Journal of Mathematical Psychology 108 (2022) 102655

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

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

State trace analysis: What it can and cannot do☆

F. Gregory Ashby [a,*], Donald Bamber [b]

[a] University of California, Santa Barbara, United States of America
[b] University of California, Irvine, United States of America

ARTICLE INFO

Article history:
Received 16 June 2021
Received in revised form 2 November 2021
Accepted 28 February 2022
Available online xxxx

Keywords:
State-trace analysis
Monotonic state-trace model
Single-versus multiple systems
Dissociations

ABSTRACT

State-trace analysis (STA) is a method for determining the number of underlying parameters or latent variables that are varying across two or more tasks. STA is based on the fact that under very weak conditions, any model in which r parameters are varying across r or more tasks predicts an r-dimensional state-trace plot. Although monotonicity assumptions can sometimes be useful in STA, they are not required. Specifically, there is no need to assume that performance in any task is a monotonic function of whichever parameters are varying. As a result, requiring STA models to assume monotonicity seriously reduces the applicability of STA. Whereas an STA can identify the number of varying parameters, it provides no information about the number of underlying systems. Similarly, STA is ill suited to examining dissociations. It can be used to test for double, but not single dissociations. In particular, a monotonic state-trace plot rules out a double dissociation but provides no information about whether or not the data contain a single dissociation.

1. Introduction

State-trace analysis (STA; Bamber, 1979) is a method for determining the complexity of a set of data in which one or more independent variables (IVs) are manipulated across experiments or conditions and, most typically, two separate dependent variables (DVs) are measured (e.g., performance in two tasks). If the complexity of the data is less than the potential variation in the IVs and DVs, then some sort of bottleneck must exist. Although a more precise description is given below, informally, the goal of STA is to measure the width of this bottleneck. The bottleneck between the IVs and DVs is due, presumably, to perceptual and cognitive processes internal to the human participant. As such, an accurate model of those processes should therefore also include the same bottleneck. In the case of mathematical models, the width of the bottleneck is typically defined by the number of parameters the model must vary to account for the data.

Fundamentally, STA is a method for determining the complexity or dimensionality of perceptual and cognitive processes that are recruited across a variety of tasks and conditions. In our opinion, it is the best available method for addressing this problem. However, during the past several decades, STA has been used for a variety of other purposes — in particular, to ask questions about the architecture of the underlying processes, that is, about whether the perceptual and cognitive processes are configured as a single system or as multiple systems, and also to ask whether the data from the various conditions provide empirical support for some kind of dissociation. As we will see, STA is poorly suited to both of these problems.

This article describes the mathematical basis of STA, with the goal of improving its current application. Some of the results presented here have been described previously (i.e., Propositions 2 and 4). However, Propositions 1, 3 and 5 are new. In particular, we extend STA from its usual two-task applications to any number of tasks, and we present new results on the inability of STA to identify the number of underlying cognitive systems or a possible dissociation between performance in two tasks. This article proceeds as follows. Section 2 establishes the mathematical foundations of STA, and extends the method to any number of tasks. Section 3 shows that a popular restriction of STA to state-trace plots that are monotonically increasing or decreasing fails to exploit the full potential of STA. Section 4 establishes the inability of STA to discriminate between single- and multiple-systems models that predict the same width bottleneck. Section 5 provides rigorous justification for using STA to test for double dissociations, but also shows that STA provides no information about whether or not the data contain a single dissociation. Finally, we close with some general conclusions.

2. The mathematical foundations of STA

STA begins by plotting performance on two DVs against each other. The DVs may come from the same or different tasks. For

☆ We thank Robin Thomas, Andrew Heathcote, and an anonymous reviewer for their helpful comments.

* Corresponding author.
E-mail address: fgashby@ucsb.edu (F.G. Ashby).

https://doi.org/10.1016/j.jmp.2022.102655

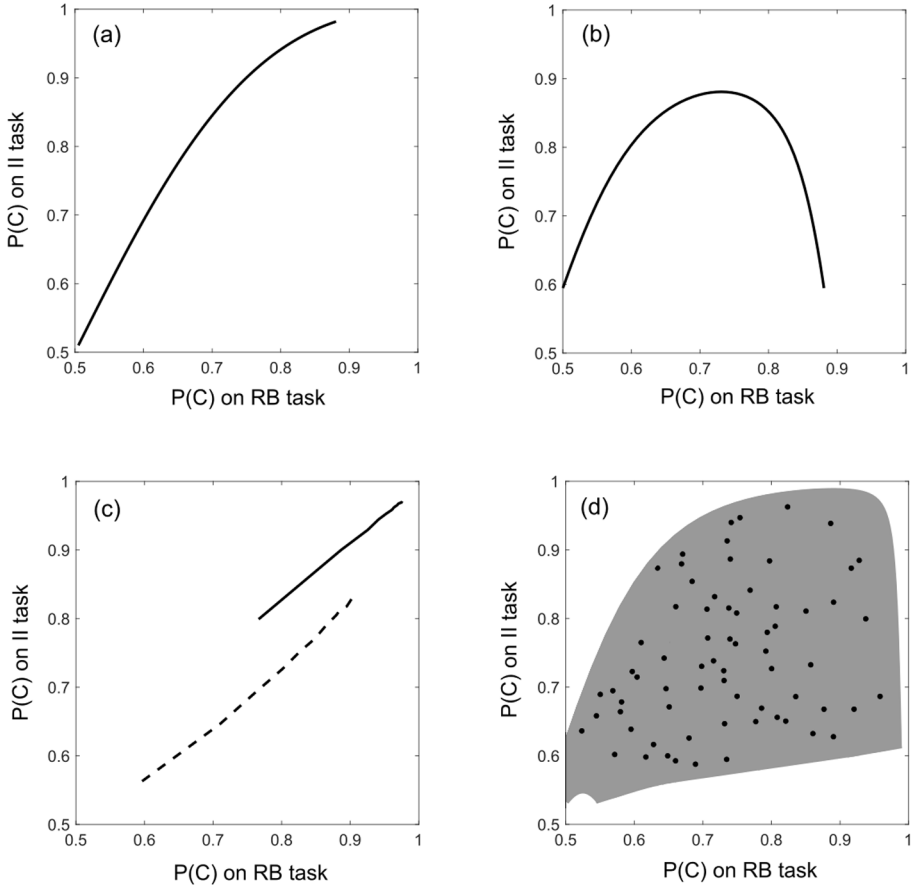

Fig. 1. Different types of state-trace results. All four panels were generated from the generalized context model (GCM; Nosofsky, 1986) in rule-based (RB) and information-integration (II) categorization tasks. The single-monotonic curve in panel (a) was generated by assuming that the GCM overall discriminability parameter c varies across tasks and participants. The single-nonmonotonic curve in panel (b) was generated by assuming that the GCM attention weight parameter w varies across tasks and participants. The double curves in panel (c) show a possible outcome of an experiment with two groups of participants, in which the parameter c varies continuously within each group and the two groups are characterized by different values of w. The scatter plot in panel (d) was generated by assuming that both c and w vary across tasks and participants.

example, an STA could be performed on an ROC curve, which plots the probability of a hit against the probability of a false alarm from a YES–NO detection task. Alternatively, the STA could be performed on data collected from two different categorization tasks, where the two DVs are the proportion of correct responses in each task (as in Figs. 1–3).

2.1. Different types of state-trace plots

To begin, we describe the various possible outcomes of an STA, which are all illustrated in Fig. 1.

Definition 1 (*Types of State Traces*)**.** Consider an STA that plots values of two DVs against each other. The plot that results includes the following types.

- In a *single-monotonic* plot, the data all fall on a single monotonic curve — that is, a curve that is either nondecreasing or nonincreasing (e.g., as in Fig. 1a);
- In a *single-nonmonotonic* plot, the data all fall on a single nonmonotonic continuous curve (e.g., as in Fig. 1b);
- In a *double* plot, the data fall on two separate curves — that is, they do not all fall on a single continuous curve (e.g., as in Fig. 1c);
- In a *scatter* plot, the data fill a region and do not fall on any one or two curves (e.g., as in Fig. 1d).

2.2. Mathematical models

All applications of STA assume that the IVs manipulated in an experiment and the DVs that are recorded are related via one or more latent or intervening variables. In psychology, latent variables are often interpreted broadly and include constructs such as hunger, personality, or intelligence. Many such latent variables are surely multidimensional, in the sense that any serious model of these broad constructs would likely include more than one parameter. The best that STA can do is determine the number of dimensions across which the latent variable or variables are varying. For example, if an STA concludes that the latent variables vary on two dimensions, it is impossible to know, on the basis of the STA alone, whether the two dimensions describe two different latent variables, or represent a single latent variable that varies on two dimensions. Therefore, to be clear, we refer to the dimensions across which the latent variables vary as *parameters*. So in an application to an experimental setting in which there are m IVs, r parameters, and n DVs, the goal of STA is to identify the numerical value of r.

Consider any of the state-trace plots described in Definition 1. Suppose that two IVs are manipulated and that performance in both tasks is mediated by a single intervening parameter θ. In this case, there must exist some function f that determines the value of θ for every possible set of values of IV_1 and IV_2 – that is, $f(IV_1, IV_2) = \theta$. Furthermore, there must also exist functions g_1 and g_2 that map θ to DV_1 and DV_2, respectively. In other words,

$g_1(\theta) = DV_1$ and $g_2(\theta) = DV_2$. We call f the *input function* and g_1 and g_2 the *output functions* (i.e., as in Dunn & Anderson, 2018).

Determining the functions f, g_1, and g_2 from the empirical state-trace plot is beyond the scope of STA. These functions could be highly complex. Future research might approximate them, but in almost all applications of STA, they should be considered unknowable. The contribution of STA to this problem is not to estimate these functions, but to identify the existence of the bottleneck — that is, to conclude that performance in the tasks must be mediated by some single varying parameter.

Although the functions f, g_1, and g_2 are unknowable, a common goal of researchers, especially mathematical psychologists, is to propose a mathematical model of the perceptual and cognitive processes thought to be active in the tasks under study. A fully-specified model should define all three functions f, g_1, and g_2, and describe the bottleneck by proposing a parameter that models the single varying dimension of the latent variable. In practice however, few if any, current models within mathematical psychology meet these goals. Current mathematical models specify the free parameters that define the latent variable space and they specify output functions that generate predicted values of various DVs for any specified set of numerical parameter values. However, they typically make weak, or no assumptions about the input functions. We call such models output-specified models.

The following definitions formalize these ideas.

Definition 2 (*Fully-specified Mathematical Model*)**.** Consider a set of one or more tasks in which m real-valued IVs (i.e., $IV_1, \ldots, IV_m$) are varied and n real-valued DVs (i.e., $DV_1, \ldots, DV_n$) are recorded. A *fully-specified mathematical model* of these tasks specifies how the value of each of the DVs (i.e., $DV_1, \ldots, DV_n$) is determined by the values of the IVs (i.e., by $IV_1, \ldots, IV_m$). This is done as follows. Let $\mathbf{I} \subseteq \mathbb{R}^m$ denote the set of potential values of the m-tuples $[IV_1, \ldots, IV_m]$, let $\mathbf{D} \subseteq \mathbb{R}^n$ denote the set of potential values of the n-tuples $[DV_1, \ldots, DV_n]$, and let $\mathbf{\Theta} \subseteq \mathbb{R}^r$ denote the set of potential values of the r-tuples $\underline{\boldsymbol{\theta}} = [\theta_1, \ldots, \theta_r]$, where $\theta_1, \ldots, \theta_r$ are real-valued intervening variables called *parameters* that mediate the effect of the IVs on the DVs. A *fully-specified mathematical model* specifies an *input function* $f : \mathbf{I} \rightarrow \mathbf{\Theta}$ that maps each $[IV_1, \ldots, IV_m] \in \mathbf{I}$ to some $\underline{\boldsymbol{\theta}} \in \mathbf{\Theta}$, and an *n-tuple output function* $G : \mathbf{\Theta} \rightarrow \mathbf{D}$ that maps each $\underline{\boldsymbol{\theta}} \in \mathbf{\Theta}$ to an n-tuple $[DV_1, \ldots, DV_n] \in \mathbf{D}$. The model predicts, for all $[IV_1, \ldots, IV_m] \in \mathbf{I}$, that

$$[DV_1, \ldots, DV_n] = G[f(IV_1, \ldots, IV_m)]. \tag{1}$$

It is often useful to express the n-tuple output function G in terms of real-valued component output functions. Thus

$$G(\underline{\boldsymbol{\theta}}) = [g_1(\underline{\boldsymbol{\theta}}), \ldots, g_n(\underline{\boldsymbol{\theta}})]. \tag{2}$$

Using these functions, the model's predictions can be rewritten as:

$$DV_i = g_i[f(IV_1, \ldots, IV_m)], \text{ for } i = 1, \ldots, n. \tag{3}$$

Definition 3 (*Output-specified Mathematical Model*)**.** Given the same experiment and notation as in Definition 2, an *output-specified mathematical model* consists of an n-tuple output function $G : \mathbf{\Theta} \rightarrow \mathbf{D}$. The model predicts that, as the values of the IVs (i.e., $IV_1, \ldots, IV_m$) are varied, it will always be the case that

$$[DV_1, \ldots, DV_n] \in G[\mathbf{\Theta}] = \{G(\underline{\boldsymbol{\theta}}) : \underline{\boldsymbol{\theta}} \in \mathbf{\Theta}\}. \tag{4}$$

Thus, given an output-specified model, its output function G could be paired with any one of a variety of different input functions f to produce a variety of different fully-specified models.

Earlier we noted that almost all mathematical models in psychology are output-specified models, in the sense that they specify output functions, but rarely say much, if anything, about input functions. For example, signal detection theory specifies exact equations that predict P(Hit) and P(FA) given values of its parameters d' and X_c. But the theory is much more vague about how the IVs manipulated in an experiment determine values of d' and X_c. It predicts that d' should increase with signal intensity, but it does not postulate a functional form for this increase, and it prescribes how X_c might change with payoffs, but an estimated value of X_c that differs from the predicted value is generally not considered strong evidence against the theory.

Historically, mathematical psychology has not considered an input function as a necessary component of a mathematical model. There are several reasons that the field has focused on output functions. First, one could speculate that the natural evolution of mathematical modeling is to first focus on identifying the output functions, and only shift attention to the input function after this first problem is largely solved. The predictions of an output function can be tested directly against observed data, so it should be possible to identify an incorrect output function via sufficient empirical testing. The predictions of input functions though are often unobservable since they predict numerical values of some hypothetical parameters. Without knowing something about the true output function, it might not even be possible to identify the appropriate parameters. If not, then it seems hopeless to try to identify the correct input function.

Second, mathematical psychology has agreed collectively on a rather small set of DVs to receive the lion's share of attention — including, for example, response accuracy and response time. Thus, the search for output functions can largely be restricted to functions that make predictions about this small set of DVs. In contrast, there is no such universal agreement about relevant IVs. In fact, there are virtually an unlimited number of potential IVs that could affect response time or accuracy. And each new IV requires specifying a new input function. So naturally, the search for input functions has lagged behind the search for output functions.

Even so, the field has allocated some attention to input functions, and we suspect that this trend will only increase in the future. For example, some general modeling principles are clearly directed at the input function. For example, consider the principle of correspondent change (Townsend & Ashby, 1983), which states that, if a model is valid, then changing some IV should only cause a correspondent change in the value of the parameter the theory associates with this IV. So if signal detection theory is valid, then increasing signal intensity should increase d' but have no effect on X_c. From the input-, output-function perspective, this is clearly an attempt to force a theory to make (e.g., ordinal) predictions about its input functions, or at least to favor theories that make some empirically supported predictions about input functions over theories that make no predictions. Furthermore, there have been some attempts to build stronger models. For example, van Ravenzwaaij, Brown, Marley, and Heathcote (2020) used Fechner's law to predict brightness and Valentin, Maddox, and Ashby (2014) proposed a model of how feedback delays affect learning rates in tasks that depend on procedural learning by modeling the time-course of the biochemical events in the striatum that mediate synaptic plasticity. These attempts are all incomplete however, since the final models still included free parameters. A fully-specified model would predict values of the DVs directly from knowledge of the IVs, without appealing to any unknown free parameters.

For these reasons, although we believe that a complete (i.e., fully-specified) model must include both input and output functions as described in Definition 2, to be consistent with popular terminology, we will use the term model even in cases when only weak assumptions are made about the input function.

2.3. The dimensionality of state-trace plots

A state-trace plot is generated by plotting values of DV_2 against values of DV_1. Suppose an empirical STA supports the inference of a single varying parameter (i.e., that $r = 1$). A proposed model of these data might vary a single parameter θ across the tasks in an attempt to model the bottleneck imposed by the single parameter. Note that for a single numerical value of θ, the model predicts a single point on a state-trace plot, namely

$$(DV_1, DV_2) = [g_1(\theta), g_2(\theta)]. \tag{5}$$

If the value of θ is changed, then the model predicts a different point. Therefore, continuously changing θ sweeps out a curve in state-trace space, and each single numerical value of θ points to a single point on this curve.[1] The curve is the state-trace plot predicted by the model under the assumption that θ is the only parameter that varies across DV_1 and DV_2.

Bamber (1979) briefly considered the possibility of generalizing STA to an arbitrary number of tasks. This possibility has not been seriously pursued (although see Dunn & Anderson, 2018; Dunn & James, 2003), but the exercise provides insights into standard applications of STA, and it also offers the possibility of extending applications of STA to new domains, and thereby increasing its usefulness and applicability.

Consider the situation described in Definition 2, in which we compare performance across n DVs (which could come from any number of tasks between one and n), rather than only two. The space of all possible outcomes of these n DVs defines the experiment's *data space* **D**, and note that any point in **D** can be indexed by the ordered n-tuple $\underline{\boldsymbol{d}} = [DV_1, DV_2, \ldots, DV_n]$. Similarly, the space of all possible values of a model's parameters, Θ, defines the model's *parameter space*, and note that any point in Θ can be indexed by the ordered r-tuple $\underline{\boldsymbol{\theta}} = [\theta_1, \theta_2, \ldots, \theta_r]$. For any specific numerical combination of its parameters, a model predicts a performance value in each task. Thus, the output function of a model maps its r-dimensional parameter space to n-dimensional data space:

$$G : \Theta \rightarrow \mathbf{D}, \tag{6}$$

where in general $G(\underline{\boldsymbol{\theta}}) = [g_1(\underline{\boldsymbol{\theta}}), g_2(\underline{\boldsymbol{\theta}}), \ldots, g_n(\underline{\boldsymbol{\theta}})]$. Note that traditional STA is a special case of this scenario in which $n = 2$.

Now consider $G(\Theta)$, the image of G. These are all possible data combinations that the model can fit perfectly. When $n > r$; that is, when there are more DVs than free parameters, then we expect that there will be possible data outcomes that the model cannot perfectly fit. In these cases, the image of G is a proper subset of **D** and it defines a manifold in data space. Therefore, call this image the *model manifold*.[2] Note that every state-trace plot predicted by the model is a subset of the model manifold — that is, all points in the predicted state-trace plot must belong to the model manifold, but we expect that some points in the model manifold will not be represented in the state-trace plot. The goal of STA is to determine the dimensionality of the model manifold.

Topologists have devised a few different ways of defining the dimension of a topological space. In the case of "well behaved" spaces (i.e., normal spaces with a countable base), these definitions all agree with each other.[3] For our purposes, the definition of dimension with the nicest properties is the *small inductive dimension*, also known as the *Menger–Urysohn dimension*. The small inductive dimension of a topological space X is denoted $\operatorname{ind} X$. A formal definition requires more topological machinery than is needed for this article because we are concerned here only with spaces X that are subsets of some Euclidean space $\mathbb{R}^s$, $s \geq 1$, and that have the usual subspace topology.[4] In this case, the small inductive dimension has the following properties, which are sufficient for our purposes.

- For every nonempty subset X of $\mathbb{R}^s$, $\operatorname{ind} X$ is defined and is equal to an integer between zero and s inclusive. In particular, $\operatorname{ind} \mathbb{R}^s = s$.
- If X and Y are nonempty subsets of $\mathbb{R}^s$ and $X \subseteq Y$, then $\operatorname{ind} X \leq \operatorname{ind} Y$.
- Suppose X is a subset of $\mathbb{R}^s$. Then $\operatorname{ind} X = s$ if and only if X is a *substantial subset* of $\mathbb{R}^s$ – that is, if and only if $X \supseteq O$, where O is some nonempty open subset of $\mathbb{R}^s$.
- Suppose that X and Y are nonempty subsets of $\mathbb{R}^s$ and $\mathbb{R}^t$ respectively, where s and t need not be equal. If X and Y are homeomorphic, then $\operatorname{ind} X = \operatorname{ind} Y$.

Proposition 1 establishes the dimensionality of the model manifold, and therefore the dimensionality of the resulting state-trace plot.

Proposition 1. *Consider applications of a model with r free parameters, $\underline{\boldsymbol{\theta}} = [\theta_1, \theta_2, \ldots, \theta_r]$, to a task or collection of tasks with n DVs, where $n \geq r$. Let $G(\underline{\boldsymbol{\theta}}) = [g_1(\underline{\boldsymbol{\theta}}), g_2(\underline{\boldsymbol{\theta}}), \ldots, g_n(\underline{\boldsymbol{\theta}})]$ denote the model's output function. Suppose that the parameter space Θ is a substantial subset of $\mathbb{R}^r$ and that the data space* **D** *is a substantial subset of $\mathbb{R}^n$. Suppose further that the model's output function G is a* homeomorphic embedding *(Engelking, 1989, p. 67) of the model's parameter space Θ into the model's data space* **D***; or in other words, that the following conditions hold:*

1. *The output function G is one-to-one (i.e., injective) — that is, if $\underline{\boldsymbol{\theta}} \neq \underline{\boldsymbol{\theta}}^*$, then $G(\underline{\boldsymbol{\theta}}) \neq G(\underline{\boldsymbol{\theta}}^*)$. (This guarantees that the output function has an inverse.)*
2. *The output function G is continuous.*
3. *Its inverse is also continuous.*

Then, the model manifold has dimension $\operatorname{ind} G(\Theta) = r$, whereas $\operatorname{ind} \mathbf{D} = n$. If $r < n$ then the model manifold has a smaller dimension than the data space. As a result, we say that the model is a homeomorphic-embedding model with a bottleneck.

Proof. Because Θ is a substantial subset of $\mathbb{R}^r$, $\operatorname{ind} \Theta = r$. Because the output function G is one-to-one, it has an inverse. And, because G and its inverse are both continuous, G is (by definition) a homeomorphism, and thus Θ and $G(\Theta)$ are homeomorphic. Therefore, $\operatorname{ind} G(\Theta) = \operatorname{ind} \Theta = r$. Finally, because **D** is a substantial subset of $\mathbb{R}^n$, $\operatorname{ind} \mathbf{D} = n$. □

This proposition has a number of important implications. First, note that it predicts that in standard two-task applications of STA, models in which only one parameter is varying must predict a single-monotonic or single-nonmonotonic state-trace curve, whereas models with two or more varying parameters must predict a scatter plot.

Second, it also suggests that in some cases it might be useful to generalize STA to three or more tasks. For example, suppose we plot performance across various conditions and/or participants on three different DVs. So points in the data space are denoted by the ordered triple $[DV_1, DV_2, DV_3]$. First, Proposition 1 tells us

[1] These arguments assume that the model satisfies the conditions of Proposition 1.

[2] Technically, it is not necessarily a manifold. Proposition 1 will describe conditions on $G(\Theta)$ that guarantee that it is a manifold or a manifold with boundary, and as we will discuss, almost all mathematical models within psychology satisfy these conditions.

[3] A topological space X is normal if any two disjoint closed sets of X have disjoint open neighborhoods.

[4] The interested reader can find the definition in either Engelking (1989, chap. 7), or Pol (2004).

that any model for which only one parameter is varying predicts a one-dimensional model manifold — that is, the state trace plot will be a one-dimensional curve through the three-dimensional data space. Second, as Bamber (1979) noted, any model for which two parameters are varying predicts a two-dimensional model manifold — in other words, a curved surface in data space. Finally, models for which three or more parameters vary predict a three-dimensional model manifold. In this case, the predicted performance combinations could fill a three-dimensional volume that is a subset of data space. Current applications of STA allow experimenters to identify scenarios in which one parameter is varying versus more than one, but current applications cannot discriminate between cases where two parameters are varying versus more than two. So adding a third task has the potential to allow identification of three possibilities: DV triples in which only one parameter is varying, DV triples in which two parameters are varying, and DV triples in which three or more parameters are varying.

Note that the conditions required for Proposition 1 to hold are all exceedingly weak. For example, in the standard two-task STA, the first condition simply means that any model in which only one parameter is varying cannot produce a state-trace curve that intersects itself. And the continuity conditions just imply that small changes in the parameter values cause small changes in predicted performance. The strongest condition is arguably that G is continuous because this condition could rule out a model that predicts a bifurcation as some parameter increases through a critical point — that is, a model that predicts a qualitative change in performance as a parameter increases from below to above some critical threshold value. A few such models have been proposed (e.g., Savi, Marsman, van der Maas, & Maris, 2019; Van der Maas & Molenaar, 1992), but the vast majority of current cognitive models do not violate any of these conditions.

It is also important to note that the conditions of Proposition 1, although weak, are necessary for STA to succeed at identifying the number of varying parameters. For example, if G is not one-to-one, then it is possible that a model with a single varying parameter could fill an entire area of the standard two-task state-trace plot. If so, then STA would conclude wrongly that two or more underlying parameters are varying. Such space-filling curves, which were first discovered by Peano (1890), are continuous but not smooth. But it is possible that in the absence of a one-to-one mapping, a single-parameter model could even produce a smooth state-trace curve that would be impossible to distinguish from an area-filling state-trace plot (i.e., a scatter plot) produced by a model with two or more parameters. For example, Bamber and Van Santen (1985) identified the Lissajous curve as an example of this phenomenon.[5]

The requirement that G is one-to-one rescues us from these scenarios.[6] Even so, this rescue is only theoretical because it is possible to construct single-parameter one-to-one mappings G that produce state-trace curves that, in practice, would be statistically impossible to distinguish from a scatter-plot state trace. For example, most space-filling curves are constructed by taking the limit of a sequence of simpler curves, each of which is a one-to-one mapping from the unit interval to the unit square, with the property that each successive curve in the sequence more closely approximates the area-filling limit. So a curve that is late in the sequence, but before the limit, is one-to-one and will fill much of the state-trace plot. Most importantly, because of statistical error (measurement, perceptual, cognitive, or individual difference), it would be impossible to discriminate from a scatter plot.

On the other hand, there are several reasons that these identifiability concerns are not serious problems. First, as already mentioned, we know of no models in the literature that produce anything close to an area-filling state-trace plot when only a single parameter is varied, and it is difficult to conceive of any future model having this problem.

The second reason why these identifiability issues should not be of concern is much more important. And this reason stems from the basic question of why we would want to use STA to identify how many parameters are varying. The answer, of course, is that our goal is to study the complexity of the underlying model — that is, we are interested in how tight the bottleneck is between the independent and dependent variables (Bamber, 2019). And we use the number of varying parameters as an operational definition of this complexity. This makes sense because in general, adding a parameter to a model increases its complexity or mathematical flexibility. In fact, popular goodness-of-fit measures like AIC and BIC define complexity or flexibility in exactly this way. But it is also widely recognized (at least within the statistics literature) that this measure of flexibility is imperfect. For example, in the field of information geometry, the mathematical flexibility of a model can be measured by the volume of its model manifold (i.e., its image in data space). Remember that each point in the model manifold is a data combination that the model can fit perfectly. Therefore, by this definition, the more different data sets the model can fit perfectly, the more complex or flexible it is. The volume of the model manifold will almost always increase with the addition of a new parameter, but models with the same number of parameters are not typically associated with the same volumes. For example, Myung, Balasubramanian, and Pitt (2000) used this approach to show that a one-parameter power function model is more complex than a one-parameter log function model (i.e., because a power function is more flexible or "bendy" than a log function). The important point here is that the number of free parameters is not a perfect measure of mathematical complexity, and that information geometry would not differentiate between a one-parameter model that produces a scatter-plot state trace versus a two-parameter model that produces the same plot. Both models would be classified as equally complex and the fact that one varies one parameter and the other varies two would be considered irrelevant. So by this account, it is still important to discriminate between a single-monotonic or single-nonmonotonic state-trace plot versus a scatter plot, regardless of whether the conditions of Proposition 1 are satisfied. The model producing the single-monotonic or single-nonmonotonic plot is less complex than the model producing the scatter plot (i.e., the bottleneck is narrower), regardless of whether or not the models differ in the number of varying parameters.

The remainder of this article considers the standard two-task STA and assumes that the conditions of Proposition 1 are met. The next result reiterates the original mathematical foundations of STA.

Proposition 2. *Suppose the conditions of Proposition 1 are met and the number of DVs is $n = 2$. Let DV_1 and DV_2 denote the two DVs that are recorded during performance of one or two tasks. Consider the STA that plots DV_2 on the ordinate against DV_1 on the abscissa.*

1. A model in which a single parameter varies across the two conditions (i.e., $r = 1$) always predicts a single-monotonic state-trace curve if performance on both tasks is monotonic with increases in the parameter. Furthermore, any such state-trace plot could have been produced by such a one-parameter model.

2. A model in which a single parameter varies across the two conditions always predicts a single, nonmonotonic state-trace curve

[5] Note that Peano curves and Lissajous curves are continuous, so the problems arise from condition 1 of Proposition 1, not from conditions 2 or 3.

[6] Although just barely. For example, Osgood curves are non-intersecting curves of positive area. Even so, they are not space-filling and so, theoretically at least, could be distinguished from state traces produced by models with two or more varying parameters.

if performance on DV_1 is monotonic and performance on DV_2 is nonmonotonic with changes in the parameter, or vice versa.

3. A model in which two or more parameters are varying across the two conditions can produce any type of state-trace plot.

Proof. Parts 2 and 3 are originally due to Bamber (1979), and part 1 is due to Dunn and Kirsner (1988). However, for completeness, and because the proofs are simple, we reproduce them here.

1. Denote the single varying parameter by θ. Then there exist some output functions g_1 and g_2 such that $DV_1 = g_1(\theta)$ and $DV_2 = g_2(\theta)$. Because g_1 is strictly monotonic in θ, g_1 has an inverse and the inverse is itself a function. Therefore,

$$\theta = g_1^{-1}[DV_1], \tag{7}$$

which implies that

$$DV_2 = g_2\left\{g_1^{-1}[DV_1]\right\}. \tag{8}$$

A function of a function is itself a function (i.e., $g_2 \circ g_1^{-1}$ is a function), and since g_2 and g_1^{-1} are both monotonic in θ, then so too is $g_2 \circ g_1^{-1}$.

If the state-trace plot is a single monotonically-increasing curve, then there must exist a monotonic increasing function $M(\cdot)$ such that $DV_2 = M(DV_1)$, for every point (DV_1, DV_2) on the curve. A model with one varying parameter θ that is consistent with this curve is the model in which $DV_1 = g_1(\theta) = \theta$ and $DV_2 = g_2(\theta) = M(\theta)$.

2. By Proposition 1, any model in which only one parameter varies across tasks predicts a one-dimensional state-trace curve. If DV_i is nonmonotonic with the single varying parameter θ, then there exists values $\theta' < \theta'' < \theta'''$ all in Θ, such that either

$$g_i(\theta') < g_i(\theta'') > g_i(\theta'''), \tag{9}$$

or else

$$g_i(\theta') > g_i(\theta'') < g_i(\theta'''). \tag{10}$$

In either case, the state-trace curve is nonmonotonic.

3. If two parameters are varying, then by Proposition 1 the model is capable of covering a region of positive area in data space. Any point in this region is a possible outcome, so by selection, one can choose points that create a plot of any type. For example, Fig. 1d was created by simultaneously varying the GCM overall discriminability (i.e., c) and attention weight (i.e., w) parameters in rule-based (RB) and information-integration (II) categorization tasks (see the Appendix for a description). Every point in the gray region is a prediction of the model with a different (c, w) combination. Note that this region includes all points on the curves shown in panels a – c. □

3. STA and monotonicity assumptions

Bamber's (1979) focus was to discriminate state-trace plots that are consistent with one varying parameter versus plots that rule out only one varying parameter. So his interest was in determining whether a state-trace plot was a single-monotonic or single-nonmonotonic curve versus a double curve or scatter plot. Some later applications of STA followed this approach (e.g., Loftus, 2002). In contrast, many more recent applications of STA attempted to discriminate between single-monotonic curves and all other types — in other words, single-nonmonotonic curves, double curves, and scatter plots were all lumped together (e.g., Prince, Brown, & Heathcote, 2012; Stephens, Matzke, & Hayes, 2019, 2020). For example, according to Prince et al. (2012) "determining whether there is one or more than one mediating latent variable is accomplished by determining whether the state-trace plot is monotonic" (p. 81).

This approach to STA, which focuses on single-monotonic curves, explicitly assumes the following default model.

Definition 4 (*Monotonic State-trace Model*)**.** Consider an experiment in which two DVs are each recorded under n different experimental conditions. Then a *monotonic state-trace model* is an output-specified model in which Θ is a single parameter that is assigned value θ_i in experimental condition i by the input function; and $g_1(\cdot)$ and $g_2(\cdot)$ are monotonic nondecreasing output functions that generate predicted performances on DVs 1 and 2, respectively.

These models are useful because, by Proposition 2, they make the strong and empirically testable prediction that the state-trace plot must be monotonic. From a statistical perspective, this implies that the plot will contain no *delta-discordant* pairs of points. This means that if (x, y) and (x', y') are two points on the state-trace plot, then it is impossible that $x > x'$ but $y < y'$. Several statistical tests of this hypothesis have been developed (Kalish, Dunn, Burdakov, & Sysoev, 2016; Prince et al., 2012), and Bamber (2019) showed how adaptive methods could be used to find pairs (IV_1, IV_2) and (IV_1', IV_2') that map to delta-discordant pairs of points, *if such pairs exist* and under the additional condition that the input and output functions are all continuous. Therefore, the advantage of focusing on single-monotonic curves – and therefore assuming that performance on both tasks is monotonic with changes in any parameters – is primarily statistical. In particular, since any empirical state-trace plot only includes a finite, and usually reasonably small number of points, discriminating between single-nonmonotonic and double state-trace plots can be challenging (Bamber, 2019). For example, to show that a plot is a double curve rather than a single-nonmonotonic curve requires showing that there are no missing points that connect the two separate curves that define a double plot. The assumption of monotonicity greatly simplifies this problem.

Bamber (2019, Section 3) argued that, although monotonicity assumptions can make it easier to test state-trace models, such assumptions are not essential to STA. This point is formalized in Proposition 1, which provides an example of a testable type of model that does not assume monotonicity — namely, homeomorphic-embedding models with bottlenecks. Furthermore, Propositions 1 and 2 together, make clear that the monotonic state-trace model, by failing to discriminate single-nonmonotonic state-trace curves from double curves or scatter plots, considerably reduces the applicability of STA because it means that a state-trace plot that is consistent with a single varying parameter can only be identified if an extra assumption is added that is both strong and unnecessary — namely that performance on both tasks is monotonic with increases in the single varying parameter. This assumption is strong because there are many highly plausible single-parameter models that violate monotonicity. The well-known and popular generalized context model (GCM; Nosofsky, 1986) with a freely varying attention weight parameter w is one clear example (for other examples, see Ashby, 2019). The assumption of monotonicity is unnecessary because Propositions 1 and 2 make clear that models with a single varying parameter can be identified without assuming monotonicity.

Stephens et al. (2019) used STA to re-analyze the results of many studies that compared performance in rule-based (RB) and information-integration (II) categorization tasks. In virtually all of these tasks, the stimuli varied on two dimensions and the categories were identical except for their orientation in stimulus space (i.e., in which the diagonal-trending II categories were created by rotating the RB categories by 45° in stimulus space). Stephens et al. (2019) concluded that in almost all cases the resulting state-trace plots were single monotonic curves, and they suggested that the GCM with a single varying attention weight parameter was a viable monotonic state-trace model for these results. Furthermore, Stephens et al. (2020) continued to

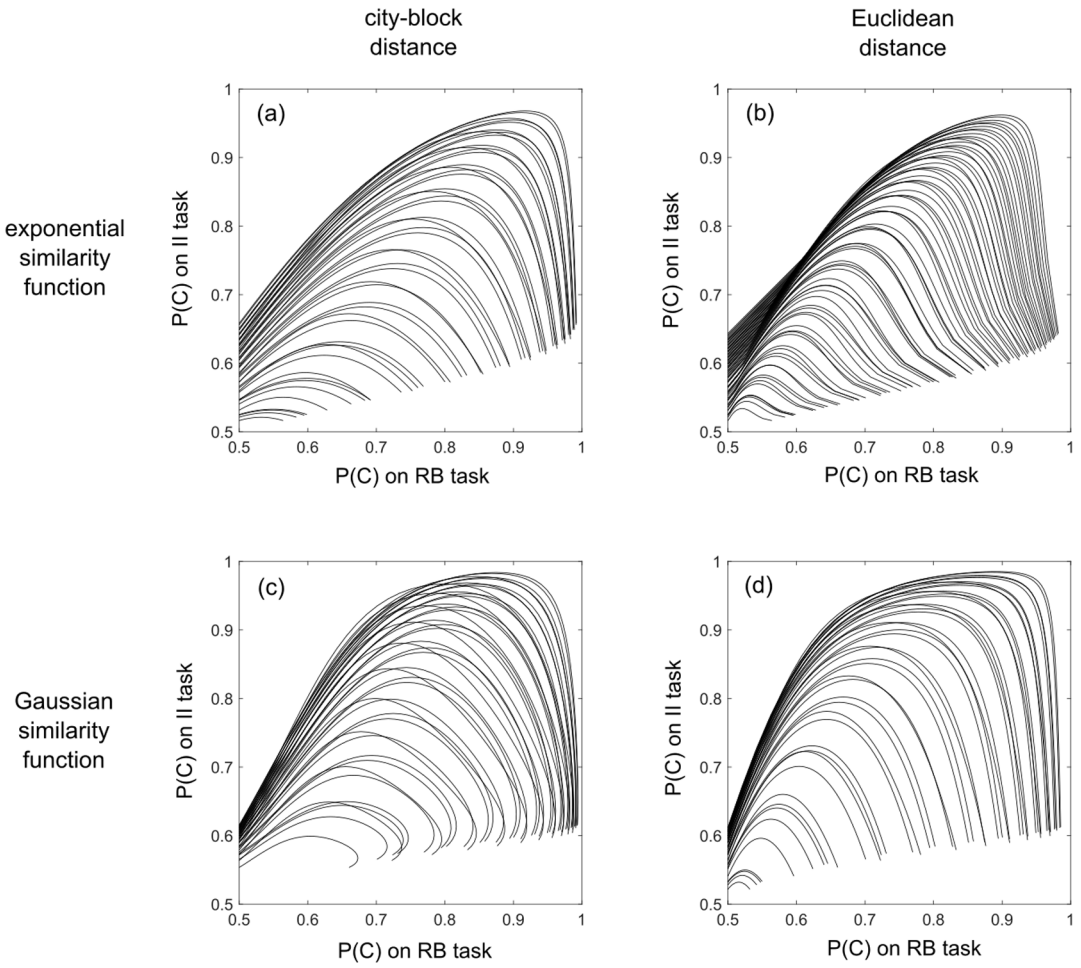

Fig. 2. State-trace plots predicted by the GCM for different similarity functions and distance measures as the attention weight w varies continuously in RB and II tasks that are identical except for the orientation of the categories in their two-dimensional stimulus space. Different curves are associated with different constant values of the GCM parameters c (overall discriminability) and β (bias toward response A). The parameter β varies from 0.2 to 0.8 (in units of 0.1) in all panels. The parameter c varies from 2 to 24 in panel (a), from 2 to 20 in panel (b), from 25 to 125 in panel (c), and from 5 to 105 in panel (d).

make this claim. Unfortunately, however, the GCM with a single varying attention weight is not a monotonic state-trace model because it violates the monotonicity assumption and therefore predicts single-nonmonotonic curves in these applications.

Consider a state-trace analysis that plots performance in an II categorization task against performance in an RB task. Suppose that the categories are identical in the two tasks except for their orientation in stimulus space, and that dimension 1 is the only relevant dimension in the RB task. Define the GCM attention parameter w as the proportion of attention allocated to dimension 1 (i.e., so the proportion allocated to dimension 2 equals $1 - w$). Then the GCM predicts that performance in the RB task increases monotonically with w (i.e., because stimulus dimension 1 is the only relevant dimension), but performance in the II task increases as w increases from 0 to .5, and then decreases as w increases from .5 to 1 (because both dimensions are equally important in the II task).

Fig. 2 shows predictions of four different versions of the GCM under these conditions. The four versions are identical except for how distance is computed (city-block versus Euclidean) and the function that relates distance to similarity (exponential or Gaussian). The Appendix describes all four versions of this model in detail. Each curve in all four panels of Fig. 2 was generated by varying only the GCM attention weight w (i.e., proportion of attention allocated to dimension 1) continuously from 0 to 1. The different curves in each panel are each associated with different fixed values of other GCM parameters — specifically, overall discriminability (i.e., the GCM c parameter) and the bias toward response A (i.e., the GCM β parameter). Fig. 2 shows that all four versions of the GCM always predict single-nonmonotonic state-trace curves under these conditions. Therefore, the GCM with a single varying attention parameter is not a monotonic state-trace model and therefore, in general, is not a viable model for single-monotonic RB versus II state-trace plots, such as those reported, for example, by Stephens et al. (2019). The only exception occurs if all of the discrete points on the empirical monotonic state-trace plots fall either to the left or right of the peaks of the curves shown in Fig. 2. In this case, the empirical state-trace plot appears monotonic because it only samples from a restricted range of the state-trace curve predicted by the underlying model. The peaks of the Fig. 2 curves occur when $w = .5$, which is the optimal value of w in the II task. In the RB task, the optimal value is $w = 1$. So for all points in an empirical state-trace plot to fall on the same side of the peak of Fig. 2 curves, every participant or group of participants would have to have been biased in the II task to allocate more attention to dimension 1 than to dimension 2, and this same bias would have to exist, for example, in all experiments for which Stephens et al. (2019) reported monotonic state-trace curves. This is a straightforward prediction to test. One simply needs to fit the GCM to each II data set that was used to generate a state-trace point, and then check whether $\hat{w} > .5$ in every case. Stephens et al. (2019) did not report the results of such a test, nor did they mention this strong prediction of the model they recommended.

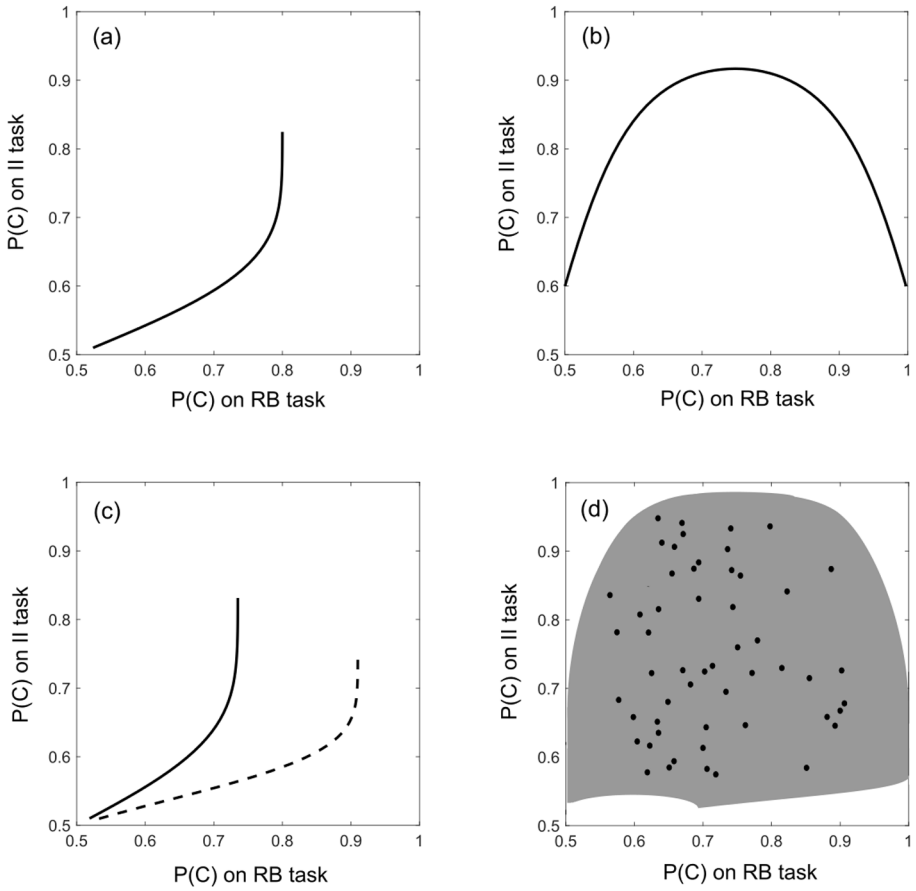

Fig. 3. Four different state-trace results predicted by a dual-systems model that assumes a simple rule-based strategy is used in RB tasks and that the GCM is used in II tasks. The single-monotonic curve in panel (a) was generated by assuming that the overall discriminability parameter c varies across tasks and participants. The single-nonmonotonic curve in panel (b) was generated by assuming that the attention weight parameter w varies across tasks and participants. The double curves in panel (c) show a possible outcome of an experiment with two groups of participants in which the parameter c varies continuously within each group, but each group is characterized by a different value of w. Panel (d) was generated by assuming that both c and w vary across tasks and participants.

The more important point however, is that lumping the single-nonmonotonic curves in Fig. 2 with double curves and scatter plots leads to the false conclusion that none of the many curves in Fig. 2 are consistent with a single varying parameter.

4. STA and the number of underlying systems or processes

Propositions 1 and 2 make no mention of the architecture of the model that produces the state-trace plot. For example, both propositions make predictions about the form of the state-trace plot predicted by a model in which only one parameter is varying, and these results hold regardless of whether that single varying parameter is embedded in a model that postulates 1, 2, or 27 underlying processes or systems. Nevertheless, a number of recent articles have attempted to use STA to identify the number of underlying cognitive systems that are mediating performance in the two tasks. Specifically, these articles claimed that the number of parameters (or latent variables) that the STA finds to be varying across the two tasks under study places a lower bound on the number of underlying cognitive systems, and as a result, a one-dimensional state-trace plot therefore can be interpreted as evidence favoring single-system models over multiple-systems models (e.g., Dunn, 2008; Dunn, Newell, & Kalish, 2012; Newell & Dunn, 2008; Newell, Dunn, & Kalish, 2011; Stephens et al., 2019, 2019). In other words, the primary goal of these articles was to use STA, not to infer whether one parameter was varying versus more than one, but rather in an attempt to infer whether that one varying parameter was embedded in a model that postulated one cognitive system or two systems.

STA is ill suited to this task. There are methods that were designed to test between alternative cognitive architectures. Townsend and Nozawa's (1995) elegant work on systems factorial technology comes immediately to mind. However, STA was not developed with the goal of identifying the underlying cognitive architecture. Furthermore, it has been known for some time that STA is incapable of identifying the number of underlying systems (Ashby, 2014), at least in the case, for example, of the models that Stephens et al. (2019) were considering. Even so, attempts to use STA for such purposes are still prevalent (Stephens et al., 2020). Although Propositions 1 and 2 should be sufficient to establish the futility of this exercise, this section makes this point even more clear.

The articles that have been directed at this purpose have used the terms "processes" and "systems" interchangeably and without any formal definition. As in their dictionary definitions, we view the term process as more general than the term system, because the latter requires reference to architecture, whereas the former does not. Even so, in the absence of formal definition, both terms are ambiguous. For example, in psychology, the word "system" is used without consensus and in a wide variety of ways. For example, plausible arguments can be made for the seemingly contradictory positions that the brain is one single giant system, or that it includes many functionally distinct systems. In the memory literature, separate systems are typically identified

operationally – by the presence of a double dissociation. In neuroscience, a system is often defined based on neuroanatomical connections.

The term "process" is enveloped in similar ambiguity. To illustrate this ambiguity, we will consider a thought experiment in which a ball is fired from a cannon. We will model the well-understood flight of this ball with two parameters and derive the resulting two-dimensional state-trace plot. Next, we will describe a disagreement among three hypothetical researchers who have differing views about the number of processes that mediate the flight of the cannonball. Researcher I thinks there is only one process. Researchers IIA and IIB believe there are two processes, but they disagree about the identity of those processes. We will argue that each of these viewpoints is reasonable, and therefore that the disagreement is because it is unclear what is meant by the term "process".

This example illustrates that because the existing literature is unclear about what is meant by a "process" or a "system", it is not meaningful to claim that the results of any STA can be used to make inferences about the number of active processes or systems.

4.1. A thought experiment

Suppose we fire a ball from a cannon that is situated at the edge of a flat, level plain, and then record three DVs:

- $y_{\max}$, the maximum altitude (measured by radar) that is reached by the cannonball;
- t_{gr}, the elapsed time from when the cannon is fired to when it hits the ground;
- x_{gr}, the horizontal distance traversed by the cannonball when it hits the ground.

The resulting STA will plot values of these three DVs while we manipulate the following three IVs:

- the angle, denoted θ_{muz}, that the muzzle of the cannon is elevated above horizontal;
- the weight of gun powder that we put in the cannon;
- the weight of the cannonball. Although we always use the same size cannonball, we can make the ball out of substances with different densities (e.g., aluminum, iron, lead, depleted uranium).

4.2. Calculation of the state trace

The state-trace for this thought experiment can be calculated using elementary physics. Assume that the effects of the Earth's curvature and rotation are negligible, as is air resistance and the decline in gravitational force with increased distance from the center of the Earth.

Let v_{muz} denote the muzzle velocity (i.e., speed) of the cannonball as it exits the cannon. This is a monotonic increasing function of the weight of gun powder in the barrel and a monotonic decreasing function of the weight of the cannon ball. The cannonball's muzzle velocity has a horizontal component v^x_{muz} and a vertical component v^y_{muz}. These are given by:

$$v^x_{\text{muz}} = (\cos\theta_{\text{muz}})\, v_{\text{muz}} \text{ and } v^y_{\text{muz}} = (\sin\theta_{\text{muz}})\, v_{\text{muz}}. \tag{11}$$

Let t denote time elapsed since the firing of the cannon. Let $x(t)$ denote the horizontal distance traveled by the cannonball at time t and $y(t)$ denote the vertical distance (i.e., altitude). Let g denote gravitational acceleration. Then

$$x(t) = v^x_{\text{muz}}\, t; \tag{12}$$

$$y(t) = v^y_{\text{muz}}\, t - (g/2)\, t^2. \tag{13}$$

Now $dy/dt = v^y_{\text{muz}} - gt$. So, $dy/dt = 0$ when $t = v^y_{\text{muz}}/g$. So, $y(t)$ is at a maximum when $t = v^y_{\text{muz}}/g$. Thus,

$$y_{\max} = y(v^y_{\text{muz}}/g) = (v^y_{\text{muz}})^2/(2\, g). \tag{14}$$

Note that $y(t) = 0$ when either $t = 0$ or $t = 2v^y_{\text{muz}}/g$. Thus,

$$t_{\text{gr}} = 2v^y_{\text{muz}}/g. \tag{15}$$

Therefore,

$$x_{\text{gr}} = x(t_{\text{gr}}) = v^x_{\text{muz}} t_{\text{gr}} = 2v^x_{\text{muz}} v^y_{\text{muz}}/g. \tag{16}$$

Recall that the DVs in our thought experiment are $y_{\max}$, t_{gr}, and x_{gr}. So, applying Eqs. (14)–(16), we see that

$$\text{ST}_{\text{thought exp}} \tag{17}$$
$$= \left\{ \left(\frac{[v^y_{\text{muz}}]^2}{2g},\ \frac{2v^y_{\text{muz}}}{g},\ \frac{2v^x_{\text{muz}} v^y_{\text{muz}}}{g} \right) \in \mathbb{R}^3 : v^x_{\text{muz}} > 0 \ \&\ v^y_{\text{muz}} > 0 \right\}.$$

Here we have expressed the state trace using two parameters: v^x_{muz} and v^y_{muz}. Alternatively, because v^x_{muz} and v^y_{muz} are functions of v_{muz} and θ_{muz} (Eq. (11)), we could have re-expressed the state trace for the thought experiment using v_{muz} and θ_{muz} as parameters.

By Proposition 1, this state trace plot is two-dimensional.

4.3. How many processes?

We have analyzed our flight-of-the-cannonball thought experiment using two parameters. Does that mean that there are two processes involved in the flight of the cannonball? Not necessarily. Different people with different intuitions about how to define a process can have different opinions about the number of processes involved in the flight of the cannonball. For example, consider the opinions of the following three people.

Person I. The arc of the cannonball through the sky is entirely explained by Newton's laws. Nothing more is needed. The cannonball's flight is therefore just one process.

Person IIA. There are two processes involved in the cannonball's flight. One process, governed by the weight of gunpowder and the weight of the cannonball, determines the initial speed imparted to the cannonball. The other process, governed by the elevation of the cannon's barrel, determines the initial direction of the cannonball's motion.

Person IIB. There are two components to the motion of the cannonball. The vertical component is affected by gravity; the horizontal component is not. Therefore, the flight of the cannonball is mediated by two processes.

Although each of these people has a sensible and tenable point of view, their opinions differ markedly. They disagree about the number of processes, and even the two people who agree that there are two processes disagree about the nature of those processes. These three people have different opinions because they have different intuitions about what constitutes a "process". The problem is that "process" is an intuitive term rather than a technical term.

In order to take seriously any claim that the number of parameters revealed by an STA is a lower bound on the number of processes or underlying systems, it is necessary to formally define "process" and/or "system". The next two sections propose new definitions of single- and multiple-systems models and then show that STA is incapable of distinguishing between these two model classes.

We formally define systems, rather than processes, for two reasons. First, previous attempts to use STA for this purpose have

tried to test between models that are widely considered to postulate one versus two cognitive systems, rather than one versus two processes. Second, the existence of models in which there is wide consensus about the number of underlying systems provides test cases that can be used to validate our definitions. In contrast, we know of no models in which there is wide consensus about the number of processes they postulate. As mentioned previously, we interpret a process to be more general than a system. Therefore, we assume that any scenario that satisfies our definition of multiple systems also implies multiple processes. However, we have no position on the reverse inference. Specifically, we make no claim that a scenario that satisfies our definition of a single system implies a single process. For example, by the definitions offered in the next section, the normal equal-variance, signal detection theory model is a single-system model. However, some could reasonably argue that it postulates separate perceptual and decisional processes.

Some may feel that our definitions can be improved upon in a way that makes it possible to show that the number of varying parameters revealed by an STA is, in fact, a lower bound on the number of underlying systems that mediate the phenomenon under study. We welcome such attempts.

4.4. Formal definitions of single-system and multiple-systems models

To begin, note that debates about whether performance is mediated by one or multiple systems will almost always compare performance across multiple tasks. For example, a multiple-systems model might predict that performance in two tasks is dominated by two different systems, whereas a single-system model must predict that the same system mediates performance in both tasks. In an STA where two different DVs are compared from the same task, all models predict that the same cognitive system or systems generated behavior on every trial, which greatly complicates any attempt to identify the number of underlying systems. As a result, in this and the next section we assume that the STA compares performance on two separate tasks T_1 and T_2. For example, Figs. 1–3 show state-trace plots in which overall accuracy is computed for two different categorization tasks, and these are exactly the types of state-trace plots that Stephens et al. (2019) used in their attempts to test between single- and multiple-systems models.

As noted above, our first problem is to define formally what we mean by multiple systems. Unfortunately, although a variety of researchers have attempted to use STA to identify the number of underlying systems, none of them has defined what they mean by system. This alone makes it difficult to interpret any claims about the ability of STA to identify the number of systems. We are skeptical that any mathematical definition of single- versus multiple-systems would accurately convey the many uses of the term "system" that are currently prevalent in the psychological literature. Instead, some restrictive assumptions are required. For example, note that even the successful efforts of Townsend and Nozawa (1995) to discriminate between alternative cognitive architectures was restricted to a small and well-defined set of competing models.

Current and prior attempts to use STA to identify the number of systems have been loosely based on the notion of a "system" as the term is used in the memory literature. The double-dissociation test of multiple systems that is commonly used in this literature is derived from a black-box architecture in which single-system models assume one black box (or series of black boxes) with the property that destroying any part of the architecture – for example, via lesion – must affect predictions of the model that describes performance of that system in all tasks. According to this account, multiple-systems models assume at least two separate black boxes that are arranged in a way that destroying one has no effect on the performance of the other. This is the type of experiment that motivated initial proposals that humans have multiple memory systems. For example, lesioning the hippocampus impairs episodic memory but not procedural memory, whereas lesions to the striatum impair procedural memory but not episodic memory (e.g., Packard & McGaugh, 1992).

The COVIS model of category learning evolved from the memory literature, so characterizing COVIS as a multiple-systems model assumes this same black-box definition of system. In particular, COVIS postulates separate rule-based and procedural-learning black boxes that are each capable of generating a categorization response on their own. So for example, COVIS predicts that a simultaneous dual-task should impair RB learning more than II learning, whereas a feedback delay should impair II learning more than RB learning (e.g., Ashby & Valentin, 2017). Therefore, this is the definition of system that Stephens et al. (2019) implicitly had in mind when they characterized the GCM as a single-system model and COVIS as a dual-systems model and then concluded that "these state-trace analyses show that the evidence for two distinct category learning systems is much more limited and inconsistent than is implied by the impressive list of dissociations presented by Ashby and Valentin (2017)" (p. 14).

For these reasons, we chose to define single- and multiple-systems models in a way that is consistent with how the term "system" is used in the memory literature – namely, that an experimental intervention that affects any part of a single-system model should affect the predictions of that model in all tasks, whereas in a multiple-systems model there should exist some interventions that affect predicted performance in some tasks, but not others. We show that these definitions define different classes of models, in the sense that the two classes make different predictions in some experiments and under certain conditions. Then, in Propositions 3 and 4, we show that this reasonable class of single-system and multiple-systems models are essentially nonidentifiable via STA.

Propositions 3 and 4 do not rule out the possibility that single-system and multiple-systems models might be defined in some different way that offers some hope that STA might contribute to discriminating between the two classes, at least under some conditions. Even so, they cast serious doubt on any possible use of STA for this purpose, and at the minimum they show that there is no mathematical basis for the current practice of applying STA to this problem.

Our next definition formalizes this notion of a system. Then we define single- and multiple-systems models.

Definition 5 (*System*)**.** Suppose

$$\underline{\boldsymbol{\theta}} = (\theta_1, \ldots, \theta_r); \tag{18}$$

$$\underline{\boldsymbol{\theta}}' = (\theta_1', \ldots, \theta_r'). \tag{19}$$

Suppose $\varepsilon > 0$. We say that $\underline{\boldsymbol{\theta}}$ and $\underline{\boldsymbol{\theta}}'$ are *ε-close axially* if, for some m, $\theta_k = \theta_k'$ for all $k \neq m$ and if $|\theta_m - \theta_m'| < \varepsilon$.

Consider two real-valued functions $g_1(\cdot)$ and $g_2(\cdot)$ defined on a subset $\boldsymbol{\Theta}$ of $\mathbb{R}^r$. We say that $g_1(\cdot)$ and $g_2(\cdot)$ *move alike locally* on $\boldsymbol{\Theta}$ if they have the following property: For every $\underline{\boldsymbol{\theta}} \in \boldsymbol{\Theta}$, there exists an $\varepsilon > 0$ such that, for every $\underline{\boldsymbol{\theta}}' \in \boldsymbol{\Theta}$ that is ε-close axially to $\underline{\boldsymbol{\theta}}$,

- either $g_1(\underline{\boldsymbol{\theta}}) = g_1(\underline{\boldsymbol{\theta}}')$ and $g_2(\underline{\boldsymbol{\theta}}) = g_2(\underline{\boldsymbol{\theta}}')$,
- or $g_1(\underline{\boldsymbol{\theta}}) \neq g_1(\underline{\boldsymbol{\theta}}')$ and $g_2(\underline{\boldsymbol{\theta}}) \neq g_2(\underline{\boldsymbol{\theta}}')$.

Furthermore, we say that the pair $[g_1(\cdot), g_2(\cdot)]$ constitutes a *system* if $g_1(\cdot)$ and $g_2(\cdot)$ move alike locally.

Our application of this definition will be to an STA that compares performance in two tasks, T_1 and T_2. In this case, a system

is a pair of functions that depend on a set of common parameters and have the property that changing any one of those parameters changes performance in both tasks (or in neither task). Given this definition, we now define single- and multiple-systems models.

Definition 6 (*Single- and Multiple-systems Models*)**.** Consider an experiment in which performance is compared in two tasks, T_1 and T_2. Let $P(T_1)$ and $P(T_2)$ denote numeric measures of performance on the relevant DVs in tasks T_1 and T_2, respectively. A single-system model is a mathematical model with parameter space $\Theta \subseteq \mathbb{R}^r$ that assumes performance in both tasks depends on the same output function $g_S(T_i, \underline{\theta})$ (the subscript S is for single), where $i = 1, 2$ and $\underline{\theta} \in \Theta$. The output function might depend on characteristics of the DV being recorded, the task (e.g., which stimuli are used), and on one or more parameters indexed in $\underline{\theta}$. Even so, the same core equations are used to derive predictions in both tasks. Specifically,

$$P(T_1) = g_S(T_1, \underline{\theta}) \text{ and } P(T_2) = g_S(T_2, \underline{\theta}), \text{ where } \underline{\theta} \in \Theta. \tag{20}$$

If the pair of functions $[g_S(T_1, \cdot), g_S(T_2, \cdot)]$ constitute a system as defined in Definition 5 (i.e., they move alike locally), then we say that Eq. (20) is a *single-system model* for the tasks T_1 and T_2.

A model with parameter space $\Theta \subseteq \mathbb{R}^r$ that assumes $N_S \geq 2$ underlying cognitive systems is a mathematical model in which performance in all tasks depends on N_S different functions, $g_{M,1}$, $g_{M,2}, \ldots, g_{M,N_S}$, and again, each may depend on characteristics of the DV being recorded, the task, and on one or more parameters indexed in $\underline{\theta}$ (the subscript M is for multiple). Suppose that each of the function pairs

$$[g_{M,j}(T_1, \cdot), g_{M,j}(T_2, \cdot)], \; j = 1, \ldots, N_S \tag{21}$$

is a system as defined in Definition 5 (i.e., they move alike locally). These systems may be quite different from one another, and we assume that at least one parameter in $\underline{\theta}$ affects some systems but not others. Specifically, we assume there exists $i \neq j$ for which

$$[g_{M,i}(T_1, \cdot), g_{M,j}(T_2, \cdot)], \tag{22}$$

do not move alike locally. In addition, some real-valued supervisory function h determines how each subsystem contributes to performance. Specifically, for $i = 1, 2$,

$$P(T_i) = h[g_{M,1}(T_i, \underline{\theta}), g_{M,2}(T_i, \underline{\theta}), \ldots, g_{M,N_S}(T_i, \underline{\theta})], \; \underline{\theta} \in \Theta. \tag{23}$$

Such a model is called a *multiple-systems model* with N_S separate systems.

Note that this definition formalizes the criteria used in the memory literature that any intervention that affects a single-system model should affect the predictions of that model in all tasks, whereas in a multiple-systems model there should exist some interventions that affect predicted performance in some tasks, but not others. By definition, single-system models assume that the same cognitive system is used in all tasks. Therefore, the predictions of a single-system model on the same DV in two different tasks will be closely related because such models will use the same equations in both tasks and the same parameters. The predictions could differ because the tasks differ. For example, the two tasks might use different stimuli. Nevertheless, because the same equations and parameters are used in both tasks, Definition 6 assumes that changes in any parameter that cause predictions to change in one task will also cause predictions to change in the other task. In contrast, multiple-systems models assume that there are multiple cognitive systems that are each capable of performing at least some tasks, and that different cognitive systems may be used in different tasks and conditions. In any multiple-systems model, different equations are used to derive predictions for each cognitive system. Furthermore, each system could have its own unique parameters that do not affect predictions of any other system. Therefore, with multiple systems, it is possible that task T_1 recruits a certain cognitive system but task T_2 does not. Note that in this case, changing a parameter that affects the predictions of that system will cause predictions to change in task T_1 but not in task T_2, which is impossible in a single-system model.[7]

More formally, consider the $r + 1$ dimensional space that has a dimension for every parameter in $\underline{\theta}$ plus $P(T_i)$. Note that any output function defines a manifold in this space with the property that any point on the manifold gives the predicted performance in task T_i for any possible combination of parameter values. We call this the output function's T_i *model manifold*. In a single-system model, there is only one output function for each task, so there is only one T_i model manifold. As a result, for any fixed set of parameter values there is only one predicted value $P(T_i)$. In a multiple-systems model however, there are N_S output functions for each task and therefore N_S different T_i model manifolds that all lie in the same $r + 1$ dimensional space. So for any fixed set of parameter values there are N_S predicted values of $P(T_i)$. The supervisory function h determines how these N_S values contribute to the observed performance in task T_i. Furthermore, note that any manifold that is flat on one dimension predicts the same value of $P(T_i)$ for all values of the parameter that defines that dimension. As a result, in a multiple-systems model, one or more manifolds could be flat on some dimension, whereas this is not allowed in a single-system model. Therefore, in addition to differing in the number of their T_i model manifolds, single- and multiple-systems models also differ in the nature of those manifolds.

Although this definition of single- versus multiple-systems models is an attempt to formalize the definition of system that is assumed implicitly in the memory literature, it is important to note that Definition 6 says nothing about double dissociations. A double dissociation describes a possible experimental outcome, and says nothing directly about the underlying model that produced that data. In contrast, Definition 6 states conditions that a mathematical model must satisfy to be classified as postulating a single system or multiple systems. In fact, it has been known for many years that models seemingly constructed from a single black box can predict double dissociations (Plaut, 1995). In agreement with this, it is straightforward to show that models classified as single system by Definition 6 can predict double dissociations. And of course, a model meeting the Definition 6 criteria for multiple systems will not predict a double dissociation in all experiments. For this reason, Definition 6 focuses on the structure of the model rather than the appearance of the data.

Note that almost all current models that are widely regarded as single-system or multiple-systems models would be correctly classified by this definition. For example, consider the single-system GCM and the dual-systems COVIS models of categorization, which are the same models that Stephens et al. (2019) had in mind when they were trying to identify the number of underlying systems. In the single-system GCM, the function g_S is described by Eq. (40)–(42) in the Appendix. These same equations are used to predict performance (i.e., accuracy) in all tasks. This is the model used to generate Figs. 1 and 2. The model makes different predictions in the RB and II tasks that define the state-trace plots in those figures, even with the same parameter values, because the stimuli are different. However, changing any of its parameter values changes its predictions in both tasks. In fact, this is true even when the two tasks are qualitatively different. In Figs. 1 and 2, both tasks are categorization and they

[7] This definition of single- and multiple-systems models agrees with the more informal definitions proposed by Ashby (2014).

differ only in the stimuli that define the two contrasting categories. The GCM though, also has been used to account for results from recognition-memory experiments (e.g., Nosofsky, 1988). Although the ways that Eq. (40)–(42) are combined to predict a recognition-memory response versus a categorization response are different, note that these different composite GCM functions still move alike locally, and therefore Definition 6 would classify the GCM as a single-system model, even when it is applied to two such different tasks. In particular, changing any GCM parameter causes the model to change predictions in both tasks.

The dual-systems model COVIS includes parameters that affect predicted performance in II tasks, but these same parameters have no effect on predicted performance in RB tasks. In the dual-systems model used to generate Fig. 3, which is a simplified version of COVIS, $g_{M1} = g_S$ from the GCM, whereas g_{M2} is described by Eq. (43)–(45) in the Appendix. In this model, the parameter X_1 only affects predicted RB performance, whereas the parameters γ and β only affect predicted II performance (see the Appendix for a description of these parameters). Among multiple-systems models currently in the literature, common choices for the supervisory function h are to choose the single most accurate system, the most confident system on each trial (Ashby, Alfonso-Reese, Turken, & Waldron, 1998), or to blend the outputs of the multiple systems, for example, in a manner proportional to each system's accuracy (Erickson & Kruschke, 1998).

It is important to note that Definition 6 identifies two separate classes of models that make different predictions under certain conditions. For example, note that these definitions show immediately that a state-trace plot that is a horizontal or vertical line is possible for a multiple-systems model, but is inconsistent with all single-system models. For example, a horizontal state-trace plot shows that the experimenter-manipulated IVs changed performance on task T_1 (by changing some underlying parameter) but had no effect on performance in task T_2. This suggests that qualitatively different models were used in the two tasks, and thereby supports a multiple-systems account over a single system. Note that a horizontal- or vertical-line state-trace plot means that a perfect dissociation occurred, in which the manipulated IV caused performance to change in one task, but not the other. Therefore, any perfect dissociation is also compatible with a multiple-systems account and incompatible with all single-system models.

On the other hand, using these different predictions as an empirical test of single versus multiple systems presents a significant statistical challenge. For example, discriminating between a horizontal state-trace plot that falsifies single-system accounts and a plot with a minuscule positive slope that does not rule out single-system models would require large sample sizes. Furthermore, note that the null hypothesis in such tests would be that the multiple-systems model is correct (e.g., that the state-trace plot is horizontal). This is opposite current practice in which the null hypothesis is that the single-system model is correct (e.g., Stephens et al., 2019). Therefore, although these differential predictions are important to establish that the model classes are different, they are only of limited empirical usefulness. In light of this, an obvious question is whether there are any other outcomes of an STA that could provide evidence about the number of underlying systems. The next section answers this question.

4.5. STA is unable to identify the number of systems

The next result shows that, except for state traces that include horizontal- or vertical-line segments, the set of all possible state traces predicted by Definition 6 single- and multiple-systems models are identical. In other words, any state-trace plot produced by any single-system model can be reproduced exactly by some multiple-systems model, and any such state-trace plot produced by any multiple-systems model can be reproduced exactly by some single-system model.

Proposition 3. *Let $P(T_1)$ and $P(T_2)$ denote numeric measures of performance in two tasks T_1 and T_2, respectively. Consider a state-trace plot that includes no horizontal or vertical line segments. Then, (a) any state-trace plot produced by any single-system model can be reproduced exactly by some multiple-systems model. Conversely, (b) any state-trace plot produced by any multiple-systems model can be reproduced exactly by some single-system model.*

Proof. *Part (a).* Suppose that $[g_S(T_1, \cdot), g_S(T_2, \cdot)]$ is a single-system model with parameter space $\Theta \subseteq \mathbb{R}^r$. In other words, Eq. (20) of Definition 6 holds.

We will now construct a multiple-systems model with the pair of systems

$$[g_{M,1}(T_1, \cdot), g_{M,1}(T_2, \cdot)] \text{ and } [g_{M,2}(T_1, \cdot), g_{M,2}(T_2, \cdot)], \tag{24}$$

which has the same parameter space and predicts the same state trace as this single-system model. There are many ways to do this. We will do it in such a way that the two systems are substantially different from each other. Define

$$g_{M,1}(T_1, \underline{\theta}) = -g_S(T_1, \underline{\theta}) \;\&\; g_{M,1}(T_2, \underline{\theta}) = g_S(T_2, \underline{\theta})^3 + g_S(T_2, \underline{\theta}) \tag{25}$$

$$g_{M,2}(T_1, \underline{\theta}) = g_S(T_1, \underline{\theta})^3 + g_S(T_1, \underline{\theta}) \;\&\; g_{M,2}(T_2, \underline{\theta}) = -g_S(T_2, \underline{\theta}). \tag{26}$$

Note that the function that maps any real z to z^3+z is monotonic increasing. Therefore, because $g_S(T_1, \cdot)$ and $g_S(T_2, \cdot)$ move alike locally, it follows that the same applies to the two functions in (25) and the two functions in (26). Thus, each of the two function pairs in (24) is a system.

Finally, we define a supervisory function h for this multiple-systems model. For any real x and y, let

$$h(x, y) = (x + y)^{1/3}. \tag{27}$$

Then note that Eq. (23) of Definition 6 holds, so this is a multiple-systems model in which the number of systems is $N_S = 2$. This model predicts the state trace

$$\begin{aligned} P(T_i) &= h[\, g_{M,1}(T_i, \underline{\theta}),\ g_{M,2}(T_i, \underline{\theta})\,] \\ &= \left([\, g_S(T_i, \underline{\theta})^3 + g_S(T_i, \underline{\theta})\,] + [\, -g_S(T_i, \underline{\theta})\,] \right)^{1/3} \\ &= g_S(T_i, \underline{\theta}), \end{aligned} \tag{28}$$

for $i = 1, 2$, and therefore it predicts the same state trace as the single-system model.

Part (b). Consider a multiple-systems model that generates a state trace $\mathcal{M} \subseteq \mathbb{R}^2$. We will construct a single-system model that predicts the same state trace. Let the parameter space of the single-system model be

$$\Theta = \{(x + y, x - y) \in \mathbb{R}^2 : (x, y) \in \mathcal{M}\}. \tag{29}$$

For $\underline{\theta} = (\theta_1, \theta_2) \in \Theta$, consider the model that predicts

$$\begin{aligned} &P(T_1) = g_S(T_1, \underline{\theta}) = (\theta_1 + \theta_2)/2 \text{ and} \\ &P(T_2) = g_S(T_2, \underline{\theta}) = (\theta_1 - \theta_2)/2. \end{aligned} \tag{30}$$

Note that, if either θ_1 or θ_2 is varied while the other is held constant, then $g_S(T_1, \underline{\theta})$ and $g_S(T_2, \underline{\theta})$ will both vary. Thus, the functions $g_S(T_1, \cdot)$ and $g_S(T_2, \cdot)$ move alike locally and, thus, the pair define a single-system model. This model generates the state trace:

$$\{[(\theta_1 + \theta_2)/2, (\theta_1 - \theta_2)/2] : (\theta_1, \theta_2) \in \Theta\}. \tag{31}$$

But $(\theta_1, \theta_2) \in \Theta$ if and only if $\theta_1 = x + y$ and $\theta_2 = x - y$ for some $(x, y) \in \mathcal{M}$. As a result, this model predicts the state trace

$$P(T_1) = \frac{(x+y)+(x-y)}{2} = x, \tag{32}$$

and

$$P(T_2) = \frac{(x+y)-(x-y)}{2} = y, \tag{33}$$

and therefore it predicts the same state trace as the multiple-systems model. □

Proposition 3 shows that if a state trace is consistent with a single-system model, it is necessarily also consistent with some multiple-systems model, and conversely if the plot is consistent with a multiple-systems model and contains no horizontal- or vertical-line segments, then it necessarily is also consistent with some single-system model. This result establishes the futility of using STA to test between single- and multiple-systems models, at least for all single- and multiple-systems models defined as in Definition 6. The only ambiguity it leaves is whether the single-system models that mimic predictions of a multiple-systems model would be widely recognized as postulating a single processing system (and vice versa). The proposition guarantees that such mimicking single-system models exist, but it says nothing about the psychological assumptions made by these mimicking models. The next result shows that the mimicking models, in both classes, need not be exotic. In particular, widely popular single-system and multiple-systems models can both account for all four types of state-trace plots.

Proposition 4. *Consider the class of models that are universally recognized as postulating a single system and the class of models that are universally recognized as postulating multiple systems. Then the following results hold.*

1. A single-system model can produce any type of state-trace plot. The conditions under which it predicts each of the four types are exactly as described in Proposition 2.

2. A multiple-systems model can produce any type of state-trace plot. The conditions under which it predicts each of the four types are exactly as described in Proposition 2.

Proof. Both of these results were established by Ashby (2014). Note that they can both be established by construction – that is, it suffices to identify single- and multiple-systems models that predict state-trace plots of each possible type.

1. All four panels of Fig. 1 were generated from the GCM, which is universally accepted as a single-system model (e.g., Stephens et al., 2019). Panel (a) was generated by assuming that the GCM overall discriminability parameter c varies continuously. In the GCM, performance increases monotonically with c in all tasks, so panel (a) follows from part 1 of Proposition 2. Panel (b) was generated by assuming that the GCM attention weight parameter w varies continuously. As described in Section 3, GCM performance in the RB task increases monotonically with w, but performance in the II task increases to a maximum when $w = .5$ and then decreases. Therefore, panel (b) follows from part 2 of Proposition 2. Panel (d) was generated by assuming that both c and w vary continuously. Every point in the gray region denotes a different possible outcome of the two tasks. The black dots represent a random sample of these possible outcomes. Panel (c) denotes one possible subset of the panel (d) points in gray.

2. All four panels of Fig. 3 were generated from a simplified version of the COVIS model, which is universally accepted as a dual-systems model (e.g., Stephens et al., 2019). The model assumed that in the RB task, participants used a simple rule-based strategy that investigated only two possible rules – a one-dimensional rule on dimension 1 and a one-dimensional rule on dimension 2. In the II task, the model assumed participants used the GCM. See the Appendix for details. The single-monotonic curve in panel (a) was generated by assuming that the discriminability parameter c varies continuously. The single-nonmonotonic curve in panel (b) was generated by assuming that the attention weight w varies continuously. The double curves in panel (c) show a possible outcome of an experiment with two groups of participants. Each group was characterized by a different value of w. Within each group, c varied continuously. Panel (d) was generated by assuming that both c and w vary continuously. The black dots denote a random sample of possible outcomes. □

Propositions 3 and 4 establish that, except for horizontal or vertical lines, single-system and multiple-systems models can both produce any type of state-trace plot, and furthermore, that if a plot was generated by a model of one type, then it necessarily also could have been generated by a model of the opposite type. Thus, STA cannot be used to make any inferences about the number of systems that mediated performance in the tasks under study. Single-system and multiple-systems models predict exactly the same state-trace plots. Therefore, an STA is useful for concluding that a single parameter is varying across the two tasks under study, or equivalently, that an appropriate model of the tasks under study should only vary one parameter, but it can provide no information about the number of cognitive systems postulated by the model in which that one varying parameter is embedded.

Our results do not rule out the possibility that single- and multiple-systems models could be defined in some different way that brings STA into play. However, Proposition 4 shows that, even in this best case scenario, STA could provide useful input about the number of underlying systems only in certain, rather unusual models, or only if the field somehow universally changed its mind that the GCM is a single-system model and COVIS is a dual-systems model. At the minimum, the results in this section require that any future claim that STA has anything to contribute to a discussion about the number of underlying systems must include a formal definition of system that is qualitatively different from Proposition 4, and a careful justification that STA has a positive role to play in discriminating between these newly defined single-system and multiple-systems models.

5. Using STA to test dissociations

One recent and increasingly common application of STA, is to test for empirical dissociations. Newell and Dunn (2008) even argued that STA "overcomes all of the flaws of dissociation logic" (p. 285). For example, Stephens et al. (2019) used STA to re-analyze data from 28 different studies that each reported some type of dissociation between RB and II categorization. Their interpretation of the results of these analyses was that "we show that many of the dissociations thought to reflect the operation of distinct processes disappear against the stricter criteria of state-trace analysis" (p. 3). This section examines the mathematical basis for such conclusions.

Dissociations are almost always defined by comparing performance on two tasks. Therefore, this section considers the case in which each of two tasks T_1 and T_2 are run under n different experimental conditions. For example, T_1 and T_2 might be RB and II categorization tasks, respectively, and the n conditions could represent n different memory loads of some simultaneous dual task.

Let $P_i(T_1)$ and $P_i(T_2)$ denote performance in tasks T_1 and T_2, respectively, in condition i. Then the possible outcomes of a dissociation experiment that includes conditions i and j are illustrated

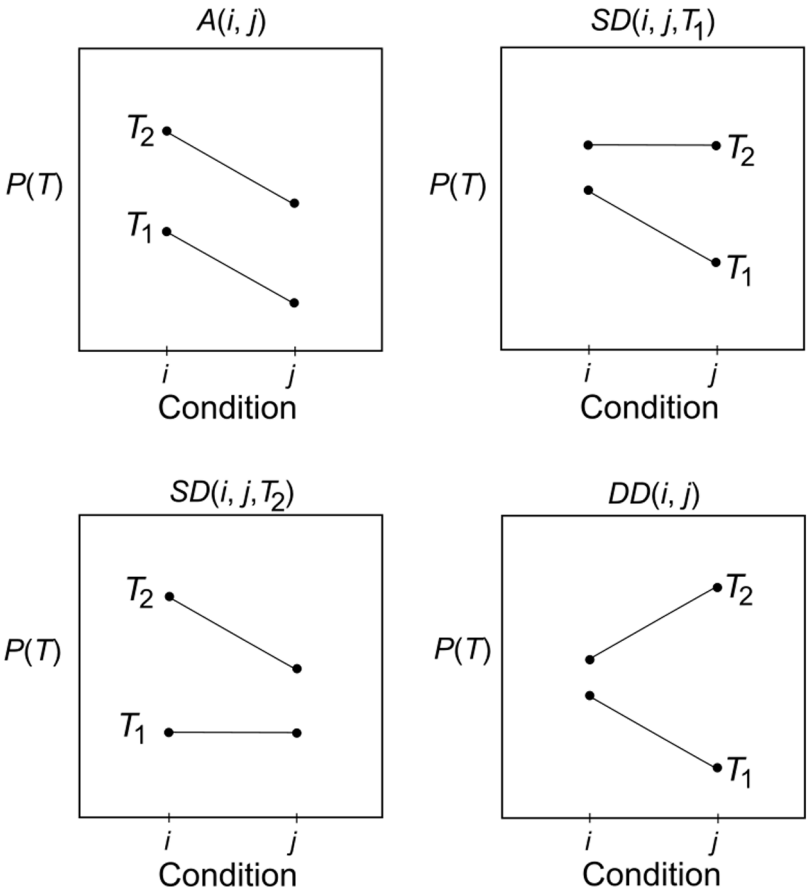

Fig. 4. Four possible outcomes of a dissociation experiment that compares performance on tasks T_1 and T_2 under different experimental conditions i and j.

in Fig. 4. Formally, the possible outcomes of this experiment are defined as follows.

Definition 7 (*Types of Dissociations*)**.** Consider a dissociation experiment in which two tasks T_1 and T_2 are run under experimental conditions i and j. Then the possible outcomes of this experiment are as follows.

- An *association of type* $A(i, j)$ occurs if $P_i(T_1) > P_j(T_1)$ and $P_i(T_2) > P_j(T_2)$;
- A *simple dissociation of type* $SD(i, j, T_1)$ occurs if $P_i(T_1) > P_j(T_1)$ and $P_i(T_2) = P_j(T_2)$;
- A *simple dissociation of type* $SD(i, j, T_2)$ occurs if $P_i(T_2) > P_j(T_2)$ and $P_i(T_1) = P_j(T_1)$;
- A *double dissociation of type* $DD(i, j)$ occurs if $P_i(T_1) > P_j(T_1)$ and $P_i(T_2) < P_j(T_2)$;
- A *null effect of type* $NE(i, j)$ occurs if $P_i(T_1) = P_j(T_1)$ and $P_i(T_2) = P_j(T_2)$.

Associations, dissociations, and null effects will all be called *effects*.

Almost all of the dissociations examined by Stephens et al. (2019) were single dissociations that were predicted *a priori* by the COVIS theory of category learning (Ashby et al., 1998). For example, one of the dissociations they examined was the prediction by COVIS that a simultaneous dual task would impair RB learning more than II learning.

To begin, it is important to note that using STA to examine dissociations requires stronger assumptions than a traditional dissociation analysis. In an STA, the experimenter plots ordered pairs $[P_i(T_1), P_i(T_2)]$ for different conditions i. But this requires that $P_i(T_1)$ and $P_i(T_2)$ are paired in some way – for example, perhaps because they were completed by the same participant or group of participants. This pairing is not a requirement of traditional dissociation analysis. For example, the points plotted in Fig. 4 could have all been estimated from different groups of participants run in different labs at different times. So obviously, for this reason, the remainder of this section is restricted to dissociation experiments that satisfy the pairing assumption required by STA.

Table 1
All possible outcomes of an experiment in which two tasks T_1 and T_2 are completed in two conditions i and j.

	$P_i(T_2) > P_j(T_2)$	$P_i(T_2) = P_j(T_2)$	$P_i(T_2) < P_j(T_2)$
$P_i(T_1) > P_j(T_1)$	$A(i, j)$	$SD(i, j, T_1)$	$DD(i, j)$
$P_i(T_1) = P_j(T_1)$	$SD(i, j, T_2)$	$NE(i, j)$	$SD(j, i, T_2)$
$P_i(T_1) < P_j(T_1)$	$DD(j, i)$	$SD(j, i, T_1)$	$A(j, i)$

Table 1 shows the effects associated with every possible outcome of a dissociation experiment. We shall be focusing on dissociations, particularly double dissociations. Note that there are $n(n-1)$ types of double dissociations, where n is the number of conditions under which the two tasks were run:

$$DD(i, j) : i = 1, \ldots, n;\ j = 1, \ldots, n;\ i \neq j. \tag{34}$$

Some of these double dissociations are incompatible with each other. Thus, $DD(i, j)$ and $DD(j, i)$ cannot both occur in the same experiment. Consequently, the number of double dissociations in an experiment can range from zero to a maximum of $n(n-1)/2$.

Our next result establishes the ability of STA to identify dissociations.

Proposition 5. *The results of a dissociation experiment are consistent with a monotonic state-trace model if and only if the experiment contains no double dissociations.*

Proof. *If* Consider a dissociation experiment that contains no double dissociations. We will construct a monotonic state-trace model for this experiment. We begin by defining a varying parameter $\boldsymbol{\theta}$ by setting

$$\theta_i = P_i(T_1) + P_i(T_2), \text{ for } i = 1, 2, \ldots, n. \tag{35}$$

We will also need to define two output functions: one for T_1 and one for T_2. We begin with the output function for T_1. We set

$$g_1(\theta_i) = P_i(T_1), \text{ for } i = 1, 2, \ldots, n. \tag{36}$$

For this to be a legitimate definition of the function $g_1(\cdot)$, we must show that, for any i and j, if

$$P_i(T_1) + P_i(T_2) = \theta_i = \theta_j = P_j(T_1) + P_j(T_2), \tag{37}$$

then

$$P_i(T_1) = g_1(\theta_i) = g_1(\theta_j) = P_j(T_1). \tag{38}$$

Assume that the definition is not legitimate. Then, for some i and j, (37) holds but $P_i(T_1) \neq P_j(T_1)$. Without loss of generality, suppose that $P_i(T_1) > P_j(T_1)$. Then (37) implies that $P_i(T_2) < P_j(T_2)$. This, in turn, implies that, contrary to supposition, there is a double dissociation in the experiment. Thus, our definition of the function $g_1(\cdot)$ is legitimate.

Next, we show that $g_1(\cdot)$ is monotonic nondecreasing. Consider any i and j for which $\theta_i = P_i(T_1) + P_i(T_2) < P_j(T_1) + P_j(T_2) = \theta_j$. Then note that the definition of $g_1(\cdot)$ requires that $g_1(\theta_i) = P_i(T_1) \leq P_j(T_1) = g_1(\theta_j)$. This must be the case, because if $P_i(T_1) > P_j(T_1)$ then it would have to be true that $P_i(T_2) < P_j(T_2)$. But if this were true then contrary to supposition, the experiment contains a double dissociation. Thus the function $g_1(\cdot)$ is monotonic nondecreasing.

Now, define an output function for T_2 by setting

$$g_2(\theta_i) = P_i(T_2), \text{ for } i = 1, 2, \ldots, n. \tag{39}$$

Using the same type of argument as above, it is seen that Eq. (39) is a legitimate definition and that the function $g_2(\cdot)$ is monotonic nondecreasing.

To summarize so far: $[\boldsymbol{\theta}, g_1(\cdot), g_2(\cdot)]$ is a monotonic state-trace model, and because of how it was constructed, it perfectly fits the state-trace plot that results from the dissociation experiment. Therefore, if an experiment has no double dissociations, then it is consistent with a monotonic state-trace model.

(*Only if*). Suppose the results of a dissociation experiment are consistent with the monotonic state-trace model $[\boldsymbol{\theta}, g_1(\cdot), g_2(\cdot)]$. Consider any two experimental conditions i and j. Note that there are three possibilities for the relative values of θ_i and θ_j

- $\theta_i > \theta_j$, which implies $P_i(T_1) \geq P_j(T_1)$ and $P_i(T_2) \geq P_j(T_2)$;
- $\theta_i = \theta_j$, which implies $P_i(T_1) = P_j(T_1)$ and $P_i(T_2) = P_j(T_2)$;
- $\theta_i < \theta_j$, which implies $P_i(T_1) \leq P_j(T_1)$ and $P_i(T_2) \leq P_j(T_2)$.

Table 2 describes these possible outcomes of the dissociation experiment (indicated by $\surd$) according to the monotonic state-trace model, along with the impossible outcomes (indicated by $\otimes$).

Comparing Table 2 to Table 1 shows that the double dissociations $DD(i, j)$ and $DD(j, i)$ are both impossible in the monotonic state-trace model. Therefore, the monotonic state-trace model predicts that the dissociation experiment will have no double dissociations. In other words, if an experiment is consistent with a monotonic state-trace model, then it contains no double dissociations. □

Table 2
Possible ($\surd$) and impossible ($\otimes$) outcomes of a dissociation experiment according to the monotonic state-trace model.

	$P_i(T_2) > P_j(T_2)$	$P_i(T_2) = P_j(T_2)$	$P_i(T_2) < P_j(T_2)$
$P_i(T_1) > P_j(T_1)$	$\surd$	$\surd$	$\otimes$
$P_i(T_1) = P_j(T_1)$	$\surd$	$\surd$	$\surd$
$P_i(T_1) < P_j(T_1)$	$\otimes$	$\surd$	$\surd$

Therefore, a single-monotonic state-trace curve rules out a double dissociation. Even so, Proposition 5 has the following immediate and important corollary.

Corollary 5.1. *A single-monotonic state-trace curve is consistent with either no dissociations* or *with one or more single dissociations. Therefore, the conclusion that a state-trace plot is a single-monotonic curve provides no information about whether or not the data include a single dissociation.*

As described earlier, Stephens et al. (2019) used STA to re-analyze data from 28 different studies that each reported some type of dissociation between RB and II categorization. All 28 studies were run to test an *a priori* prediction of COVIS about some single dissociation between RB and II learning or performance. Stephens et al. (2019) concluded that almost all of the resulting state-trace plots were single-monotonic curves, and they concluded from this result that "we show that many of the dissociations thought to reflect the operation of distinct processes disappear against the stricter criteria of state-trace analysis" (p. 3). Proposition 5 shows that this conclusion has no logical or mathematical basis. The finding that almost all of the state-trace plots from the 28 studies were single-monotonic curves provides no information about the presence or absence of the predicted dissociations.

Proposition 5 establishes the validity of using STA to test for double dissociations. And, as previously mentioned, double dissociations are often used, especially in the memory and cognitive neuroscience literatures, as operational tests of multiple systems. So at first glance, Proposition 5 might seem to provide support for the use of STA to identify the number of systems. However, it has long been known that single-system models can also account for double dissociations (Plaut, 1995). For example, Fig. 1d shows that the GCM can account for virtually *any* state-trace plot, even one consistent with a double dissociation. Furthermore, multiple-systems models do not predict double dissociations in all experiments. For example, in the many experiments re-analyzed by Stephens et al. (2019), the multiple-systems model COVIS always predicted single dissociations, and never predicted a double dissociation. So the presence or absence of a double dissociation allows no general inferences to be drawn about the number of underlying systems, even though the presence or absence of a double dissociation could be theoretically valuable (e.g., because it could confirm or disconfirm predictions of some specific model).

6. Conclusions

STA is arguably the best available method for determining the number of underlying parameters or latent variables that are varying across two or more tasks or conditions – that is, for measuring the width of the bottleneck between IVs and DVs. STA is based on the fact that under very weak conditions, any model in which r parameters are varying across r or more DVs or tasks predicts an r-dimensional state-trace plot (i.e., or model manifold). Thus, in the standard two-task STA, a model with one varying parameter predicts a one-dimensional state-trace plot –

that is, either a single-monotonic or single-nonmonotonic curve, whereas models with two or more varying parameters predict a scatter plot or a double plot. Note that this application does not require any monotonicity assumptions. Specifically, there is no need to assume that performance in any task is a monotonic function of whichever parameters are varying. Since many popular cognitive models include parameters that violate monotonicity (e.g., see Fig. 2), the practice of lumping single-nonmonotonic curves, double curves, and scatter plots together seriously reduces the ability of STA to identify phenomena that are mediated by a single varying parameter. Monotonicity is typically assumed in an attempt to increase statistical power. However, in many applications, the possible increase in statistical power comes at a high cost. For example, in the categorization tasks examined by Stephens et al. (2019), optimal performance required selective attention to one stimulus dimension in the RB tasks and equal attention to both dimensions in the II tasks. Even so, the Stephens et al. (2019) monotonicity assumption made it virtually impossible for them to identify data sets in which only a single attention weight was varying across the two tasks.

Whereas STA is among the best available methods for identifying the number of parameters that vary or the number of underlying latent variables, it provides virtually no information about the number of systems that characterize the underlying model or mechanisms that relate the IVs and DVs. The only exception is that a state trace that includes horizontal or vertical line segments rules out all single-system models. In the absence of this result, however, an STA might be used to conclude that the data are not complex enough to rule out a single varying parameter, but there is no possible outcome of any other STA that could be used to learn whether that single parameter varies in a model or architecture constructed from a single system or multiple systems. Single- and multiple-systems models predict exactly the same state-trace plots.

Similarly, STA is a poor choice if the goal is to examine dissociations. STA can be used to test for double dissociations, but not for single dissociations. In particular, a single-monotonic state-trace curve rules out a double dissociation but provides no information about whether or not the data contain a single dissociation.

Appendix. Description of the dual-systems model

The two tasks used to generate Figs. 1–3 were rule-based (RB) and information-integration (II) categorization tasks that used stimuli that varied on two stimulus dimensions. Therefore, each stimulus can be described by a point in a two-dimensional stimulus space, and each category is defined by a cluster of points in this space. In the RB task, each of the two equally-likely categories contained 10 stimuli, which were equally spaced on dimension 2 and shared the same value on dimension 1 (i.e., 0.6 for category A and 0.4 for category B). Thus, the optimal strategy is to ignore dimension 2 and respond A or B depending on whether the value of the presented stimulus on dimension 1 is large or small, respectively. The II task was identical, except the clusters were rotated 45° clockwise in stimulus space. In this condition, the category A and B clusters are perfectly partitioned by a diagonal line with slope 1. Therefore, the optimal strategy in the II task requires equal attention to both dimensions.

The dual-systems model was a simplified version of COVIS (Ashby et al., 1998). This model assumed that performance in the II task is mediated by the standard GCM (Nosofsky, 1986), augmented with the γ parameter introduced by Ashby and Maddox (1993). Ashby and Rosedahl (2017) showed that the GCM is a special case of the COVIS procedural system. Specifically, this version of the GCM assumes that the probability of responding A on a trial when stimulus k is presented equals

$$P(\mathrm{A}|k) = \frac{\beta\left(\sum_{i\in C_A}\eta_{ik}\right)^{\gamma}}{\beta\left(\sum_{i\in C_A}\eta_{ik}\right)^{\gamma} + (1-\beta)\left(\sum_{i\in C_B}\eta_{ik}\right)^{\gamma}}, \tag{40}$$

where C_A and C_B are sets containing the stimuli in categories A and B, respectively, η_{ik} is the similarity between stimuli i and k, β is a parameter that reflects the participant's bias toward responding A, and γ is a parameter that determines whether the participant probability matches ($\gamma = 1$), over-matches ($\gamma > 1$), or under-matches ($\gamma < 1$). Similarity is assumed to be inversely related to the weighted Minkowski distance between the perceptual representations of the stimuli. More specifically, the distance between the perceptual representations of stimuli i and k, denoted δ_{ik}, equals:

$$\delta_{ik} = \left(w|x_{i1} - x_{k1}|^r + (1-w)|x_{i2} - x_{k2}|^r\right)^{1/r}, \tag{41}$$

where w is the proportion of attention allocated to dimension 1, x_{ij} is the coordinate value of stimulus i on the jth perceptual dimension, and r is chosen to be 1 or 2 to produce city-block or Euclidean distance, respectively. The dual-systems model assumes Euclidean distance. Similarity is inversely related to distance via:

$$\eta_{ik} = \exp(-c\delta_{ik}^{\alpha}) \tag{42}$$

where c is a parameter that increases with the overall discriminability of the stimuli, and α is 1 or 2, which produces the exponential or Gaussian similarity function, respectively. The dual-systems model assumes a Gaussian similarity function. This is the same model used to produce Fig. 1. Note that there are four possible versions of the model (depending on whether $r = 1$ or 2 and $\alpha = 1$ or 2) and each has four free parameters (β, γ, w and c).

In the RB task, the dual-systems model assumed participants use a completely different rule-based system, which is a simplification of the COVIS explicit system. The model assumed that on each trial, the participant used one of two decision rules:

R_1: "Respond A if $x_{i1} > X_1$; otherwise respond B"
or

R_2: "Respond A if $x_{i2} > X_2$; otherwise respond B",

where x_{i1} and x_{i2} are the values of the current stimulus on dimensions 1 and 2, respectively, and X_1 and X_2 are parameters that represent the response criteria.

The former of these rules is optimal for the RB tasks considered in this article. If there is normally distributed perceptual or criterial noise, then the predicted probability correct for this rule equals (e.g., Ashby & Valentin, 2018)

$$P(C|\mathrm{R}_1) = .5\left[1 - \Phi\left(\frac{X_1 - .6}{\sigma}\right)\right] + .5\Phi\left(\frac{X_1 - .4}{\sigma}\right), \tag{43}$$

where Φ is the cumulative Z distribution function and σ^2 is the noise variance. Ashby and Maddox (1993) showed that σ is inversely related to the discriminability parameter c of the GCM. Therefore, we can eliminate a parameter by setting $\sigma = 1/c$, and so

$$P(C|\mathrm{R}_1) = .5\left\{1 - \Phi\Big[c(X_1 - .6)\Big]\right\} + .5\Phi\Big[c(X_1 - .4)\Big]. \tag{44}$$

The incorrect rule predicts that $P(C|\mathrm{R}_2) = .5$. The model assumes that use of the correct rule increases with w – that is, with the proportion of attention allocated to the relevant dimension. Specifically, the model assumes that

$$P(C) = wP(C|R_1) + (1-w)P(C|R_2)$$
$$= wP(C|R_1) + .5(1-w). \quad (45)$$

All four panels of Fig. 3 were generated with $\beta = .5$, $\gamma = 1$, and X_1 set to its optimal value. In panel (a), w was set to .6, and c varied from 1 to 40. In panel (b), c was set to 60, and w varied from 0 to 1. In panel (c), c varied from 1 to 40 in both groups. In one group, w was set to .47 and in the other it was set to .82. Finally, in panel (d), w varied from 0 to 1, and c varied from 5 to 105.

References

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

Ashby, F. G. (2019). State-trace analysis misinterpreted and misapplied: Reply to Stephens, Matzke, and Hayes (2019). *Journal of Mathematical Psychology*, *91*, 195–200.

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*(3), 442–481.

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, *37*(3), 372–400.

Ashby, F. G., & Rosedahl, L. (2017). A neural interpretation of exemplar theory. *Psychological Review*, *124*(4), 472–482.

Ashby, F. G., & Valentin, V. V. (2017). Multiple systems of perceptual category learning: Theory and cognitive tests. In H. Cohen, & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science* (2nd ed.). (pp. 157–188). Elsevier.

Ashby, F. G., & Valentin, V. V. (2018). The categorization experiment: Experimental design and data analysis. In E. J. Wagenmakers, & J. T. Wixted (Eds.), *Methodology*: *Vol. 5*, *Stevens' handbook of experimental psychology and cognitive neuroscience* (4th ed.). (pp. 307–347). New York: Wiley.

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

Bamber, D. (2019). Safeguarding against bad luck when attempting to discredit a state-trace model. *Journal of Mathematical Psychology*, *90*, 76–87.

Bamber, D., & Van Santen, J. P. (1985). How many parameters can a model have and still be testable? *Journal of Mathematical Psychology*, *29*(4), 443–473.

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

Dunn, J. C., & Anderson, L. (2018). Signed difference analysis: Testing for structure under monotonicity. *Journal of Mathematical Psychology*, *85*, 36–54.

Dunn, J. C., & James, R. N. (2003). Signed difference analysis: Theory and application. *Journal of Mathematical Psychology*, *47*(4), 389–416.

Dunn, J. C., & Kirsner, K. (1988). Discovering functionally independent mental processes: The principle of reversed association. *Psychological Review*, *95*(1), 91–101.

Dunn, J. C., Newell, B. R., & Kalish, M. L. (2012). The effect of feedback delay and feedback type on perceptual category learning: The limits of multiple systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(4), 840–859.

Engelking, R. (1989). *General topology* (Revised & completed edition). Berlin: Heldermann Verlag.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, *127*(2), 107–140.

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.

Loftus, G. R. (2002). Analysis, interpretation, and visual presentation of experimental data. In J. Wixted (Ed.), *Methodology in experimental psychology*: *Vol. 4*, *Stevens' handbook of experimental psychology* (3rd ed.). (pp. 339–390). New York: Wiley.

Myung, I. J., Balasubramanian, V., & Pitt, M. A. (2000). Counting probability distributions: Differential geometry and model selection. *Proceedings of the National Academy of Sciences*, *97*(21), 11170–11175.

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

Newell, B. R., Dunn, J. C., & Kalish, M. (2011). Systems of category learning: Fact or fantasy? *Psychology of Learning and Motivation-Advances in Research and Theory*, *54*, 167–215.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57.

Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(4), 700–708.

Packard, M. G., & McGaugh, J. L. (1992). Double dissociation of fornix and caudate nucleus lesions on acquisition of two water maze tasks: Further evidence for multiple memory systems. *Behavioral Neuroscience*, *106*(3), 439–446.

Peano, G. (1890). Sur une courbe, qui remplit toute une aire plane. *Mathematische Annalen*, *36*(1), 157–160.

Plaut, D. C. (1995). Double dissociation without modularity: Evidence from connectionist neuropsychology. *Journal of Clinical and Experimental Neuropsychology*, *17*(2), 291–321.

Pol, E. (2004). Dimension of metrizable spaces. In K. P. Hart, J. Nagata, & J. E. Vaughan (Eds.), *Encyclopedia of general topology* (pp. 314—317).

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

Savi, A. O., Marsman, M., van der Maas, H. L., & Maris, G. K. (2019). The wiring of intelligence. *Perspectives on Psychological Science*, *14*(6), 1034–1061.

Stephens, R. G., Matzke, D., & Hayes, B. K. (2019). Disappearing dissociations in experimental psychology: Using state-trace analysis to test for multiple processes. *Journal of Mathematical Psychology*, *90*, 3–22.

Stephens, R. G., Matzke, D., & Hayes, B. K. (2020). State-trace analysis – Misrepresented and misunderstood: Reply to Ashby (2019). *Journal of Mathematical Psychology*, *96*, Article 102342.

Townsend, J. T., & Ashby, F. G. (1983). *Stochastic modeling of elementary psychological processes*. New York: Cambridge University Press.

Townsend, J. T., & Nozawa, G. (1995). Spatio-temporal properties of elementary perception: An investigation of parallel, serial, and coactive theories. *Journal of Mathematical Psychology*, *39*(4), 321–359.

Valentin, V. V., Maddox, W. T., & Ashby, F. G. (2014). A computational model of the temporal dynamics of plasticity in procedural learning: Sensitivity to feedback timing. *Frontiers in Psychology*, *5*(643).

Van der Maas, H. L., & Molenaar, P. C. (1992). Stagewise cognitive development: An application of catastrophe theory. *Psychological Review*, *99*(3), 395–417.

van Ravenzwaaij, D., Brown, S. D., Marley, A., & Heathcote, A. (2020). Accumulating advantages: A new conceptualization of rapid multiple choice. *Psychological Review*, *127*(2), 186–215.