In this section, these challenges are explored, and a potential solution is presented.

As in the simulated example, we compared restricted versions of the unidimensional and bidimensional models. Specifically, in the unidimensional and bidimensional models, the only orderings considered were the ones in which subjects performed worst in the Not Studied condition and best in the Binocularly studied condition. Ten chains were run, using 1000 samples of warmup and 500 posterior draws, each. The unidimensional model exhibited good convergence (as measured by *split* $-\hat{R}$ and lack of divergences), but the bidimensional model failed the *split* $-\hat{R}$ diagnostic. In particular, the chains sampled disparate arrangements of condition means (β_c), resulting in a highly multimodal posterior distribution (each arrangement of condition means resulted in a mode that chains did not transition between). Although multiple modes are not necessarily problematic to Bayesian inference, the presence of multiple modes will frustrate the ability of the No-U-Turn Sampler, and the resulting draws from the chains are not guaranteed to adequately represent the posterior distribution. Model comparison would be inappropriate without better convergence.

Further investigation revealed that this multimodality reflected a form of model mimicry. Mimicry occurs because data that conform to a non-monotonic arrangement of means can be captured through a mixture of different monotonic orders (or vice versa: a monotonic order could reflect a mixture of non-monotonic orders). In our model, this can occur even within an individual considering that the mixing across orders occurs at the trial level. We present a simulation to illustrate model mimicry when dealing with a mixture across orders. To simplify the example, the data were generated without any subject- or item-effects. Additionally, the data were continuous-valued (i.e., generated from a bivariate normal), and the trial-level correlation was set to 0.9. The condition means were non-monotonically arranged, with values of $((-2, -2), (-1, 1), (1, -1), (2, 2))$. Each condition had 600 trials and the data are shown in Fig. 7A. The monotonic model was fit to these data, allowing only for the two monotonic orders in which the first and fourth conditions were fixed as the best and worst, respectively. Simulated data were then sampled from the posterior distribution. This resulted in a posterior predictive distribution, shown in Fig. 7B. Although none of the mixing components in the unidimensional model included non-monotonically arranged condition means, our implementation of the unidimensional model is nevertheless able to mimic non-monotonic data. It achieves this by determining that both of the orders are equally probable (the mixing parameter was highly concentrated around 0.5). Generatively, this can be understood as the model determining that each order is equally likely for the given trial. Across samples in the posterior predictive distribution, the model therefore produces each order equally often. Although the model is unable to capture the mode of conditions 3 and 4, the positive correlation in the data enabled the model to place predictions on high density regions of the data.

Application of the bidimensional model can therefore result in a kind of non-identifiability. When the data are non-monotonic, the bidimensional model can account for this by either (A) placing high

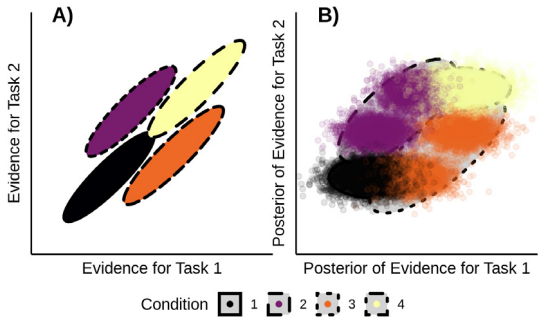

Fig. 7. Flexibility of monotonic model as applied to non-monotonic data. Continuous data for which the means are non-monotonically arranged (A) were fit with a unidimensional model, the posterior predictive distribution of which is shown in (B). Although none of the components of this model contain a non-monotonic order, the model can reproduce a degree of non-monotonicity by placing equal weight on the different allowable orderings; the model can either generate from the lower left cloud of condition 2 and the upper right cloud for condition 3, or the lower right cloud in condition 3 and the upper left cloud for condition 2. Because the data are positively correlated, the model effectively captures much of the data with a mixture of these two arrangements. Ellipses in (B) show the estimation of a 2d-gaussian fit to the posterior samples.

probability on the mixture component associated with the non-monotonic order, or (B) placing approximately equal probability on the mixture components associated with the two monotonic orders. The reverse situation could occur if the data were monotonic. Moreover, the values of the condition mean parameters (within each order) that are deemed likely will depend on which orders are determined to be likely. This trade-off between condition means and order probabilities represents the source of the multimodality observed when the bidimensional model was applied to real data. Presumably, this non-identifiability was not observed in the model recovery simulations due to the relatively large effect sizes (magnitude of non-monotonicity) that were modeled, and the relative uniformity of the simulated subjects (all subjects were simulated to follow the same set of condition means).

The notion that combinations of data that are monotonically arranged can appear non-monotonic was explored by Davis-Stober et al. (2016) and Prince et al. (2012), under the assumption that different subjects have different orders. However, our model suggests the possibility of something more pernicious, with mixtures of orders occurring within an individual. Our first attempt to solve the convergence problem for the bidimensional model relaxed the fixed effects assumption, allowing that different subjects have different mixtures across the orders, with different conditions means for each subject (see the Appendix for further details). However, this decoupling of subjects only exacerbated the convergence problem for the bidimensional model.

A more straightforward modification to the proposed technique is to compare models that contain only monotonic or non-monotonic orders, rather than using nested modeling between the unidimensional and bidimensional models. That is, rather than comparing the unidimensional model (which has two allowed orders for the present dataset) against the bidimensional model (which has four allowed orders), one could compare the unidimensional model to a model with the remaining two orders (i.e. only the non-monotonic orders). Indeed, application of the proposed mixture model exhibits good convergence for both models and the results suggested that the models faired equally well (i.e., there was no clear winner). However, as outlined in the introduction, our theoretical question of interest is not monotonic versus non-monotonic but rather the dimensionality underlying the data. In light of this goal, *any* reliable instance of non-monotonicity (e.g., clearly non-monotonic data for some subjects) could falsify the unidimensional model, even if the dataset as a whole is equivocal when comparing the monotonic and non-monotonic models. Thus, this comparison between the monotonic and non-monotonic models does not differentiate between the unidimensional and bidimensional models. For instance, it is not clear whether equivalence between the monotonic and non-monotonic models indicates that the data were insufficient for determining the latent dimensionality (e.g., noisy data or too few observations) or whether this indicates that some aspects of the data (e.g., some subjects) were non-monotonic, while other aspects were monotonic (e.g., other subjects). If the latter is true, this would favor the bidimensional model.

Considering the convergence problems for the bidimensional model when mixing across orders, we instead formulated the model comparison in terms of each order on its own merits (i.e., no mixing of orders, which eliminated the λ parameters), asking whether each order provided the best account of the dataset as a whole and also the extent to which each subject was best explained by that order. Because each order is an instance of monotonicity or non-monotonicity, this allowed an assessment of whether the data contained specific instances of non-monotonicity. In the current case, there were four models, where the pairs of condition means for No Study and Binocular are constrained to be lowest and highest, respectively, and each model specifies a particular ordering of the two middle conditions for the two dependent variables. In changing the model comparison to a comparison amongst orders, we allowed that each subject has their own set of condition means under the constraints of that order (see Appendix).

In addition to comparing the four orders using WAIC, we also compared the different orders using the Pseudo Bayesian Model Averaging Plus (PseudoBMA+) method proposed by Yao, Vehtari, Simpson, and Gelman (2018). This technique is similar to AIC model weights in that it estimates the probability that a model will best predict future data, given the observed data (Akaike, 1978; Wagenmakers & Farrell, 2004). However, the estimation procedure is substantially different, not least because it relies on the full posterior distribution and the weights are regularized based on the uncertainty of the estimated predictive ability of a model. PseudoBMA+ weights were calculated with the R package, loo 2.0.0 (Vehtari et al., 2018). Use of the PseudoBMA+ allowed us to compare the four models for each subject, even though the models were applied to the entire dataset.

Models that allowed for each of the four orders were fit separately using 10 chains that were allowed 1000 warmup samples and 500 samples from the posterior. The PseduoBMA+ weights revealed substantial support for the order that matched the raw condition means. Specifically, in the supported model, the condition means for the 2AFC task were ordered as No Study $<$ Word $<$ CFS $<$ Binocular, while the condition means for the Naming task were ordered as No Study $<$ CFS $<$ Word $<$ Binocular. This represents a non-monotonic arrangement, and it was given a PseudoBMA+ weight of 0.99. The WAIC scores between this most preferred order and the second most preferred order (a monotonic order in which the condition means followed Not Studied $<$ CFS $<$ Word $<$ Binocular for both dependent variables) revealed a difference of 2.8 standard errors.

The primary concern that prompted development of this technique was that the presently available methods do not account for correlations between the dependent variables, and that failing to acknowledge these dependencies may bias the model comparison process. One advantage of the proposed technique is that the utility of modeling these dependencies can be assessed by comparing a model with and without these correlations. A comparison of the WAIC for the best-fitting order to that same order with correlation parameters fixed to 0 revealed that the model with correlation parameters was preferred by 7 standard errors.

4. Limitations and further modeling considerations

In this section we briefly mention extensions and discuss limitations of the proposed technique. One extension is application to continuous data rather than binary decisions. This could be achieved with a bivariate normal distribution, rather than a bivariate probit. Most of the remaining structure would be the same, except that the variance on the priors would need to be increased and the model would require a prior on the trial-level variance.

One limitation of the proposed technique is its computational demands. On a modern though not state-of-the-art computer (2.40 GHz), the longest running chain required 2 h to finish. In comparison, the technique of Kalish et al. (2016) only takes a few seconds to run and the Bayesian approach of Davis-Stober et al. (2016) is also quite fast. Although the slow speed of the proposed technique is acceptable for comparison between a small number of orders, this computational burden makes it challenging to explore whether the results are robust to model assumptions (e.g., to assess the impact of different priors). Likewise, since each order must be modeled separately, designs that involve more conditions may require substantial computational power. This increase with the number of conditions is unfortunate, given that the inclusion of many conditions can benefit a state-trace analysis by more fully mapping out the functions along each dependent measure. However, advances in computer technology (e.g., GPUs) may ultimately overcome the computational limitations of this technique. Moreover, access to a computing cluster can be beneficial, as each chain and each model can be run in parallel.

Another potential limitation may arise if different participants have different orders. Our model comparison between orders allowed that each participant could have their own set of conditions means, but those means were subject to the same order constraint for all participants. Additional model recovery studies are needed to assess whether the proposed technique is robust to violations of this fixed order constraint. In the meantime, the current results can be used to assess heterogeneity across subjects by calculating a PseudoBMA+ weight for each participant (Fig. 8). That is, rather than calculating a single weight for each model based on the log-likelihood of all trials across all subjects, one can calculate a weight for each subject from the log-likelihoods that correspond to trials contributed by that subject. Doing so provides an approximate way to ask which order will be best able to predict new data for each subject. This approach is more constrained than fitting each subject separately in that: (1) item effects can only be determined by modeling the dataset across subjects; (2) pooling data across subjects provides increased precision when estimating the trial-by-trial dependencies in Ω_c, which are assumed to be the same for all subjects; and (3) correlations between the two measures across subjects will induce shrinkage in subject-specific condition means.

The order of subjects for both panels of Fig. 8 is according to how strongly the data of each subject favored the 2 non-monotonic orders as compared to the 2 monotonic orders, as seen in Fig. 8A. Although there are some participants whose data seem to be predictable with just a single order (Fig. 8B), the results favoring the first order (the non-monotonic order seen in the condition means) are not as clear-cut as they were for the dataset as a whole, which exhibited a PseudoMBA+ weight of.99 for this order. This is unsurprising, and likely reflects the relatively low trial count (32) for each condition for a given subject. Across these subject differences, the first order, as shown by the black bars, fared better than the other orders more often than would be expected by chance, and collectively across the dataset this provided strong support for this ordering of the condition means.

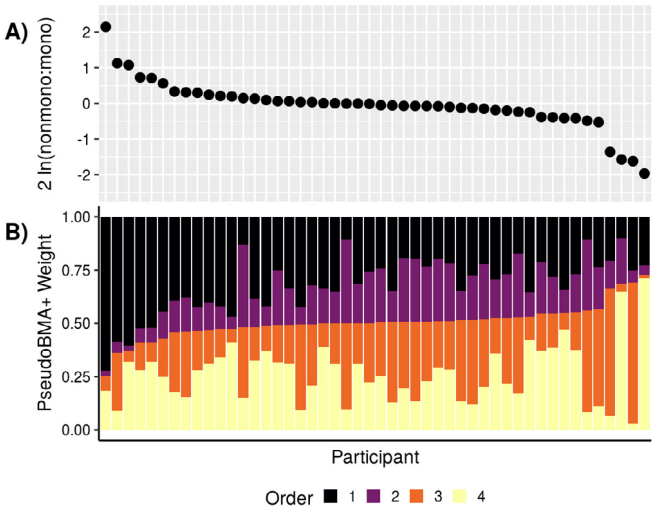

Fig. 8. Subject-level PseudoBMA+ Weights. (A) $2log_e$(odds) of non-monotonic to monotonic orders, for each participant. Higher values indicate larger support for a non-monotonic order. (B) PseudoBMA+ weights for each of the four orders, with the list of participants in the same sequence as in panel A. The first and second orders are non-monotonic, with condition means in the 2AFC task following Not Studied < Word < CFS < Binocular and Not Studied < CFS < Word < Binocular, respectively, and following Not Studied < CFS < Word < Binocular and Not Studied < Word < CFS < Binocular in the Naming Task (see Fig. 6 for data). The third and fourth orders are non-monotonic, with the condition means following Not Studied < Word < CFS < Binocular and Not Studied < CFS < Word < Binocular, respectively.

5. Conclusion

State-trace analyses are a powerful, relatively non-parametric method for distinguishing between models that differ in latent dimensionality (e.g., Bamber, 1979; Dunn et al., 2014; Prince et al., 2012). However, despite making few assumptions, current state-trace analysis techniques are not completely assumption free. Specifically, current approaches assume independence between the dependent variables. The first goal of this paper was to highlight that correlations between the dependent measures of a state-trace analysis can critically influence the analysis (i.e., they can produce model selection biases when incorrectly assuming independence). The second goal was to show that this assumption can be relaxed by modeling the joint distribution of data in a hierarchical framework. Explicitly modeling the joint distribution of the data avoids bias in the analysis, enables estimation of the dependencies, and can increase the sensitivity to uncover the underlying non-monotonicity of the data.

It is important to clarify which aspects of the technique introduced here are essential and which are nonessential. Central to the technique was a model of the joint distribution of dependent variables in a hierarchical framework, to avoid the assumption of independence between the dependent variables. Uncertainty in the monotonicity of the true model was initially instantiated through a mixture of possible orderings of condition effects, but model comparison ultimately required fitting each plausible order individually. Other details are nonessential. For example, the model was fit using Stan's No-U-Turn Sampler algorithm, but any suitably robust algorithm could have been applied. Similarly, model evaluation proceeded via comparison of PseudoBMA+ and WAIC, but other measures could be suitable. Likewise, the choice of a bivariate probit was nonessential; other distributions that provide a reasonable description of the data could be utilized (e.g., a bivariate normal distribution if the data are continuous).

A final point to consider is that for some datasets, the independence assumption may be adequate even if untrue. For instance,

if one adopted the technique presented here but set the correlation parameters to be zero, the posterior distribution would still capture positive or negative dependencies between the subject-, item-, or trial-effects, if the data were sufficient to overwhelm this prior assumption of independence. We were motivated to develop this technique because our experimental design could easily have induced correlations between the dependent measures and because the number of data points collected for each individual was relatively low (insufficient for overwhelming the prior), given the limited number of available items (items can only be used once per participant in this learning task). Furthermore, the power analyses we conducted by applying the Kalish et al. (2016) technique to simulated data revealed that unacknowledged dependencies in the data can cause biases. Accordingly, we developed a method that captures dependencies at the subject, item, trial levels, and we demonstrated that this technique can recover ground truth values of these dependencies across a range of simulated values.

Funding

The work was supported by NSF BCS-1431147 (Huber), NSF CAREER Award 1554871 (Cowell), and NIMH RF1MH114277 (Cowell and Huber).

Appendix

In this appendix, a modification to the model is presented that allows for modeling of subjects with individualized mixing proportions and condition effects. In this model, the condition effects (β_c) and subject-effects are removed. instead, a separate $\beta^{intercept}$, β^{raw}, and ζ is drawn for each subject, s. The resulting, subject-dependent, condition means are then defined as in Eq. (9):

$$\beta_{c,s} = \beta_s^{intercept} + \beta_s^{raw} \sum_{i=1}^{n\_conditions-1} \zeta_{i,s} \qquad c \in 2, \ldots, n\_conditions - 1 \tag{A.1}$$

As before, each β^{raw} is sampled from a univariate normal distribution, so the β^{raw} values across tasks are sampled independently. Any suitable prior can be used for the ζ, though they must sum to 1 for each subject. The $\beta^{intercept}$ are drawn from bivariate normal distributions.

$$\beta_s^{intercept} \sim N\left(\begin{pmatrix}\eta_x\\ \eta_y\end{pmatrix}, \begin{pmatrix}\tau_x & 0\\ 0 & \tau_y\end{pmatrix} \Omega \begin{pmatrix}\tau_x & 0\\ 0 & \tau_y\end{pmatrix}\right) \tag{A.2}$$

As in the main text, sampling is done with the non-centered parameterization. In contrast to the model in the main text, the means of these subject effects, η_x and η_y are not fixed to 0, but are instead given separate univariate normal priors, $N(0, 0.5^2)$ — the same prior as the $\beta^{intercept}$ received in the main text. The scale parameters, τ_x and τ_y, are given gamma priors with shape of 2 and rate of 4. Item-effects are sampled in the same way as described in the main text. To be explicit, the resulting trial-level mean is then defined by

$$\mu_{c,i,s} = item\_intercept_i + \beta_{c,s}, \tag{A.3}$$

The final requirement is that the mixing parameters, λ, be defined uniquely for each subject. The same priors described in the main text can be used to construct a λ for each subject. This results in a model that assigns probability of a given outcome according to

$$p\left(\begin{pmatrix}x\\ y\end{pmatrix}_{c,s,i} | \begin{pmatrix}\mu_x\\ \mu_y\end{pmatrix}_{c,s,i}, \Omega_c\right) = \sum_{j=1}^{n\_orders} \lambda_{j,s}\psi^2\left(\begin{pmatrix}x\\ y\end{pmatrix}_{c,s,i} | \begin{pmatrix}\mu_x\\ \mu_y\end{pmatrix}_{j,c,s,i}, \Omega_c\right). \tag{A.4}$$

Eq. (A.4) differs from 12 only in the presence of a subscript s on the λ and in how the means, $\begin{pmatrix}\mu_x\\ \mu_y\end{pmatrix}_{j,c,s,i}$ were constructed (i.e., with Eqs. (A.1)–(A.3)).

References

Akaike, H. (1978). On the likelihood of a time series model. *The Statistician*, *27*(3/4), 217. http://dx.doi.org/10.2307/2988185.

Bamber, D. (1979). State-trace analysis: A method of testing simple theories of causation. *Journal of Mathematical Psychology*, *19*(2), 137–181. http://dx.doi.org/10.1016/0022-2496(79)90016-6.

Betancourt, M. (2017). A Conceptual Introduction to Hamiltonian Monte Carlo. Retrieved from http://arxiv.org/abs/1701.02434.

Betancourt, M., & Girolami, M. (2015). Hamiltonian Monte Carlo for hierarchical models. In S. K. Upadhyay, U. Singh, D. K. Dey, & A. Loganathan (Eds.), *Current trends in Bayesian methodology with applications*. CRC Press, Retrieved from http://arxiv.org/abs/13120906.

Bürkner, P.-C. (2017). Brms : An R package for Bayesian multilevel models using stan. *Journal of Statistical Software*, *80*(1), http://dx.doi.org/10.18637/jss.v080.i01.

Bürkner, P. C., & Charpentier, E. (preprint). Monotonic effects: A principled approach for including ordinal predictors in regression models. PsyArXiv Preprints, 1–20. http://dx.doi.org/10.31234/OSF.IO/9QKHJ.

Davis-Stober, C. P., Morey, R. D., Gretton, M., & Heathcote, A. (2016). Bayes factors for state-trace analysis. *Journal of Mathematical Psychology*, *72*, 116–129. http://dx.doi.org/10.1016/j.jmp.2015.08.004.

Duane, S., Kennedy, A. D., Pendleton, B. J., & Roweth, D. (1987). Hybrid monte carlo. *Physics Letters. B*, *195*(2), 216–222.

Dunn, J. C. (2008). The dimensionality of the remember-know task: A state-trace analysis. *Psychological Review*, *115*(2), 426–446. http://dx.doi.org/10.1037/0033-295X.115.2.426.

Dunn, J. C., & James, R. N. (2003). Signed difference analysis: Theory and application. *Journal of Mathematical Psychology*, *47*(4), 389–416. http://dx.doi.org/10.1016/S0022-2496(03)00049-X.

Dunn, J. C., & Kalish, M. L. (2018). *State-trace analysis*. Springer.

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 (2014). *Psychonomic Bulletin & Review*, *21*(4), 947–954. http://dx.doi.org/10.3758/s13423-014-0637-y.

Dunn, J. C., & Kirsner, K. (1988). Discovering functionally independent mental processes: The principle of reversed association. *Psychological Review*, *95*(1), 91–101. http://dx.doi.org/10.1037/0033-295X.95.1.91.

Gelman, A. (2014). How do we choose our default methods? In *Past, present, and future of statistical science* (pp. 293–301). Chapman and Hall/CRC, http://dx.doi.org/10.1201/b16720-31.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472. http://dx.doi.org/10.1214/ss/1177011136.

Greene, W. H. (2017). *Econometric analysis* (8th ed.). Pearson.

Jang, Y., Lee, H., & Huber, D. E. (2019). How many dimensions underlie judgments of learning and recall redux: Consideration of recall latency reveals a previously hidden non-monotonicity. *Journal of Mathematical Psychology*, http://dx.doi.org/10.1016/j.jmp.2018.10.006.

Kalish, M. L., Dunn, J. C., Burdakov, O. P., & Sysoev, O. (2016). A statistical test of the equality of latent orders. *Journal of Mathematical Psychology*, *70*, 1–11. http://dx.doi.org/10.1016/j.jmp.2015.10.004.

Kruschke, J. K. (2014). Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan, second edition. Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan, Second Edition. http://dx.doi.org/10.1016/B978-0-12-405888-0.09999-2.

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*(9), 1989–2001. http://dx.doi.org/10.1016/j.jmva.2009.04.008.

Loftus, G. R., Oberg, M. A., & Dillon, A. M. (2004). Linear theory, dimensional theory, and the face-inversion effect. *Psychological Review*, *111*(4), 835–863. http://dx.doi.org/10.1037/0033-295X.111.4.835.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc..

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology*, *4*(2), 61–64. http://dx.doi.org/10.3758/s13414-012-0291-2.

Neal, R. M. (1996). Bayesian Learning for Neural Networks, 118. http://dx.doi.org/10.1007/978-1-4612-0745-0.

Pan, K. (2017). An analytical expression for bivariate normal distribution. *SSRN Electronic Journal*, http://dx.doi.org/10.2139/ssrn.2924071.

Pratte, M. S., & Rouder, J. N. (2012). Assessing the dissociability of recollection and familiarity in recognition memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *38*(6), 1591–1607. http://dx.doi.org/10.1037/a0028144.

Prince, M., Brown, S., & Heathcote, A. (2012). The design and analysis of state-trace experiments. *Psychological Methods*, *17*(1), 78–99. http://dx.doi.org/10.1037/a0025809.

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. http://dx.doi.org/10.3758/BF03196750.

Sadil, P., Potter, K., Huber, D. E., & Cowell, R. A. (submitted). Connecting the dots without top-down knowledge: Evidence for the separability of levels within the visual processing hierarchy.

Stan Development Team. (2017a). RStan: the R interface to Stan. Retrieved from http://mc-stan.org.

Stan Development Team. (2017b). Stan Modeling Language Users Guide and Reference Manual. Retrieved from http://mc-stan.org.

Tsuchiya, N., & Koch, C. (2005). Continuous flash suppression reduces negative afterimages. *Nature Neuroscience*, *8*(8), 1096–1101. http://dx.doi.org/10.1167/4.8.61.

Vehtari, A., Gabry, J., Yao, Y., & Gelman, A. (2018). loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models. Retrieved from https://cran.r-project.org/package=loo.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. http://dx.doi.org/10.1007/s11222-016-9696-4.

Wagenmakers, E., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*(1), 192–196. http://dx.doi.org/10.3758/BF03206482.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research (JMLR)*, *11*, 3571–3594.

Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018). Using stacking to average bayesian predictive distributions (with discussion). *Bayesian Analysis*, *13*(3), 917–1003. http://dx.doi.org/10.1214/17-BA1091.