



UvA-DARE (Digital Academic Repository)

Disappearing dissociations in experimental psychology: Using state-trace analysis to test for multiple processes

Stephens, R.G.; Matzke, D.; Hayes, B.K.

DOI

[10.1016/j.jmp.2018.11.003](https://doi.org/10.1016/j.jmp.2018.11.003)

Publication date

2019

Document Version

Final published version

Published in

Journal of Mathematical Psychology

License

Article 25fa Dutch Copyright Act (<https://www.openaccess.nl/en/policies/open-access-in-dutch-copyright-law-taverne-amendment>)

[Link to publication](#)

Citation for published version (APA):

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. <https://doi.org/10.1016/j.jmp.2018.11.003>

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (<https://dare.uva.nl>)



Disappearing dissociations in experimental psychology: Using state-trace analysis to test for multiple processes

Rachel G. Stephens^{a,*}, Dora Matzke^b, Brett K. Hayes^a

^a School of Psychology, University of New South Wales, Sydney NSW 2052, Australia

^b Department of Psychology, University of Amsterdam, Postbus 15906, 1001 NK Amsterdam, Netherlands

HIGHLIGHTS

- Current evidence for multiple psychological processes depends on strong assumptions.
- Much support for dual-processes in reasoning disappears under realistic assumptions.
- The majority of evidence for two systems of category learning similarly disappears.
- State-trace analysis offers a rigorous foundation for claims about latent processes.

ARTICLE INFO

Article history:

Received 1 May 2018

Received in revised form 11 October 2018

Available online 28 December 2018

Keywords:

Dissociations
Dual-process theories
State-trace analysis
Reasoning
Category learning

ABSTRACT

Dissociations have served as a key source of evidence for theory development in experimental psychology. Claims about the existence of multiple distinct psychological processes or systems are often based on demonstrations that manipulations such as working memory load, mood or instructions have differential effects on task performance. For example, a manipulation may have a larger effect on performance in one task, and a smaller or no detectable effect in another, as identified by statistical models like analysis of variance. However, inferring distinct underlying processes based on such interaction effects can be misleading. Such an inference depends on the strong – and probably false – assumption that underlying psychological variables map linearly onto the observable dependent variables. Fortunately, state-trace analysis offers an alternative approach to test for multiple underlying variables, avoiding the linearity assumption. We apply state-trace analysis to databases of studies from reasoning and from category learning that have been cited as evidence for qualitatively distinct processes. We show that many of the dissociations thought to reflect the operation of distinct processes disappear against the stricter criteria of state-trace analysis. We argue that it is important for experiments to be designed with state-trace analysis in mind, and highlight the need for the development and more widespread use of similar techniques. This will lead to a more rigorous foundation for theoretical claims about distinct underlying psychological mechanisms.

© 2018 Elsevier Inc. All rights reserved.

1. Introduction

Psychological science is in the midst of a “crisis of confidence” (Pashler & Wagenmakers, 2012, p. 528). The crisis has been building for quite some time and has been bolstered by large-scale failures to replicate widely-publicized effects (e.g., Doyen, Klein, Pichon, & Cleeremans, 2012; Klein et al., 2014; Open Science Collaboration, 2012; Shanks et al., 2013) and by reports about the prevalence of questionable research practices, such as optional stopping, hypothesizing after the results are known (HARKing), and the selective reporting of statistically significant results (e.g., Bones,

2012; Ioannidis, 2005; John, Loewenstein, & Prelec, 2012; Kerr, 1998; Simmons, Nelson, & Simonsohn, 2011; Wagenmakers, Krypotos, Criss, & Iverson, 2012; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012). These unfortunate events have generated various recommendations aimed at increasing the credibility of the published literature and led to the widespread acknowledgment of the importance of replications and the transparency of the scientific process (Chambers, 2013; Koole & Lakens, 2012; Matzke et al., 2015; Munafò et al., 2017; Nosek & Lakens, 2014).

Replications and research transparency are no doubt necessary to ensure and facilitate the healthy progress of psychological science. However, following Rotello, Heit, and Dubé (2015), we argue that these measures are not sufficient. A more fundamental and insidious issue arises when conclusions – even from highly replicable results – depend on flawed assumptions about the chosen dependent variables. For instance, Rotello and colleagues showed that

* Corresponding author.

E-mail addresses: r.stephens@unsw.edu.au (R.G. Stephens), d.matzke@uva.nl (D. Matzke), b.hayes@unsw.edu.au (B.K. Hayes).

seemingly straightforward conclusions about accuracy differences between conditions in text-book examples involving eyewitness memory, deductive reasoning, and racial biases in suspect discrimination have often depended on faulty dependent variables (Dube, Rotello, & Heit, 2010; Rotello et al., 2015). Such systematic conceptual errors cannot be alleviated by replications and may result in wasteful research effort and misguided theoretical conclusions. In fact, it has been argued that replications with faulty dependent variables may have adverse effects on the field as they increase the number of errors in the literature and may decrease the research community's willingness to explore alternative hypotheses (Rotello et al., 2015).

This paper is concerned with related problems that arise when statistical analysis of observable dependent variables is used to make inferences about the underlying cognitive processes (or systems, representations, modules, etc.). In many areas of psychology, systematic differences in performance between particular tasks within a domain (e.g., memory, reasoning, categorization) are often seen as implying the existence of *multiple* distinct psychological processes or systems (see Melnikoff & Bargh, 2018, for a review). Well-known examples include declarative versus non-declarative memory and learning (e.g., McLaren et al., 2014; Schacter & Tulving, 1994; Squire, 1992), holistic and featural processing of faces (Tanaka & Gordon, 2011), visuo-spatial versus phonological working memory (Baddeley, 2012), distinct modes of thinking for utilitarian versus deontological moral judgments (Paxton & Greene, 2010), and two areas that will be our main focus: intuitive "Type 1" versus reflective "Type 2" reasoning (e.g., De Neys, 2015; Evans & Stanovich, 2013), and explicit rule-based versus procedural category learning (Ashby, Alfonso-Reese, Turken, & Waldron, 1998).

A key form of evidence for inferring multiple psychological processes has been the demonstration of functional dissociations: cases where particular experimental factors have differential effects on performance on two tasks within a domain. For example, according to Evans and Stanovich (2013, p. 232), important evidence for distinct Type 1 versus Type 2 reasoning processes includes the finding that "...belief bias has been shown to be increased and logical accuracy decreased when people operate under time pressure... or concurrent working memory load...". For the time pressure example, they cited an experiment by Evans and Curtis-Holmes (2005), in which participants made binary decisions about the validity of syllogisms that varied in logical validity and in the believability of the conclusion based on prior knowledge (see Table 1 for examples). Decisions were made with or without time pressure. Evans and Curtis-Holmes theorized that assessing arguments based on background knowledge involves relatively fast Type 1 processing, whereas assessment of logical validity requires a slower, deliberative Type 2 processing. Evaluating arguments under time pressure should reduce the contribution of Type 2 processing and increase the contribution of Type 1 processing. Fig. 1a shows their key results. Examining the proportion of arguments endorsed as "valid" (i.e., endorsement rates), they found smaller effects of validity and larger effects of believability under time pressure compared to no pressure.

The general (tacit) rationale behind interpreting such interaction effects as evidence for multiple processes is summarized in Fig. 2a. There are two key observable *dependent variables*, defined by both the performance measure used (e.g., proportion correct, endorsement rates) and an experiment factor that demarcates two tasks, with performance on each task thought to reflect greater involvement of a different kind of process. For instance, in the time pressure example the dependent variables are based on endorsement rates under time pressure or under no time pressure, but can alternatively be based on other performance measures and tasks (e.g., proportion correct under high vs. low working memory load). Thus, under a dual-process reasoning account, it is assumed

that performance on the two dependent variables is driven by two different unobservable *latent variables*, reflecting the output of distinct "Type 1" versus "Type 2" processing. Additionally, it is assumed that reasoning under time pressure will depend more on the "Type 1 processing" latent variable, with response function f translating changes in the latent variable to observable changes in the dependent variable. In contrast, reasoning without time pressure is free to depend more on the "Type 2 processing" latent variable, via response function g . In turn, Type 1 processing is strongly affected by the *independent variable* of believability, while Type 2 processing is more sensitive to the independent variable of validity. Such an interpretation is consistent with observed statistical interactions, such as smaller effects of validity and larger effects of believability under time pressure compared to no pressure. Effects like these are typically revealed using linear models such as Analysis of Variance (ANOVA), or in the case of Evans and Curtis-Holmes (2005), via t-tests of difference scores (i.e., valid–invalid or believable–unbelievable) for time-pressure versus no-pressure groups.

However, the observed interaction effects may also be consistent with a single underlying latent variable, as illustrated in Fig. 2b. In the case of deductive reasoning, both dependent variables may be driven by a single latent variable such as the subjective strength of an argument (cf. Rips, 2001). Crucially, the rejection of this account based on interaction effects identified by linear models depends on the strong – and probably false – assumption that the single latent variable has the same *linear mapping* (i.e., for f and g) onto observed response probabilities for the two dependent variables. We will refer to this as the *linearity assumption*. Indeed, if the response measure has a real limit or bound on at least one end of the scale (as is the case for standard measures like proportion correct, Likert scales, response times, etc.) then the mappings cannot be linear, unless we can assume the same bounding on the latent variable. The assumption of linear mappings has long been criticized (Loftus, 1978; see also Anderson, 1961; Bogartz, 1976), but continues to be tacitly relied upon in the interpretation of interaction effects (Wagenmakers, Krypotos et al., 2012). Instead, often the best we can assume is that the mappings are *monotonic*: as the latent variable increases, there is a monotonically increasing relationship if the dependent variable always increases or stays the same, and a monotonically decreasing relationship if the dependent variable always decreases or stays the same.¹ Thus, the key issue is that linear models such as ANOVA do not distinguish removable interactions – interactions that can be undone by a monotonic transformation of the measurement scale – from non-removable interactions, leading to potentially misleading conclusions about the underlying latent variables. As illustrated by Loftus (1978) and Wagenmakers, Krypotos et al. (2012), if the mappings to response probabilities are actually non-linear, apparent interaction effects may reflect equal shifts in the latent variable across conditions (e.g., across believable and unbelievable conditions).

The troubling consequence is that many interaction effects cited as evidence for multiple processes may have been over-interpreted. However, it is currently unclear *how much* evidence depends on the linearity assumption and thus if this assumption is relaxed so that only monotonic mappings are assumed, how many of the effects remain? Our goal is to address this question and re-examine two separate databases of studies from the research literatures on reasoning and perceptual category learning. The data will be re-analyzed using *state-trace analysis* (Bamber, 1979; Dunn, 2008), which tests for effects that are consistent with non-removable interactions. The two domains were selected because

¹ Note that even if the data are transformed prior to application of ANOVA (e.g., logit transformation, Hélié and Cousineau (2015)), the exact response mapping functions remain unclear.

Table 1
Examples of Syllogisms used by Evans and Curtis-Holmes (2005).

| Validity | Believability of conclusion | |
|----------|---|---|
| | Believable | Unbelievable |
| Valid | No astronauts are unhappy. Some healthy people are unhappy. Therefore some healthy people are not astronauts. | No healthy people are unhappy. Some astronauts are unhappy. Therefore some astronauts are not healthy people. |
| Invalid | No healthy people are unhappy. Some astronauts are unhappy. Therefore some healthy people are not astronauts. | No astronauts are unhappy. Some healthy people are unhappy. Therefore some astronauts are not healthy people. |

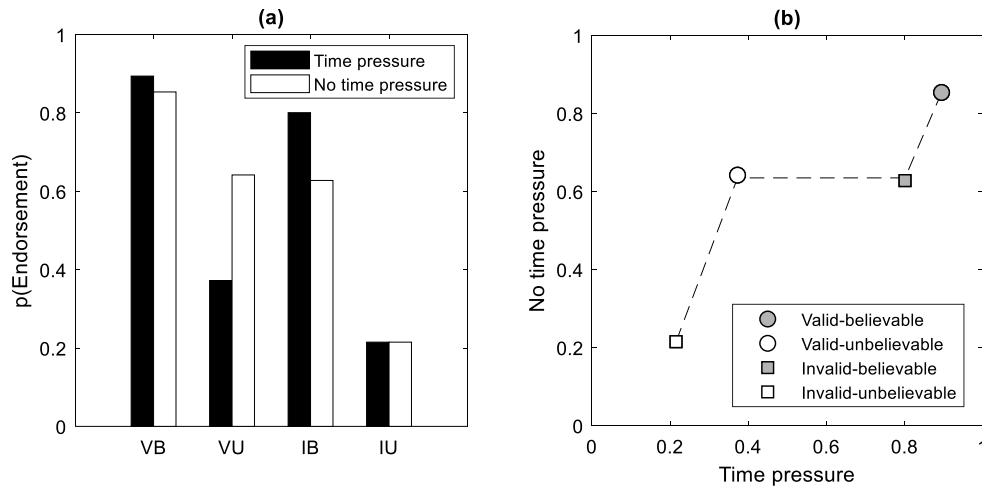


Fig. 1. (a) Dissociation found by Evans and Curtis-Holmes (2005) and (b) a corresponding state-trace plot. Dashed line shows the best-fitting monotonic points. V = valid; I = invalid, B = believable; U = unbelievable.

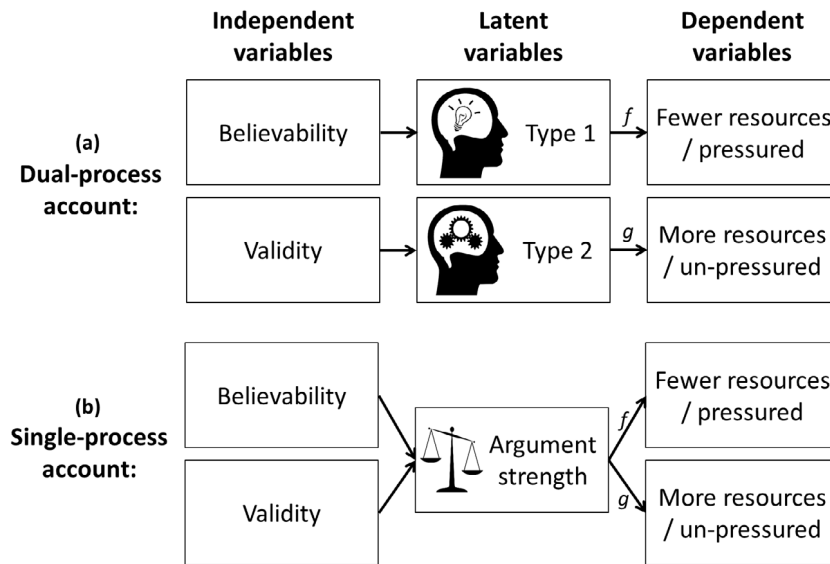


Fig. 2. Two different models of the effects of two independent variables (e.g., believability and validity) on the two dependent variables: reasoning judgments when few versus many cognitive resources are available (e.g., speeded vs. un-speeded tasks; participants with low vs. high working memory capacity). (a) A possible dual-process model, with two underlying latent variables based on the output of Type 1 or Type 2 processing. (b) A single-process model, with one underlying latent variable based on subjective argument strength. For both models, functions f and g map the latent variable(s) onto the dependent variables. State-trace analysis assumes only that these functions are monotonic and not necessarily identical, while standard dissociation logic makes the stronger assumption that they are both linear.

they are more familiar to the authors, but the re-analyses are intended to serve as examples that speak to a widespread issue in experimental psychology.

Each database is based on a list of differential effects or dissociations (i.e., selective effects on only one task), cited by dual-process proponents as key evidence for their theories. The reasoning database includes data from nine papers cited by Evans and

Stanovich (2013) as representing crucial experimental evidence for distinct Type 1 and 2 processing. The category learning database includes data from 28 papers reviewed by Ashby and Valentin (2017), which demonstrate dissociations between so-called “rule-based” and “information-integration” category learning tasks, supporting claims of separate explicit and procedural learning systems. Neither of the databases is intended to be an exhaustive set of studies that have examined the relevant effects, but instead is focused on datasets that have been identified as important for supporting the respective dual-process theories. To foreshadow the results, we find that many of the effects cited as evidence for multiple processes disappear when the linearity assumption is relaxed and state-trace analysis is applied. This highlights the need for a more critical appraisal of dissociation logic and data analysis in psychological science, and for the development and more extensive use of statistical tools like state-trace analysis.

2. State-Trace analysis

State-trace analysis (STA) offers a logically coherent approach to testing whether data across two dependent variables reflect the operation of more than one latent variable (for an accessible introduction to STA, see Newell & Dunn, 2008; for further details and software, see Dunn & Kalish, 2018). Rather than assuming that under a single-process model, each dependent variable is a linear function of the latent variable (as assumed in standard applications of ANOVA), STA relies on the milder assumption that each dependent variable is an unknown (and potentially different) monotonic function of the latent variable (see Dunn, 2008; Dunn, Kalish, & Newell, 2014). Note that this monotonicity assumption implies that, in turn, the two dependent variables must themselves be monotonically related. Thus STA tests for observed violations of a monotonic relationship between the dependent variables. In contrast, the removable interactions often identified via ANOVA do not necessarily imply a non-monotonic relationship between these variables (see Newell & Dunn, 2008). In this sense, STA demands a higher threshold of evidence than ANOVA for inferring the operation of more than one latent variable.

A key tool in STA is the *state-trace plot*, in which the mean of each condition (i.e., averaged over participants, for the current application) for one dependent variable (e.g., validity endorsement rates under time pressure) is plotted against the corresponding mean for the other dependent variable (e.g., validity endorsement rates for no time pressure) — see Fig. 1b for an example, based on the data from Evans and Curtis-Holmes (2005). The crucial question is whether the state-trace is “one-dimensional”, with all data points falling on a single monotonically increasing (or decreasing) curve. If so, the data points are consistent with a single underlying latent variable. For example, changes in endorsement rates across both time-pressure and no-pressure conditions could be driven by shifts in a single common mechanism for assessing argument strength (Fig. 2b). If instead the state-trace is two-dimensional (i.e., some of the data points reliably depart from monotonicity), the data are inconsistent with any model based on a single latent variable. Thus, they may support a dual-process account (e.g., Fig. 2a) — particularly if the violations to monotonicity occur across theoretically relevant experimental manipulations, as predicted by the model. Notice that when the means from Evans and Curtis-Holmes (2005) are re-plotted in a state-trace plot in Fig. 1b, they fall very close to a monotonic curve, shown by the dashed line. Thus, although these data have been interpreted as supporting a distinction between Type 1 and 2 processing, they are not compelling evidence against a single-process account.

To test whether there is statistically significant evidence of a two-dimensional state-trace, the conjoint monotonic regression (CMR) test recently developed by Kalish, Dunn, Burdakov, and

Sysoev (2016); also see Dunn and Kalish (2018) can be applied. In essence, this test examines whether all of the experimental conditions are in the same order (e.g., in Fig. 1b: invalid-unbelievable \leq valid-unbelievable \leq invalid-believable \leq valid-believable) across a set of dependent variables (e.g., endorsement rates for time-pressure vs. no-pressure). First, the best-fitting (maximum likelihood) monotonic approximation of the observed data is found (shown as the dashed line in Fig. 1b) via a custom optimization algorithm. Second, a frequentist hypothesis test is conducted where the null hypothesis is that the monotonic approximation (consistent with a single latent variable) provides an adequate fit to the data. In this test, a *p*-value is obtained by comparing the observed fit of the single-latent-variable model to a bootstrap sample distribution of fit values obtained under the hypothesis that this model is true.

3. Reasoning database analysis

In this section, we apply state-trace analysis to the experimental evidence for dual-process theories of reasoning cited by Evans and Stanovich (2013). Although there are multiple variants of dual-process theories, for example with Type 1 and 2 processing proceeding sequentially (Evans, 2008; Kahneman & Frederick, 2002), or in parallel (e.g., Handley & Trippas, 2015; Sloman, 1996, 2014), they commonly assume that there are key characteristics that differentiate the two types of processing. Specifically, reflective Type 2 processing is distinguished from intuitive Type 1 processing by its reliance on working memory resources and that it is involved with mental simulation or hypothetical thinking (Evans & Stanovich, 2013). Based on these characteristics, Evans and Stanovich offered a set of “illustrative examples” of experimental manipulations that “make the case for qualitatively distinct types of processing...” (p. 232). These manipulations include increasing Type 2 processing effort via instruction, or by suppressing it via concurrent working memory tasks or response deadlines that allow little time for reflective thought.

We are not the first to question the evidence used to support dual-process accounts of reasoning. Some critiques have been conceptual (e.g., Keren & Schul, 2009; Kruglanski & Gigerenzer, 2011; Osman, 2004, 2013), while others have involved making stronger assumptions about response mappings and testing quantitative models for particular tasks (e.g., Lassiter & Goodman, 2015; Rotello & Heit, 2009). Our use of state-trace analysis, however, allows us to go beyond these critiques and quantify the extent to which a broader range of key behavioral evidence for dual-process accounts has depended upon the — probably false — linearity assumption.

3.1. Method

We extracted data from nine papers (16 datasets; only studies with reported condition means could be included) cited by Evans and Stanovich (2013) in their *Experimental manipulations* section (p. 232). Details of the experiments are listed in Table 2 and the database is available in the supplemental materials. The datasets included three different types of tasks: (1) a validity judgment task for syllogisms or causal conditional arguments, (2) the Wason selection task, and (3) the conjunction fallacy task. Accordingly, a range of performance measures were used for pairs of dependent variables, including proportion correct (e.g., based on the acceptance of valid arguments and rejection of invalid arguments), the proportion of endorsements as “valid”, or the proportion of selections of an option (e.g., “matching bias” choices in the Wason selection task).

Although in standard statistical analyses, all experiment factors are simply included as factors in ANOVA, in STA a factor must be selected to define the two dependent variables (and thus the two axes of the state-trace plot). In testing dual-process accounts of reasoning, this should be the factor that – motivated by theory – is thought to most strongly reflect differential contributions of Type 1 and Type 2 processing (e.g., responding under different levels of working memory load or capacity; see Fig. 2a). The other experiment factors are then treated as independent variables that can influence the latent variable(s) (see Fig. 2a), and define different conditions (or data points in the state-trace plot) across the two dependent variables. Hence, in Fig. 1b, the time-pressure factor defined the dependent variables, and validity and believability factors defined the four conditions.

For each experiment in our database, the factor chosen to define the two dependent variables was based on the theoretical factors highlighted by Evans and Stanovich (2013) as enhancing or suppressing Type 2 processing. These factors were: high/low working memory load, high/low working memory capacity or cognitive ability, speeded/unspeeded judgments, pragmatic/deduction instructions, and naïve/expert reasoners. For the experiments of De Neys (2006b) and De Neys, Schaeken, and d'Ydewalle (2005b), both working memory load and capacity factors were available to use as sensible dependent variables, so we analyzed the data in two different arrangements, such that each factor could define the dependent variables in turn (see Table 2, signaled as “factors inverted”). Note that Evans and Stanovich did not specifically describe a critical factor for the experiments reported in Evans, Handley, and Harper (2001) nor in Klauer, Musch, and Naumer (2000), but we included their data for completeness, and chose the most sensible factor for the dependent variables: pragmatic/deduction instructions, and naïve/expert reasoners, respectively. After selecting the factor for the dependent variables, the remaining experiment factors defined the conditions; this included believability, argument or problem type, perceived base rate, working memory load, and working memory capacity.

The means, sample sizes (N_s) and – where available – standard deviations (SDs) for each experimental condition were extracted from tables or figures (using plot digitizing software; with scores set to a 0–1 proportion scale for consistency) and *Participants* sections in each publication. These summary statistics were used by the CMR algorithm to find the best-fitting monotonic function, and the SDs were used to estimate the variability of the data for the bootstrapping procedure – thus our bootstrapping was necessarily parametric, assuming that observations were normally distributed for each dependent variable in each condition. However, most of the papers in the reasoning database did not report SDs (nor standard errors [SEs]) for the relevant experimental conditions, so for these datasets we performed the CMR tests based on three levels of possible SDs : .10, .20 and .30. These SDs are plausible (with .10 being optimistic) based on the mean of reported SDs from the validity judgment experiments of Evans, Handley, Neilens, and Over (2010) of .25, the De Neys et al. (2005b) mean SD of .15, and the SDs in our own reasoning data (e.g., Stephens, Dunn, & Hayes, 2018b, Experiment 1 mean SD of .21). We also had to use approximate cell Ns for each condition of De Neys (2006b) and Experiment 3 of De Neys (2006a); in these studies only the total experiment Ns were reported, so we assumed approximately even allocation to conditions (see supplemental materials for the values we used).

3.2. Results and discussion

The results of the CMR test for each dataset are shown in Table 2, and sample state-trace plots are shown in Fig. 3, along with the best-fitting monotonic functions (state-trace plots for all datasets

are included in the supplemental materials). Many of the dissociations identified in the original ANOVA (or t-test, etc.) analyses disappear under the stricter criteria of STA, with the majority of datasets failing to indicate a two-dimensional state-trace. These null state-trace results include the Evans and Curtis-Holmes (2005) data discussed above (Fig. 1), and the De Neys (2006b) data, in which there had been an observed interaction between syllogism type (i.e., whether or not validity conflicted with believability) and working memory load, and between syllogism type and working memory span (e.g., see Fig. 3a, and Fig. 3b for null results from Evans, Handley, Neilens, & Over, 2010). The CMR results for the three datasets with known SDs are all non-significant, indicating that only a single latent variable is needed to explain the data. If we assume optimistically low SDs of .10 for the datasets with unknown SDs , only six out of the 16 tests (37.5%) are statistically significant. These “positive” results suggesting more than one latent variable are from De Neys et al. (2005b), Klauer et al. (2000) and Roberts and Newton (2001) (e.g., see Fig. 3c-d). If, however, we assume SDs of .20 or .30 for these datasets, none of the CMR tests return a significant result.

Although many reliable interaction effects were reported from the original statistical analyses of studies in the reasoning database, reflecting the functional dissociations predicted by dual-process accounts, STA found limited evidence in favor of multiple latent variables. Crucially, one-dimensionality could not be rejected even for several datasets that included key factors that should reflect a core distinction between Type 1 and 2 processing: working memory load (De Neys, 2006a, 2006b; De Neys, Schaeken, & d'Ydewalle, 2005a), working memory span (De Neys, 2006b), and cognitive ability (Evans et al., 2010). It appears that much of the experimental evidence cited by Evans and Stanovich (2013) as supporting two qualitatively distinct types of processing depends upon the linearity assumption; much of the evidence does not hold up against STA. Therefore, under more realistic assumptions, the data do not typically rule out alternative accounts under which a single latent variable drives changes in performance across conditions. Such a variable could be based on the subjective strength of a conclusion in deductive reasoning tasks, or the strength of a possible response option in Wason selection tasks or conjunction fallacy tasks. For example, drawing on a signal detection framework (see e.g., Rips, 2001; Stephens, Dunn, & Hayes, 2018a), patterns of validity judgments in the deductive reasoning experiments may be driven by a single parameter that captures people's ability to discriminate valid and invalid inferences (i.e., the relative strength of valid and invalid arguments); this discriminability is simply increased or decreased by factors such as time pressure, working memory load, and degree of expertise.

Nevertheless, our results do depend to some extent on the assumed SDs , so further examination and replication is warranted based on the potentially “positive” results from De Neys et al. (2005b), Klauer et al. (2000) and Roberts and Newton (2001). For readers interested in this reasoning domain, we briefly describe each of these varied datasets in turn. First, the De Neys et al. (2005b) Experiment 1 data involved conclusion evaluations of causal conditional arguments, for participants with low versus high working memory span, as measured by an operation span task. The three Klauer et al. (2000) datasets were based on validity judgments of categorical syllogisms, for naïve versus expert reasoners, and included manipulations of the perceived base rate of valid arguments. As noted above, Evans and Stanovich (2013) did not cite the Klauer et al. (2000) data as a comparison of naïve versus expert reasoning, so this is possibly a novel and interesting result. Lastly, the two Roberts and Newton (2001) datasets examined the numbers of card selections in variants of the Wason task, for those with or without time pressure.

Table 2
Details of the reasoning database and the State-Trace Analysis results.

| Source | Task | Dependent variable 1 | Dependent variable 2 | Factors that form the conditions | <i>p</i> (if SDs were reported) | <i>p</i> (SDs = .10) | <i>p</i> (SDs = .20) | <i>p</i> (SDs = .30) |
|---|--|---|---|--|---------------------------------|----------------------|----------------------|----------------------|
| De Neys (2006a) Experiments 2 & 3 | Conjunction fallacy | Proportion correct - WM load (Tapping or Dot memory) | Proportion correct - No/low WM load | 1) Linda problem 1/Bill problem/Linda problem 2/Job problem | – | .249 | .441 | .549 |
| De Neys (2006a) Experiments 2 & 4 | Wason selection task | “Matching bias” choices - WM load (Tapping or Dot memory) | “Matching bias” choices - No/low WM load | 1) Indicative/Deontic 2) Experiment 2/4 | – | .089 | .203 | .262 |
| De Neys (2006b) | Argument evaluation (syllogisms) | Proportion correct - High WM load | Proportion correct - No WM load | 1) No conflict/Conflict 2) Low/medium/high span | – | .688 | .850 | .923 |
| De Neys (2006b) [with span/load factors inverted] | Argument evaluation (syllogisms) | Proportion correct - Low WM span | Proportion correct - High WM span | 1) No conflict/Conflict 2) High/low/no load | – | .118 | .456 | .677 |
| De Neys et al. (2005a) | Counterexample generation for conditionals | Examples generated - High WM load | Examples generated - No WM load | 1) Weak/strong/ strongest associates | – | 1.000 | 1.000 | 1.000 |
| De Neys et al. (2005b) Experiment 1 | Argument evaluation (conditionals) | Confidence ratings - Low WM span | Confidence ratings - High WM span | 1) Many/few counterexamples 2) MP/DA/MT/AC | – | .043 | .472 | .769 |
| De Neys et al. (2005b) Experiments 1–2 | Argument evaluation (conditionals) | Confidence ratings - Low WM span | Confidence ratings - High WM span | 1) No load/load 2) MP/DA/MT/AC 3) Few/many alternatives, few/many disablers | .117 | – | – | – |
| De Neys et al. (2005b) Experiments 1–2 [with span/load factors inverted] | Argument evaluation (conditionals) | Confidence ratings - WM load | Confidence ratings - No WM load | 1) Low/high span 2) MP/DA/MT/AC 3) Few/many alternatives, few/many disablers | .414 | – | – | – |
| Evans and Curtis-Holmes (2005) | Argument evaluation (syllogisms) | Endorsement rate - Time pressure | Endorsement rate - No time pressure | 1) Valid/invalid 2) Believable/unbelievable | – | .267 | .397 | .501 |
| Evans et al. (2001) | Argument evaluation (syllogisms) | Endorsement rate - Possibility instructions | Endorsement rate - Necessity instructions | 1) Believable/abstract/unbelievable 2) Necessary/possible strong/possible weak/impossible | – | .671 | .932 | .985 |
| Evans et al. (2010) | Argument evaluation (conditionals) | Endorsement rate - Low cognitive ability | Endorsement rate - High cognitive ability | 1) Deductive/pragmatic 2) High/low belief 3) MP/DA/AC/MT | .529 | – | – | – |
| Klauer et al. (2000) Study 2 | Argument evaluation (syllogisms) | Endorsement rate - Naïve reasoners | Endorsement rate - Logic experts | 1) Believable/Unbelievable 2) Low/medium/high perceived base rate 3) Valid/Invalid | – | .002 | .162 | .448 |

(continued on next page)

Table 2 (continued).

| Source | Task | Dependent variable 1 | Dependent variable 2 | Factors that form the conditions | <i>p</i> (if SDs were reported) | <i>p</i> (SDs = .10) | <i>p</i> (SDs = .20) | <i>p</i> (SDs = .30) |
|--|----------------------------------|--|---|--|---------------------------------|----------------------|----------------------|----------------------|
| Klauer et al. (2000) Study 4 | Argument evaluation (syllogisms) | Endorsement rate - Naïve reasoners | Endorsement rate - Logic experts | 1) Believable/Unbelievable 2) Low/medium/high perceived base rate 3) Valid/Invalid | – | .009 | .211 | .484 |
| Klauer et al. (2000) Study 6 | Argument evaluation (syllogisms) | Endorsement rate - Naïve reasoners | Endorsement rate - Logic experts | 1) Believable/Unbelievable 2) Low/medium/high perceived base rate 3) Valid/Invalid | – | <.001 | .117 | .474 |
| Roberts and Newton (2001) Experiment 2 | Wason selection task | Proportion of selections - Time pressure | Proportion of selections - No time pressure | 1) 4 problem types 2) True/false antecedent, true/false consequent | – | <.001 | .104 | .555 |
| Roberts and Newton (2001) Experiment 3 | Wason selection task | Proportion of selections - Time pressure | Proportion of selections - No time pressure | 1) 4 problem types 2) True/false antecedent, true/false consequent | – | .025 | .611 | .910 |

Note. Significant *p*-values ($p < .05$) are in bold font. WM = working memory; MP = modus ponens; DA = denying the antecedent; MT = modus tollens; AC = accepting the consequent.

4. Category learning database analysis

We now turn to the domain of perceptual category learning, using STA to re-examine a database of studies cited by [Ashby and Valentin \(2017\)](#) as key behavioral evidence for two distinct learning systems. According to the COVIS model (Competition between Verbal and Implicit Systems; e.g., [Ashby et al., 1998](#); [Ashby, Paul, & Maddox, 2011](#); [Cantwell, Crossley, & Ashby, 2015](#)), category learning is driven by two competing neurobiological systems: an explicit or verbal hypothesis-testing system that depends heavily on working memory and executive attention, and a procedural system that is relatively automatic and depends on a reward signal to strengthen associations between stimuli and responses. This distinction has been supported by studies comparing performance on *rule-based* (RB) and *information-integration* (II) category learning tasks. Typical stimuli for these tasks have been Gabor patches, circular sine-wave gratings that vary in bar width (i.e., spatial frequency) and orientation. In the simplest design, the stimuli are divided into two categories. RB structures are amenable to easily verbalizable rules, so it is thought that they can be solved successfully by the verbal system. For example, the RB category boundary is a vertical division along the width/frequency dimension, so that items with a lower frequency are assigned to Category A, and the items with a higher frequency are assigned to Category B (e.g., [Maddox, Ashby, & Bohil, 2003](#)). In contrast, II structures are defined by a complex combination of properties on at least two stimulus dimensions (e.g., a weighted linear combination of width and frequency values). Since no simple verbal rule correctly parses such stimuli into the correct categories, it is claimed that the procedural system will take over the categorization process for such tasks.

As critical behavioral evidence for these distinct explicit and procedural systems, a large number of studies have identified various relevant functional dissociations: factors that are thought to differentially affect the two systems, and thus the learning performance on RB and II tasks. [Ashby and Valentin \(2017\)](#) presented a long list of such empirical results, stating that collectively they “provide strong evidence that learning in these tasks is mediated by separate systems” (p. 175). In the domain of category learning, the rationale behind this argument maps onto [Fig. 2a](#) as follows. The dependent variables are learning performance (proportion of correct classification) on the RB and II tasks. The latent variables are the explicit and procedural learning systems, with the explicit system primarily driving RB learning, and the procedural system primarily driving II learning. The independent variables include the factors identified by [Ashby and Valentin \(2017\)](#), such as self-regulation depletion (before the category learning tasks, participants write a story without using the letters “a” and “n” vs. write a story with no restrictions; [Minda & Rabi, 2015](#)) and sequencing of training difficulty (i.e., first learn exemplars that are near vs. far from the category boundary; [Spiering & Ashby, 2008](#)).

Previously, some of the evidence used to support COVIS has been challenged based on issues such as the learning criterion adopted ([Tharp & Pickering, 2009](#)), the inclusion of non-learners ([Newell, Dunn, & Kalish, 2010](#)), collapsing across subgroups of learners ([Stephens & Kalish, 2018](#)), and the relative perceptual discriminability of RB versus II stimuli ([Stanton & Nosofsky, 2007](#)) – see [Newell, Dunn, and Kalish \(2011\)](#) for a review. However, our focus is on the more general issue that – as in the reasoning domain – the dissociation evidence has often been based on interactions identified by linear models like ANOVA, which rely on the linearity assumption. If this assumption is relaxed and STA is applied, a single underlying latent variable may also often account for RB and II learning. Indeed, [Newell and colleagues](#) have demonstrated that some of the dissociations reported as support for the COVIS

model disappear under STA (e.g., [Dunn, Newell, & Kalish, 2012](#); [Newell et al., 2011](#)). Our goal is to use STA to perform a more comprehensive examination of the extent to which evidence for COVIS depends on the linearity assumption.

4.1. Method

We extracted data from 28 of the 38 papers cited by [Ashby and Valentin \(2017\)](#); 63 datasets were extracted). Across three tables (their Tables 7.1 to 7.3), they listed empirical dissociations that: (1) affect II learning more than RB learning, (2) affect RB learning more than II learning, or (3) affect each in “different (often opposite) ways” (p. 175; capitalization ignored). Only studies with reported condition means for proportion correct for both RB and II tasks could be included; nine papers were excluded because they failed to meet this criterion. We also excluded the data from [Dunn et al. \(2012\)](#) because STA had already been applied, and the only two-dimensional state-trace out of the four experiments reported in that paper was subsequently found to be an artifact of averaging over subgroups of learners (see [Stephens & Kalish, 2018](#)).

The datasets that we re-analyzed are summarized in [Tables 3–5](#), and the database is available in the supplemental materials. The dependent variables were RB and II learning performance (proportion of correct classification), mostly based on 2-category structures, though some were 4-category structures (i.e., so chance-level accuracy is .5 or .25, respectively). Eleven of the datasets based on 2-category structures involved a more complex conjunctive rule for the RB task, involving both stimulus dimensions (e.g., the correct rule was “Respond A if the bars are wide AND the orientation is steep; otherwise, respond B”; e.g., see [Spiering & Ashby, 2008](#)). The remaining experiment factors were treated as independent variables (see [Tables 3–5](#)). The stimuli for most studies were Gabor patches, though some used single lines that varied in length and orientation, Munsell color patches that varied in brightness and saturation, or yellow pixels and tones.

To be as inclusive as possible, sometimes particular scores were re-used across multiple conditions either within or between datasets from a given publication, because of the possible ways that RB and II tasks could be sensibly paired together. In other words, we simply treated the relevant conditions as having been notionally repeated within or between experiments. These instances are signaled in the supplemental materials. For example, [Ashby, Queller, and Berretty \(1999\)](#) included two versions of the RB tasks (categories are defined by line length vs. by orientation) and of the II tasks (diagonal category boundaries with positive vs. negative slope), so we performed four STA tests in which we paired each RB task with each II task. Similarly, [Maddox et al. \(2003\)](#) included both RB and II tasks in Experiment 1, but only follow-up RB conditions in Experiment 2, so for the Experiment 2 data we paired the RB scores with the corresponding II scores from Experiment 1.

The means, cell *Ns* and *SEs* were extracted from tables or figures (using plot digitizing software; with scores set to a 0–1 proportion scale for consistency) and *Participants* sections in each publication. Again, the *SEs* were used for parametric bootstrapping (*SEs* were converted to *SDs*). Unlike the reasoning database, relevant measures of variability (cell *SEs*) were reported in most of the category learning papers. For the few papers that did not report these statistics, we used a *SE* of 0.02 (see [Tables 3–5](#)), which was the mean score based on the other studies included in the database. We also had to use approximate *Ns* for conditions of five papers ([Ell, Cosley, & McCoy, 2011](#); [Grimm & Maddox, 2013](#); [Markman, Maddox, & Worthy, 2006](#); [Nadler, Rabi, & Minda, 2010](#); [Zeithamova & Maddox, 2006](#)); only the total experiment *Ns* were reported in these papers, so we assumed approximately even allocation to conditions (see supplemental materials for the values we used).

Table 3
 Details of the category learning database and the State-Trace Analysis results: Manipulations that affect Information-Integration Learning.

| Manipulation | Source | N. cat. | Stimuli | Exp | Factors that form the conditions | Dataset details (if applicable) | <i>p</i> | \wedge # <i>a</i> |
|------------------------|-----------------------------------|---------|---|----------|--|---------------------------------|------------|---------------------|
| Unsupervised learning | Ashby et al. (1999) | 2 | Single line: length \times orientation | 1A | 1) Block | RB-length vs. II-positive | .19 | \wedge |
| | | | | | | RB-length vs. II-negative | .43 | \wedge |
| | | | | | | RB-orientation vs. II-positive | .87 | \wedge |
| | | | | | | RB-orientation vs. II-negative | .48 | \wedge |
| | Ell, Ashby, and Hutchinson (2012) | 2 | Munsell color patches: brightness \times saturation | 1 | 1) Block 2) Day | RB-vertical vs. II-positive | .53 | |
| | | | | | | RB-vertical vs. II-negative | .79 | |
| | | | | | | RB-horizontal vs. II-positive | .86 | |
| | | | | | | RB-horizontal vs. II-negative | .91 | |
| | 2 | | 2 | 1) Block | RB-vertical vs. II-positive | .63 | | |
| | | | | | RB-vertical vs. II-negative | .56 | | |
| | | | | | RB-horizontal vs. II-positive | 1.00 | | |
| | | | | | RB-horizontal vs. II-negative | .87 | | |
| Observational training | Ashby, Maddox, and Bohil (2002) | 2 | Single line: length \times orientation | 1 | 1) Observational/feedback training 2) Block | | .85 | \wedge |
| | | | | | 1) Observational/feedback training 2) Response/No response | | .41 | |
| Delayed feedback | Maddox et al. (2003) | 2 | Gabor: frequency \times orientation | 1 | 1) Delayed/immediate feedback 2) Delay duration 3) Block | | .01 | |
| | | | | | 1) Delayed/immediate feedback 2) Block | Collapsed across duration | .75 | |

(continued on next page)

Table 3 (continued).

| Manipulation | Source | N. cat. | Stimuli | Exp | Factors that form the conditions | Dataset details (if applicable) | <i>p</i> | \hat{a} |
|----------------------------|---|---------|--|-----|---|---------------------------------|-----------------|-----------|
| | | | | 1–2 | 1) Delayed/immediate feedback 2) Block | $d' = 3.5$ | .14 | |
| | | | | 1–2 | 1) Delayed/immediate feedback 2) Block | $d' = 4.6$ | .19 | |
| | Maddox and Ing (2005) | 4 | Gabor: frequency \times orientation | 1 | 1) Delayed/immediate feedback 2) Block | | .58 | |
| Response mapping | Ashby, Ell, and Waldron (2003) | 2 | Single line: length \times orientation | 1 | 1) Control/Hand-switch/Button-switch 2) Block | | <.001 | |
| | | | | 2 | 1) Control/Button-switch 2) Block | | .39 | |
| | Maddox, Bohil and Ing (2004) | 2 | Gabor: frequency \times orientation | 1 | 1) $Y - N/A - B$ training 2) Block | | .51 | |
| | Maddox, Lauritzen, and Ing (2007) | 2 | Single line: length \times orientation | 1 | 1) Control/Button-switch 2) Block | | .81 | |
| | | | | 2 | 1) Switch probability 2) Block | | .98 | |
| Category label/location | Maddox, Glass, O'Brien, Filoteo, and Ashby (2010) | 4 | Single line: length \times orientation | 1–2 | 1) Control/Category label/Response location 2) Block | | .84 | |
| Analogical transfer | Maddox, Filoteo, Lauritzen, Connally, and Hejl (2005) | 2 | Single line: length \times orientation | 1 | 1) Discontinuous/no spread 2) Block | | .90 | |
| | Casale, Roeder, and Ashby (2012) | 2 | Gabor: frequency \times orientation | 1 | 1) Analogical transfer/Control 2) Block | | .11 | |
| | | | | 2 | 1) Block | | .002 | |
| | | | | 3 | 1) Block | | <.001 | |
| Transfer to same-different | Hélie and Ashby (2012) | 2 | Gabor: frequency \times orientation | 1 | 1) Block | RB-width vs. II | .98 | # |
| | | | | 1 | 1) Block | RB-orientation vs. II | .83 | # |

(continued on next page)

Table 3 (continued).

| Manipulation | Source | N. cat. | Stimuli | Exp | Factors that form the conditions | Dataset details (if applicable) | <i>p</i> | $\hat{\#}^a$ |
|-----------------------------------|---------------------------------------|--------------|--|-----|--|--|-----------------|--------------|
| | | 2 (conj.) | | 2 | 1) Block | RB-width vs. II; Transfer to same/different | .96 | $\hat{\#}$ |
| | | | | 2 | 1) Block | RB-width vs. II; Transfer to A/B | .98 | $\hat{\#}$ |
| | | | | 2 | 1) Block | RB-conj. vs. II; Transfer to same/different | .81 | $\hat{\#}$ |
| | | | | 2 | 1) Block | RB-conj. vs. II; Transfer to A/B | .95 | $\hat{\#}$ |
| Category discontinuity | Maddox, Filoteo, and Lauritzen (2007) | 4 | Single line: length \times orientation | 1 | 1) Continuous/Discontinuous 2) Block | | .92 | |
| | | 2 (com.) | | 2 | 1) Simple/Complex 2) Block | | .98 | |
| Sequencing of training difficulty | Spiering and Ashby (2008) | 2 (conj.) | Gabor: frequency \times orientation | 1–2 | 1) Hard-Easy/Easy-Hard/Random 2) Block | | .004 | |
| | | | | 1–2 | 1) Hard-Easy/Easy-Hard/Random 2) Training/Transfer 3) Easy/Medium/Hard | | <.001 | |
| Category separation | Ell and Ashby (2006) | 2 | Gabor: frequency \times orientation | 1–3 | 1) Medium-low/Medium/Medium-high/High 2) Block | | .02 | \hat{a} |

Note. N. cat. = number of categories; Exp = Experiment; conj./com. = conjunction/complex structure used for the RB task. $\hat{\#}$ = SEs were not reported, so were set to .02. # = included same/different judgments rather than classification judgments. *a* = approximate CMR fits were used. Significant *p*-values ($p < .05$) are in bold font.

Two datasets (i.e., Grimm & Maddox, 2013, with 48 conditions and Eil & Ashby, 2006, with 148 conditions) were too large to find the optimal solution using the full CMR algorithm developed by Kalish et al. (2016). The problem is one of combinatorial explosion – there were simply too many combinations of data points to check. Instead, we adopted an approximate solution that was based on resolving any violations to monotonicity independently, rather than conjointly with every other violation. While this produces a solution that is consistent with a single latent variable, it is not necessarily the best possible fit (note that failing to find the best solution would produce a bias in favor of multiple latent variables). However, Stephens et al. (2018a) found that the approximate solution was consistently close (or even identical) to the optimal solution. Applying the approximate fit algorithm, we could then use the same bootstrap statistical test as described by Kalish et al. (2016) for optimal fits.

4.2. Results and discussion

The results of the CMR test for each dataset are shown in Tables 3–5, and sample state-trace plots are shown in Fig. 4, along with the best-fitting monotonic functions (state-trace plots for all datasets are included in the supplemental materials). Again, many of the dissociations disappear when STA is applied, with the datasets failing to indicate a two-dimensional state-trace (e.g., Fig. 4a–b). For example, this includes data from H elie and Cousineau (2015), in which there had been significant interactions between RB/II task and the duration with which exemplars were presented, and between task and block within some levels of presentation duration (Experiment 1). They had also identified interactions between task and the degree of opacity of a mask presented over exemplars (Experiment 2). The key finding is that overall, only 10 out of 63 datasets (15.9%) return a significant result on the CMR test.

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). As with the reasoning database, the STA results suggest that much of the cited experimental evidence depends upon the linearity assumption. Instead of two learning systems, a single latent variable such as “degree of learning” – or a dimension-weighting parameter as included in the Generalized Context Model of categorization (Nosofsky, 1986) – would often be sufficient to account for the results. For readers interested in this category learning domain, in the next section we separately examine the results for dissociations said to: (1) affect II learning more than RB learning, (2) affect RB learning more than II learning, or (3) affect each in different (often opposite) ways. Note that the first two kinds of dissociations have been referred to as “single dissociations”, and the third as “double dissociations”, which are traditionally seen as stronger evidence for separate processes, especially if performance on the two tasks is affected in opposite directions (see Dunn & Kirsner, 1988).

4.2.1. Dissociations that primarily affect II

For the dissociations that primarily influence II learning (Table 3), the STA CMR tests find evidence for more than one latent dimension in only seven of the 40 (17.5%) datasets. Notably, STA finds no evidence of a selective effect on II performance for many of the factors included in the Ashby and Valentin (2017) review. These null results are found for manipulations of unsupervised learning, observational training, shifts in category label and response location, transfer of category learning to same/different judgments, and category discontinuity (within-category subgroups or clusters of exemplars).

Table 3 shows that three other factors (delayed feedback, switching response mappings between training and test, “analogical transfer” of learning to novel exemplars) produce inconsistent evidence for a two-dimensional state-trace. For example, the dataset from Ashby et al. (2003) Experiment 1, which is based on switching response mappings in a transfer phase (participants were asked to reverse which hands or buttons corresponded to Category A/B responses in training/test), produces reliable evidence for a two-dimensional state-trace. However, this result is not replicated when we analyze data from a follow-up experiment from the same paper, which attempted to equate pre-transfer expertise across RB and II tasks. Likewise, no evidence of more than one latent variable is found in other studies manipulating response mapping (Maddox, Bohil et al., 2004; Maddox, Lauritzen et al., 2007).

Two examples of dissociations that primarily affect II may be more promising as evidence of two-dimensionality. Spiering and Ashby (2008, e.g., see Fig. 4c) demonstrate effects of the sequencing of training difficulty (based on stimuli near to or far from the category boundary) for the two ways they presented their data (conditions broken down by block in their Figures 2 and 5, or by item difficulty in their Figures 3 and 6). Also, Eil and Ashby (2006) manipulated category separation (the between-category overlap of exemplars) and show a two-dimensional state-trace when we examine their three experiments (which were each RB or II only) as one large dataset.

4.2.2. Dissociations that primarily affect RB

For the effects that were listed as dissociations that primarily affect RB learning (Table 4), only two of the ten (20%) datasets show evidence for more than one dimension according to STA. The significant instances are from two of the three datasets from Eil et al. (2011), in which participants were presented with a social stressor task immediately before the category learning task (a third dataset from the same study produced a marginal result on the CMR test, $p=.07$). All three datasets are from one experiment, but participants were split into groups in different ways, based on their physiological responses or threat-appraisal ratings, suggesting some consistency across different ways of assessing stress. Notably, however, no evidence for more than one dimension is found in STA tests of any of the other factors listed in the table.

4.2.3. Dissociations that affect RB and II in different ways

Finally, we turn to the effects that were listed by Ashby and Valentin (2017) as double dissociations that differentially affect RB and II (Table 5). This set of factors should represent the highest proportion of significant results under STA, if indeed they affect RB and II learning in “opposite” ways; this is the type of pattern most likely to produce a non-monotonic state-trace. However, only one of the 13 (7.7%) datasets show significant evidence of more than one latent variable according to STA. The one significant instance is from Markman et al. (2006), which contributed only one dataset (see Fig. 4d). The key manipulation was that some participants were put under performance-pressure; they were told that they must exceed a performance criterion in order for both themselves and a “partner” to receive a monetary bonus. This performance-pressure aided II learning but impaired RB learning. In contrast, we find that the various other factors listed in the table do not dissociate RB and II learning according to STA.

5. General discussion

As part of recent attempts to improve the credibility of the published literature in psychological science, we have argued that researchers must also carefully consider the assumptions underlying inferences about psychological mechanisms, processes or systems.

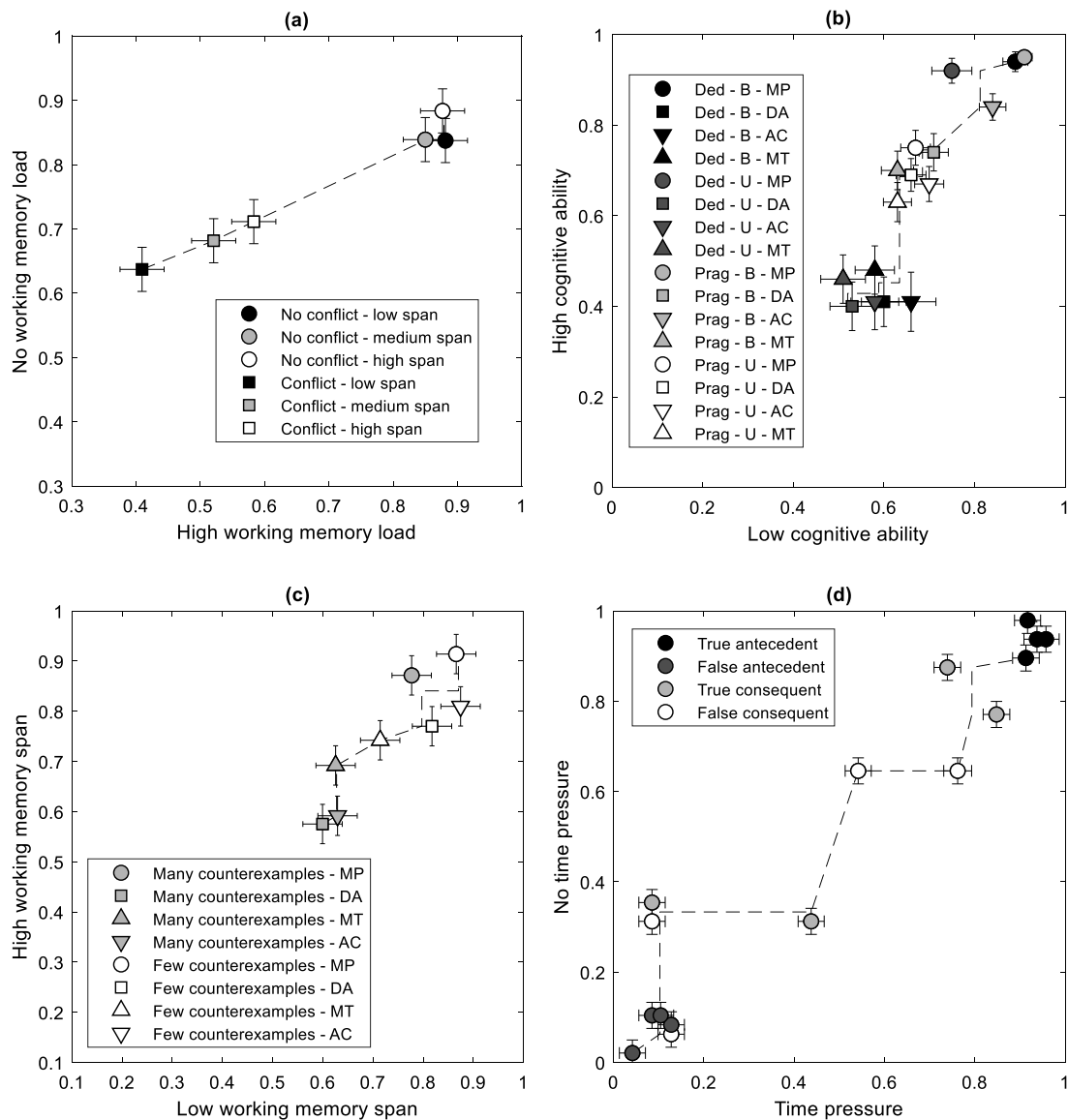


Fig. 3. State-trace plot for four sample datasets in the reasoning database: (a) De Neys (2006b); (b) Evans et al. (2010); (c) De Neys et al. (2005b) Experiment 1; (d) Roberts and Newton (2001) Experiment 2. Error bars show SEs, which are based on assuming $SD = .2$ for plots (a), (c), and (d). Each dashed line shows the best-fitting monotonic points. The null hypothesis of a one-dimensional state-trace cannot be rejected for plots (a) and (b), but is rejected for plots (c) and (d) if $SD = .1$. (Ded = deductive; Prag = pragmatic; B = believable, U = unbelievable; MP = modus ponens; DA = denying the antecedent; MT = modus tollens; AC = accepting the consequent.)

The research community's focus on replicability and transparency is necessary, but it does not address conceptual errors that arise when common statistical analyses of observable dependent variables are used to make inferences about the underlying cognitive processes. To illustrate the potential impact of such errors, we re-analyzed two databases of studies previously cited as strong behavioral evidence for distinct kinds of processes underlying reasoning (by Evans & Stanovich, 2013), and for distinct systems of category learning (by Ashby & Valentin, 2017). Much of this evidence has relied on interaction effects as identified by linear models such as ANOVA. Our main finding was that when the linearity assumption is relaxed and state-trace analysis is applied, the evidence favoring multiple latent variables or processes is greatly reduced. Our results show that even replicable effects such as the effect of working memory resources on deductive reasoning do not necessarily imply the operation of a multiple underlying processes.

We see these findings as making important contributions to methods and theorizing in the reasoning and category learning domains. They suggest that the traditional approach of relying

on interaction effects or task dissociations to infer underlying processes is insufficient. Instead, such claims require methods that make more realistic assumptions about mappings from latent to dependent variables, such as state-trace analysis. Our analyses suggest that many of the claims about factors that engage different underlying cognitive processes in reasoning (e.g., validity, believability or problem-type factors for dependent variables defined by different levels of working memory load, cognitive ability, time pressure, etc.) and category learning (e.g., unsupervised learning, feedback delay, mood, or richness-of-feedback factors for RB versus II tasks) are, at best exaggerated or inconsistent, and in many cases, lacking in any evidential support according to STA. Instead, as we discussed above, performance changes across conditions may often be accounted for by a single latent variable — for example, the variable could be subjective argument strength in deductive reasoning, or the weighting given to different stimulus features in category learning.

For those who wish to preserve dual-process accounts, our analysis points to a small sub-set of factors that may be more promising

Table 4
Details of the category learning database and the State-Trace analysis results: Manipulations that affect Rule-Based learning.

| Manipulation | Source | N. cat. | Stimuli | Exp | Factors that form the conditions | <i>p</i> | ^ |
|----------------------------------|--|-----------|--|-------|--|-------------|---|
| Concurrent working memory task | Zeithamova and Maddox (2006) | 2 | Gabor: frequency × orientation | 1 | 1) Control/Stroop 2) Block | .30 | |
| Stress | Elliott et al. (2011) | 2 | Gabor: frequency × orientation | 1 | 1) Low/high threat appraisal | .003 | ^ |
| | | | | | 1) Decrease/increase resistance 1) Decrease/increase cardiac output | .07 | ^ |
| Feedback processing interference | Maddox, Ashby, Ing, and Pickering (2004) | 2 | Gabor: frequency × orientation | 1A-1B | 1) Control/Long/Short 2) Block | .10 | ^ |
| | Filoteo, Lauritzen, and Maddox (2010) | 2 (conj.) | Single line: length × orientation × location | 1 | 1) 2D/3D/3D-working memory 2) Block | .11 | |
| | Zeithamova and Maddox (2007) | 2 | Gabor: frequency × orientation | 1 | 1) Control/Long/Short 2) Block | .92 | |
| Positive mood | Nadler et al. (2010) | 2 | Gabor: frequency × orientation | 1 | 1) Positive/neutral/negative mood 2) Block | .65 | |
| Self-regulation depletion | Minda and Rabi (2015) | 2 | Gabor: frequency × orientation | 1 | 1) Control/Ego depletion 2) Block | .80 | |
| Category number | Maddox, Filoteo, Hejl, and Ing (2004) | 2/4 | Single line: length × orientation | 1 | 1) Two/four categories 2) Block | .61 | |

Note. N. cat. = number of categories; Exp = Experiment; conj. = conjunction structure used for the RB task. ^ = Some/all SEs were not reported or unclear in figure, so were set to .02. Significant *p*-values ($p < .05$) are in bold font.

as evidence for multiple latent variables in the respective domains. For reasoning, assuming low SDs, this includes conditional reasoning for those with low/high working memory span (De Neys et al., 2005b), syllogistic reasoning for naïve/expert reasoners (Klauer et al., 2000), and performance on the Wason selection task under no/high time pressure conditions (Roberts & Newton, 2001). For category learning, this includes sequencing of training difficulty (Spiering & Ashby, 2008), category separation (Elliott & Ashby, 2006), social stress (Elliott et al., 2011), and performance-pressure (Markman et al., 2006). It may be more fruitful for future research in these areas to focus on these types of manipulations. However, given the large number of STA tests that we ran within each domain, it is possible that some of these “positive” results are Type 1 errors.

There are some caveats on our STA results. First, the current work was not an exhaustive meta-analysis of all papers that have examined the relevant factors in the reasoning and category learning literatures. Instead, we followed the lead of prominent dual-process theorists in each domain, focusing on the papers that they cite as key forms of evidence for their claims. It is possible that other types of factors or combinations of factors could produce stronger evidence for multiple latent dimensions, especially if the experiments are designed with STA in mind (cf. Stephens et al., 2018a). Certainly, some experiments in the database may have been under-powered against the stricter evidence criteria of STA, particularly if there were few conditions (e.g., Evans & Curtis-Holmes, 2005, had only four conditions; for discussion see e.g., Prince, Brown, & Heathcote, 2012).

Second, the databases consisted of summary statistics only, and – especially for the reasoning database – sometimes SDs or SEs were not available to capture the exact variability of the data. Our reliance on summary statistics meant that the bootstrapping was parametric; it assumed that observations were normally distributed for each dependent variable in each condition. When the full data are available, non-parametric bootstrapping

can instead be applied, and the CMR software by Dunn and Kalish (2018) also offers a version of the test for binary data at the individual-participant level. Despite these potential limitations, we feel that our results are sufficient to demonstrate that the warnings about interpreting interactions presented by authors such as Loftus (1978) and Wagenmakers, Kryptos et al. (2012) need to be taken seriously.

5.1. Implications for other types of evidence in reasoning and category learning

Claims regarding the existence of multiple processes in reasoning and category learning have not relied solely on evidence from functional dissociations across experimental manipulations (see Ashby & Valentin, 2017; Evans & Stanovich, 2013). In reasoning, it has been suggested that individual differences in working memory, intelligence, and motivation can produce different patterns of performance on different types of reasoning tasks (see Stanovich, West, & Toplak, 2014, for a review). Such dissociations have typically been interpreted as evidence for dual-process accounts. Additionally, in both reasoning (e.g., Goel, 2007; Goel & Dolan, 2004; Osherson et al., 1998) and category learning (see Ashby & Maddox, 2011; Ashby & Valentin, 2017), dual-process accounts have also been bolstered by findings that performance on different types of tasks is associated with different patterns of neural activity (as detected via neuroimaging or event-related potentials), or with different patterns of spared and impaired functioning in clinical cases. Notably, the analyses of these data have also generally been based on approaches (e.g., ANOVA, multiple regression) that assume a linear mapping between latent variables and observed responses.

Although our current focus was on the application of STA to dissociations cited as being based on experimental manipulations, STA can be applied to individual differences, and neuroimaging and neuropsychological data. Indeed, we applied STA to reasoning

Table 5
 Details of the category learning database and the State-Trace analysis results: Manipulations that affect both tasks.

| Manipulation | Source | N. cat. | Stimuli | Exp | Factors that form the conditions | Dataset details (if applicable) | <i>p</i> | ^a |
|-------------------------------|---|-----------|---|-----|--|---------------------------------|-----------------|--------------|
| Pressure | Markman et al. (2006) | 2 | Gabor: frequency × orientation | 1 | 1) Low/high pressure 2) Block | | <.001 | ^ |
| Richness of feedback | Maddox, Love, Glass, and Filoteo (2008) | 4 | Single line: length × orientation | 1 | 1) Minimal/full feedback 2) Block | | .12 | |
| Within versus across modality | Smith et al. (2014) | 2 | Pixels & tones: pixel density × tone duration | 1 | 1) Block | RB-vertical vs. II-minor | 1.00 | ^ |
| | | | Pixels & tones: pixel density × tone pitch | 2 | 1) Block | RB-vertical vs. II-major | .47 | ^ |
| | | | | | | RB-horizontal vs. II-minor | .92 | ^ |
| | | | | | | RB-horizontal vs. II-major | .18 | ^ |
| | | | | | | RB-vertical vs. II-minor | .80 | ^ |
| | | | | | | RB-vertical vs. II-major | .38 | ^ |
| | | | | | | RB-horizontal vs. II-minor | .96 | ^ |
| | | | | | | RB-horizontal vs. II-major | .37 | ^ |
| Dimensional priming | Grimm and Maddox (2013) | 2 (conj.) | Single line: length × orientation × position | 1–2 | 1) Position/Length/Orientation/Control 2) Block | | .77 | ^a |
| Visual masking | Hélie and Cousineau (2015) | 2 (conj.) | Gabor: frequency × orientation | 1 | 1) Target duration 2) Block | | .97 | |
| | | | | 2 | 1) Mask opacity 2) Block | | .50 | |

Note. N. cat. = number of categories; Exp = Experiment; conj. = conjunction structure used for the RB task. ^ = SEs were not reported, so were set to .02. *a* = approximate CMR fits were used. Significant *p*-values (*p* < .05) are in bold font.

datasets that separated individuals into groups based on tests of general intelligence or cognitive ability (Evans et al., 2010) and working memory span (De Neys, 2006b; De Neys et al., 2005b) (see Table 2; for another example of this approach, see Hayes, Stephens, Ngo, & Dunn, 2018). Similarly, Kalish, Newell, and Dunn (2017) performed STA on RB and II category learning data with groups based on working memory capacity, and Newell et al. (2011) also applied STA to previously published category learning data from clinical neuropsychological patients with Huntington's or Parkinson's disease (Filoteo, Maddox, & Davis, 2001; Filoteo, Maddox, Salmon, & Song, 2005). None of these applications of STA found compelling evidence for a two-dimensional state-trace. Likewise, the different neural signatures associated with reasoning tasks that emphasize deductive accuracy or responding based on prior knowledge (e.g., Goel & Dolan, 2004; Osherson et al., 1998) could be investigated by using neural data as the relevant dependent variables (see Staresina, Fell, Dunn, Axmacher, & Henson, 2013, for an example of an application of STA to patterns of neural activation during different recognition memory tasks).

Although comprehensive state-trace analyses have yet to be conducted, the above findings suggest caution in making strong claims about the existence of multiple processes in reasoning and category learning based on individual differences, neuropsychological patient studies, or neural activation data. It seems likely that when the linear mapping assumption is relaxed, at least some of these datasets could also be consistent with a one-dimensional state-trace.

5.2. What we can and cannot conclude from STA

It is important to recognize what can and cannot be learned from STA. STA tests whether the dependent variables are influenced by more than one underlying latent variable. Note that if the null-hypothesis of a one-dimensional state-trace cannot be rejected for a given data set, this does not *prove* the single-process account, nor conclusively rule-out the existence of multiple underlying processes. For example, a one-dimensional state-trace could be observed if two underlying latent variables happen to *not* differentially influence the two chosen dependent variables across conditions (for further discussion, see Dunn et al., 2014). Nevertheless, if one-dimensionality cannot be rejected, it suggests that there is no evidence that compels multiple processes; in such cases a single-process theoretical account may be preferred on the grounds of parsimony (Cassimatis, Bello, & Langley, 2008; Vandekerckhove, Matzke, & Wagenmakers, 2015). However, the present frequentist framework cannot be used to quantify evidence in favor of the null-hypothesis of a one-dimensional state-trace (but for a Bayesian approach, see e.g., Davis-Stober, Morey, Gretton, & Heathcote, 2016; Prince, Hawkins, Love, & Heathcote, 2012).

As with more conventional statistical approaches, another issue is that caution must be taken in dealing with aggregate data. Our analyses were necessarily based on averages across participants, and indeed our goal was to investigate claims about dual-processes based on data at this level. However, it is important to check that averages are a reasonable reflection of individual participants, if individual cognition is of primary interest. It is possible to find an aggregate-level two-dimensional state-trace because there are different subgroups of participants — in effect there is a second latent variable that governs the proportion of participants in each subgroup, which differentially effects the dependent variables. For example, Dunn et al. (2012) initially found (albeit inconsistent) evidence for more than one latent variable in RB and II category learning when they manipulated feedback delay, but this was shown to be the result of averaging across clear subgroups of learners (Stephens & Kalish, 2018). When low and high-performing participants were appropriately separated into different data points for

STA, the data were consistent with a single latent variable. We note that some of the two-dimensional state-traces we have identified in the databases may be similarly attributable to a mixture of one-dimensional state-traces at the subgroup or individual level, further reducing any evidence for multiple processes.

Conversely, in principle, an aggregate-level one-dimensional state-trace could arise from an underlying mixture of different two-dimensional state-traces. Note however that this would require a fortuitous and exact cancellation of observed values. Alternatively, an aggregate-level one-dimensional state-trace might be observed if only a few individuals violate monotonicity, especially if the violation is small (see Davis-Stober et al., 2016). In future research, individual-level analyses could explore this possibility. Regardless, it is important to note that to the extent that averaging may distort the picture that emerges from STA, the same is equally true for standard dissociation approaches.

If there is sufficient evidence of a two-dimensional state-trace, the test itself — as is the case for any statistical test — is silent on the psychological interpretation of the underlying variables. Such a result may imply that multiple distinct processes are producing the observed data. Alternatively, the finding of multiple latent variables may also be consistent with “single-process” accounts that posit multiple parameters. In reasoning about arguments like those in Table 1 for example, the distinction between single-process and dual-process accounts can be interpreted as a distinction between the number of qualitatively different ways that one assesses the strength of an argument. In other words, the question is whether there are one or more latent variables that govern the discrimination of stronger and weaker arguments. A dual-process account may assume that assessments can be based on the output of Type 1 or Type 2 processing. In contrast, a single-process account assumes only a single kind of assessment, but may still invoke additional parameters to reflect changes in performance factors such as the setting of a response threshold (e.g., see Rotello & Heit, 2009; Stephens et al., 2018a). Likewise, in category learning a two-dimensional state-trace could reflect the dimension-weighting and response-bias parameters in the Generalized Context Model (Nosofsky, 1986), which would generally be interpreted as a “single-process” account.

Therefore, when two-dimensions are revealed in STA, competing theoretical accounts may still need to be distinguished, based on further modeling and examination of *how* each model is able to account for the critical effects. Furthermore, an extension to STA — signed difference analysis — has been developed that can also test competing models with two or more core theoretical parameters (see Dunn & James, 2003; Stephens et al., 2018a). Such analyses that are based on the minimal monotonicity assumption can be used to guide and constrain models that make stronger assumptions, such as Gaussian response distributions in signal detection models (for such an approach in recognition memory, see Heathcote, Bora, & Freeman, 2010; Heathcote, Freeman, Etherington, Tonkin, & Bora, 2009). These considerations highlight the value in deriving formal quantitative models of competing theories, so that their predictions (and assumptions) can be made explicit and testable.

5.3. Using STA for new experiments in future

Moving forward, we recommend that researchers think more carefully about the assumptions that underlie interpretations of their statistical tests, and acknowledge the assumptions in their research reports so that these issues can be understood more widely. Ideally, studies could be designed for STA rather than ANOVA and other traditional tests, with demonstration of a two-dimensional state-trace as the intended benchmark, rather than a standard interaction effect.

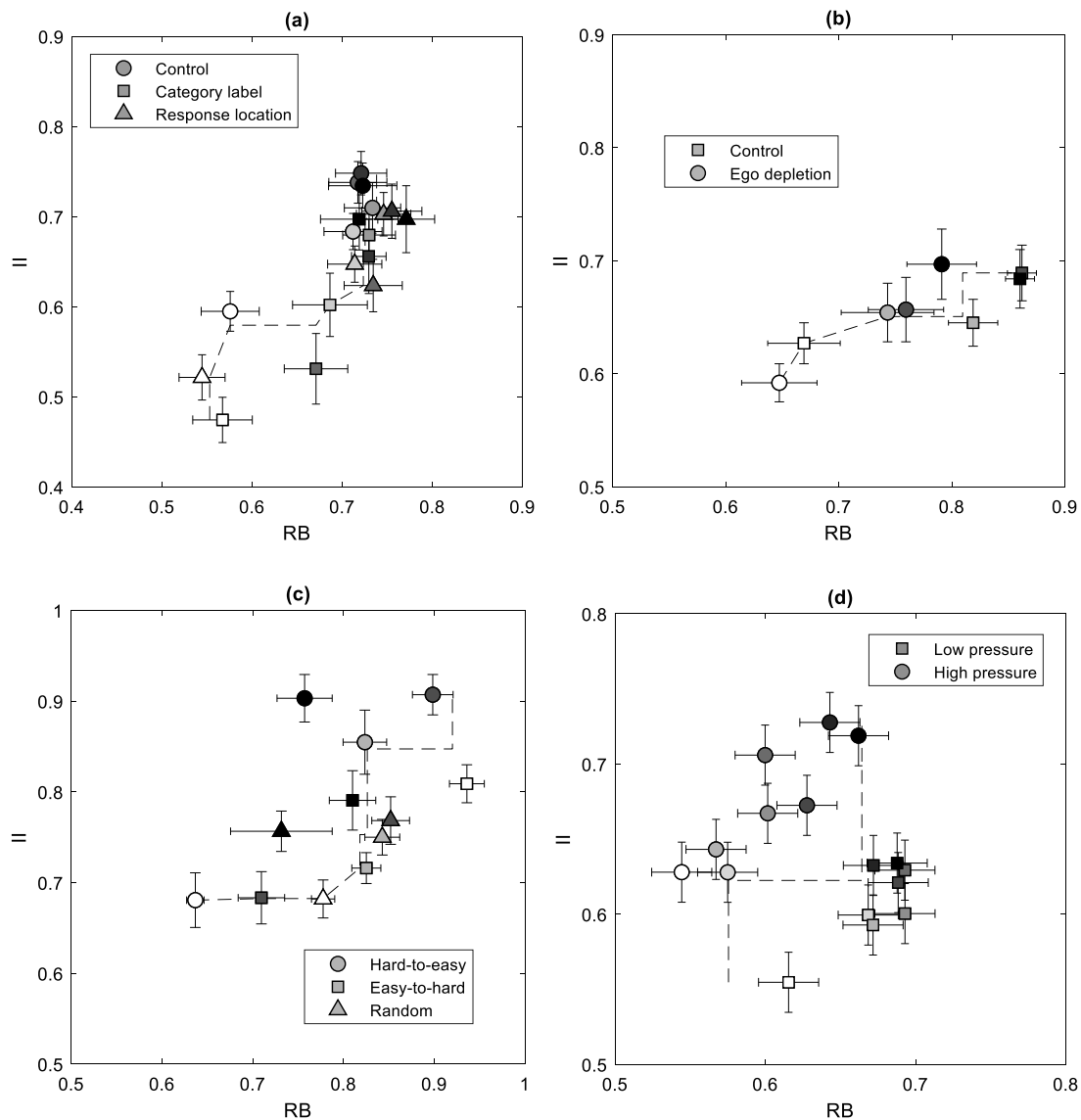


Fig. 4. State-trace plot for four sample datasets in the category learning database: (a) Maddox et al. (2010); (b) Minda and Rabi (2015); (c) Spiering and Ashby (2008; from Figures 2 & 5) (d) Markman et al. (2006). Darker shades of gray reflect increasing block number. Error bars show SEs, which are based on assuming $SE = .02$ for plot (d). Each dashed line shows the best-fitting monotonic points. The null hypothesis of a one-dimensional state-trace cannot be rejected for plots (a) and (b), but is rejected for plots (c) and (d)

In designing experiments for STA, it is important to consider that STA demands different data patterns to ANOVA in order to reject the null hypothesis, and thus different designs may be required to maximize statistical power. Power can be assessed in several ways. For example, Kalish et al. (2016) demonstrated how data simulations can be used to estimate power for an STA against any given effect size and particular experiment design. Even better, if theorists have in mind particular quantitative models that predict a two-dimensional state-trace, the models can be used to simulate predicted data for testing against STA, as an existence proof that two latent dimensions will be detected if the model is correct (for an example of this approach, see Hayes et al., 2018, footnote 4). Furthermore, Prince, Brown et al. (2012) presented useful advice for designing state-trace experiments, such as ensuring that there are enough conditions, and that the conditions overlap on at least one dependent variable so that STA can be diagnostic of dimensionality.

For future research, software for performing STA tests is freely available, such as from Dunn and Kalish (2018) and Bayesian versions by Davis-Stober et al. (2016) and Prince, Hawkins et al.

(2012). As noted above, when the full data are available, the CMR software by Dunn and Kalish uses non-parametric bootstrapping, and thus the STA test is robust in the sense that it does not rely on the ANOVA assumptions of normal residuals and homogeneity of variance. The Dunn and Kalish software (and the Bayesian versions) also includes an option to increase power by including order restrictions on the possible best-fitting monotonic curve (e.g., this might include the restriction that accuracy should never decrease across training blocks). Decisions about such order restrictions should be made a priori and be based on a compelling and uncontroversial motivation (hence we have not included them in our analyses), and ideally would be included in the pre-registration of research plans. It is important that new statistical tools and methodologies are further developed to assist researchers with applying STA logic, which is the focus of the subsequent papers in the current special issue of the *Journal of Mathematical Psychology*.

5.4. Broader implications

Although we have focused on the domains of reasoning and category learning, our results have implications for any area of psychological science in which theorists wish to draw inferences about latent psychological processes from behavioral data. This is likely to include the areas of research that we mentioned at the outset, such as declarative versus non-declarative memory and learning (e.g., McLaren et al., 2014; Schacter & Tulving, 1994; Squire, 1992), visuo-spatial versus phonological working memory (Baddeley, 2012), and distinct modes of thinking for utilitarian versus deontological moral judgments (Paxton & Greene, 2010). The current work suggests that many of the key dissociations supporting distinct processes depend heavily on the linearity assumption and will disappear under state-trace analysis. Indeed, this has been found in previous applications of STA to dissociation data in recognition memory (Dunn, 2008; Hayes, Dunn, Joubert, & Taylor, 2017), face perception (Loftus, Oberg, & Dillon, 2004), and inductive versus deductive reasoning tasks (Hayes et al., 2018; Stephens et al., 2018a). Further uptake and advancements in techniques such as STA will therefore lead to a more rigorous foundation for theoretical claims about the existence of distinct underlying psychological processes.

Acknowledgments

This research was supported by Australian Research Council Discovery Grant DP150101094 to BKH. DM was supported by a Veni grant (451-15-010) from the Netherlands Organization of Scientific Research (NWO). We thank Natali Dilevski for her assistance with building the category learning database, and John Dunn for his statistical advice and software, which was also developed by Michael Kalish and Luke Finlay.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.jmp.2018.11.003>.

References

Note: References marked with a single asterisk indicate studies included in the reasoning database. References marked with two asterisks indicate studies included in the category learning database.

- Anderson, N. H. (1961). Scales and statistics: Parametric and nonparametric. *Psychological Bulletin*, 58, 305–316. <http://dx.doi.org/10.1037/h0042576>.
- 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, 442–481. <http://dx.doi.org/10.1037/0033-295X.105.3.442>.
- Ashby, F. G., Ell, S. W., & Waldron, E. M. (2003). Procedural learning in perceptual categorization. *Memory & Cognition*, 31, 1114–1125. <http://dx.doi.org/10.3758/BF03196132>**.
- Ashby, F. G., & Maddox, W. T. (2011). Human category learning 2.0. *Annals of the New York Academy of Sciences*, 1224, 147–161. <http://dx.doi.org/10.1111/j.1749-6632.2010.05874.x>.
- Ashby, F. G., Maddox, W. T., & Bohil, C. J. (2002). Observational versus feedback training in rule-based and information-integration category learning. *Memory & Cognition*, 30, 666–677. <http://dx.doi.org/10.3758/BF03196423>**.
- Ashby, F. G., Paul, E. J., & Maddox, W. T. (2011). COVIS. In E. M. Pothos, & A. J. Wills (Eds.), *Formal approaches in categorization* (pp. 65–87). New York, NY: Cambridge University Press.
- Ashby, F. G., Queller, S., & Berretty, P. M. (1999). On the dominance of unidimensional rules in unsupervised categorization. *Perception & Psychophysics*, 61, 1178–1199. <http://dx.doi.org/10.3758/BF03207622>**.
- 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). Cambridge, MA: Elsevier. <http://dx.doi.org/10.1016/B978-0-08-101107-2.00007-5>.

- Baddeley, A. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology*, 63, 1–29. <http://dx.doi.org/10.1146/annurev-psych-120710-100422>.
- Bamber, D. (1979). State-trace analysis: a method of testing simple theories of causation. *Journal of Mathematical Psychology*, 19, 137–181. [http://dx.doi.org/10.1016/0022-2496\(79\)90016-6](http://dx.doi.org/10.1016/0022-2496(79)90016-6).
- Bogartz, R. S. (1976). On the meaning of statistical interactions. *Journal of Experimental Child Psychology*, 22, 178–183. [http://dx.doi.org/10.1016/0022-0965\(76\)90099-0](http://dx.doi.org/10.1016/0022-0965(76)90099-0).
- Bones, A. K. (2012). We knew the future all along: Scientific hypothesizing is much more accurate than other forms of precognition—A satire in one part. *Perspectives on Psychological Science*, 7, 307–309. <http://dx.doi.org/10.1177/1745691612441216>.
- Cantwell, G., Crossley, M. J., & Ashby, F. G. (2015). Multiple stages of learning in perceptual categorization: Evidence and neurocomputational theory. *Psychonomic Bulletin & Review*, 22, 1598–1613. <http://dx.doi.org/10.3758/s13423-015-0827-2>.
- Casale, M. B., Roeder, J. L., & Ashby, F. G. (2012). Analogical transfer in perceptual categorization. *Memory & Cognition*, 40, 434–449. <http://dx.doi.org/10.3758/s13421-011-0154-4>**.
- Cassimatis, N. L., Bello, P., & Langley, P. (2008). Ability, breadth, and parsimony in computational models of higher-order cognition. *Cognitive Science*, 32, 1304–1322. <http://dx.doi.org/10.1080/03640210802455175>.
- Chambers, C. D. (2013). Registered reports: A new publishing initiative at Cortex. *Cortex*, 49, 609–610. <http://dx.doi.org/10.1016/j.cortex.2012.12.016>.
- Davis-Stober, C. P., Morey, R. D., Gretton, M., & Heathcote, A. (2016). Bayes factors for state-trace analysis. *Journal of Mathematical Psychology*, 72, 116–129. <http://dx.doi.org/10.1016/j.jmp.2015.08.004>.
- De Neys, W. (2006a). Automatic-heuristic and executive-analytic processing during reasoning: Chronometric and dual-task considerations. *Quarterly Journal of Experimental Psychology*, 59, 1070–1100. <http://dx.doi.org/10.1080/02724980543000123>**.
- De Neys, W. (2006b). Dual processing in reasoning: Two systems but one reasoner. *Psychological Science*, 17, 428–433. <http://dx.doi.org/10.1111/j.1467-9280.2006.01723.x>**.
- De Neys, W. (2015). Heuristic bias and conflict detection during thinking. In B. Ross (Ed.), *Psychology of learning and motivation* (Vol. 62) (pp. 1–32). Burlington: Academic Press. <http://dx.doi.org/10.1016/bs.plm.2014.09.001>.
- De Neys, W., Schaeken, W., & d'Ydewalle, G. (2005a). Working memory and counterexample retrieval for causal conditionals. *Thinking & Reasoning*, 11, 123–150. <http://dx.doi.org/10.1080/13546780442000105>**.
- De Neys, W., Schaeken, W., & d'Ydewalle, G. (2005b). Working memory and everyday conditional reasoning: Retrieval and inhibition of stored counterexamples. *Thinking & Reasoning*, 11, 349–381. <http://dx.doi.org/10.1080/13546780442000222>**.
- Doyen, S., Klein, O., Pichon, C.-L., & Cleeremans, A. (2012). Behavioral priming: it's all in the mind, but whose mind? *PLoS One*, 7, e29081. <http://dx.doi.org/10.1371/journal.pone.0029081>.
- Dube, C., Rotello, C. M., & Heit, E. (2010). Assessing the belief bias effect with ROCs: It's a response bias effect. *Psychological Review*, 117, 831–863. <http://dx.doi.org/10.1037/a0019634>.
- Dunn, J. C. (2008). The dimensionality of the remember-know task: A state-trace analysis. *Psychological Review*, 115, 426–446. <http://dx.doi.org/10.1037/0033-295X.115.2.426>.
- Dunn, J. C., & James, R. N. (2003). Signed difference analysis: Theory and application. *Journal of Mathematical Psychology*, 47(4), 389–416. [http://dx.doi.org/10.1016/S0022-2496\(03\)00049-X](http://dx.doi.org/10.1016/S0022-2496(03)00049-X).
- Dunn, J. C., & Kalish, M. L. (2018). *State-trace analysis*. Springer.
- Dunn, J. C., Kalish, M. L., & Newell, B. R. (2014). State-trace analysis can be an appropriate tool for assessing the number of cognitive systems: A reply to Ashby. *Psychonomic Bulletin & Review*, 21, 947–954. <http://dx.doi.org/10.3758/s13423-014-0637-y>.
- Dunn, J. C., & Kirsner, K. (1988). Discovering functionally independent mental processes: The principle of reversed association. *Psychological Review*, 95, 91–101. <http://dx.doi.org/10.1037/0033-295X.95.1.91>.
- Dunn, J. C., Newell, B. R., & Kalish, M. L. (2012). The effect of feedback delay and feedback type on perceptual category learning: The limits of multiple systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38, 840–859. <http://dx.doi.org/10.1037/a0027867>.
- Ell, S. W., & Ashby, F. G. (2006). The effects of category overlap on information-integration and rule-based category learning. *Perception & Psychophysics*, 68, 1013–1026. <http://dx.doi.org/10.3758/BF03193362>**.
- Ell, S. W., Ashby, F. G., & Hutchinson, S. (2012). Unsupervised category learning with integral-dimension stimuli. *The Quarterly Journal of Experimental Psychology*, 65, 1537–1562. <http://dx.doi.org/10.1080/17470218.2012.658821>**.

- Ell, S. W., Cosley, B., & McCoy, S. K. (2011). When bad stress goes good: Increased threat reactivity predicts improved category learning performance. *Psychonomic Bulletin & Review*, 18, 96–102. <http://dx.doi.org/10.3758/s13423-010-0018-0>**.
- Evans, J. St. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, 59, 255–278. <http://dx.doi.org/10.1146/annurev.psych.59.103006.093629>.
- Evans, J. St. B. T., & Curtis-Holmes, J. (2005). Rapid responding increases belief bias: Evidence for the dual-process theory of reasoning. *Thinking & Reasoning*, 11, 382–389. <http://dx.doi.org/10.1080/13546780542000005>*.
- Evans, J. St. B. T., Handley, S. J., & Harper, C. N. J. (2001). Necessity, possibility and belief: A study of syllogistic reasoning. *Quarterly Journal of Experimental Psychology*, 54, 935–958. <http://dx.doi.org/10.1080/713755983>*.
- Evans, J. St. B. T., Handley, S. J., Neilens, H., & Over, D. (2010). The influence of cognitive ability and instructional set on causal conditional inference. *The Quarterly Journal of Experimental Psychology*, 63, 892–909. <http://dx.doi.org/10.1080/17470210903111821>*.
- Evans, J. St. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, 8, 223–241. <http://dx.doi.org/10.1177/1745691612460685>.
- Filoteo, J. V., Lauritzen, S., & Maddox, W. T. (2010). Removing the frontal lobes: The effects of engaging executive functions on perceptual category learning. *Psychological Science*, 21, 415–423. <http://dx.doi.org/10.1177/0956797610362646>**.
- Filoteo, J. V., Maddox, W. T., & Davis, J. D. (2001). A possible role of the striatum in linear and nonlinear categorization rule learning: Evidence from patients with Huntington's disease. *Behavioral Neuroscience*, 115, 786–798. <http://dx.doi.org/10.1037/0735-7044.115.4.786>.
- Filoteo, J. V., Maddox, W. T., Salmon, D. P., & Song, D. D. (2005). Information-integration category learning in patients with striatal dysfunction. *Neuropsychology*, 19, 212–222. <http://dx.doi.org/10.1037/0894-4105.19.2.212>.
- Goel, V. (2007). Anatomy of deductive reasoning. *Trends in Cognitive Sciences*, 11, 435–441. <http://dx.doi.org/10.1016/j.tics.2007.09.003>.
- Goel, V., & Dolan, R. J. (2004). Differential involvement of left prefrontal cortex in inductive and deductive reasoning. *Cognition*, 93, B109–B121. <http://dx.doi.org/10.1016/j.cognition.2004.03.001>.
- Grimm, L. R., & Maddox, W. T. (2013). Differential impact of relevant and irrelevant dimension primes on rule-based and information-integration category learning. *Acta Psychologica*, 144, 530–537. <http://dx.doi.org/10.1016/j.actpsy.2013.09.005>**.
- Handley, S. J., & Trippas, D. (2015). Dual processes and the interplay between knowledge and structure: A new parallel processing model. *Psychology of Learning and Motivation*, 62, 33–58. <http://dx.doi.org/10.1016/bs.plm.2014.09.002>.
- Hayes, B. K., Dunn, J. C., Joubert, A., & Taylor, R. (2017). Comparing single- and dual-process models of memory development. *Developmental Science*, 20, e12469. <http://dx.doi.org/10.1111/desc.12469>.
- Hayes, B. K., Stephens, R. G., Ngo, J., & Dunn, J. C. (2018). The dimensionality of reasoning: inductive and deductive inference can be explained by a single process. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, published online ahead of print. <http://dx.doi.org/10.1037/xlm0000527>.
- Heathcote, A., Bora, B., & Freeman, E. (2010). Recollection and confidence in two-alternative forced choice episodic recognition. *Journal of Memory and Language*, 62, 183–203. <http://dx.doi.org/10.1016/j.jml.2009.11.003>.
- Heathcote, A., Freeman, E., Etherington, J., Tonkin, J., & Bora, B. (2009). A dissociation between similarity effects in episodic face recognition. *Psychonomic Bulletin & Review*, 16, 824–831. <http://dx.doi.org/10.3758/PBR.16.5.824>.
- Hélie, S., & Ashby, F. G. (2012). Learning and transfer of category knowledge in an indirect categorization task. *Psychological Research*, 76, 292–303. <http://dx.doi.org/10.1007/s00426-011-0348-1>**.
- Hélie, S., & Cousineau, D. (2015). Differential effect of visual masking in perceptual categorization. *Journal of Experimental Psychology: Human Perception and Performance*, 41, 816–825. <http://dx.doi.org/10.1037/xhp0000063>**.
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2, 696–701. <http://dx.doi.org/10.1371/journal.pmed.0020124>.
- John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23, 524–532. <http://dx.doi.org/10.1177/0956797611430953>.
- Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49–81). Cambridge: Cambridge University Press.
- Kalish, M. L., Dunn, J. C., Burdakov, O. P., & Sysoev, O. (2016). A statistical test of the equality of latent orders. *Journal of Mathematical Psychology*, 70, 1–11. <http://dx.doi.org/10.1016/j.jmp.2015.10.004>.
- Kalish, M. L., Newell, B. R., & Dunn, J. C. (2017). More is generally better: Higher working memory capacity does not impair perceptual category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43, 503–514. <http://dx.doi.org/10.1037/xlm0000323>.
- Keren, G., & Schul, Y. (2009). Two is not always better than one: A critical evaluation of two-system theories. *Perspectives on Psychological Science*, 4, 533–550. <http://dx.doi.org/10.1111/j.1745-6924.2009.01164.x>.
- Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, 2, 196–217. http://dx.doi.org/10.1207/s15327957pspr0203_4.
- Klauer, K. C., Musch, J., & Naumer, B. (2000). On belief bias in syllogistic reasoning. *Psychological Review*, 107, 852–884. <http://dx.doi.org/10.1037/0033-295X.107.4.852>*.
- Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., ..., & Nosek, B. A. (2014). Investigating variation in replicability: a many labs replication project. *Social Psychology*, 45, 142–152. <http://dx.doi.org/10.1027/1864-9335/a000178>.
- Koole, S. L., & Lakens, D. (2012). Rewarding replications: A sure and simple way to improve psychological science. *Perspectives on Psychological Science*, 7, 608–614. <http://dx.doi.org/10.1177/1745691612462586>.
- Kruglanski, A. W., & Gigerenzer, G. (2011). Intuitive and deliberate judgments are based on common principles. *Psychological Review*, 118, 97–109. <http://dx.doi.org/10.1037/a0020762>.
- Lassiter, D., & Goodman, N. D. (2015). How many kinds of reasoning? Inference, probability, and natural language semantics. *Cognition*, 136, 123–134. <http://dx.doi.org/10.1016/j.cognition.2014.10.016>.
- Loftus, G. R. (1978). On interpretation of interactions. *Memory & Cognition*, 6, 312–319. <http://dx.doi.org/10.3758/BF03197461>.
- Loftus, G. R., Oberg, M. A., & Dillon, A. M. (2004). Linear theory, dimensional theory, and the face-inversion effect. *Psychological Review*, 111, 835–863. <http://dx.doi.org/10.1037/0033-295X.111.4.835>.
- Maddox, W. T., Ashby, F. G., & Bohil, C. J. (2003). Delayed feedback effects on rule-based and information-integration category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29, 650–662. <http://dx.doi.org/10.1037/0278-7393.29.4.650>**.
- Maddox, W. T., Ashby, F. G., Ing, A. D., & Pickering, A. D. (2004). Disrupting feedback processing interferes with rule-based but not information-integration category learning. *Memory & Cognition*, 32, 582–591. <http://dx.doi.org/10.3758/BF03195849>**.
- Maddox, W. T., Bohil, C. J., & Ing, A. D. (2004). Evidence for a procedural-learning-based system in perceptual category learning. *Psychonomic Bulletin & Review*, 11, 945–952. <http://dx.doi.org/10.3758/BF03196726>**.
- Maddox, W. T., Filoteo, J. V., Hejl, K. D., & Ing, A. D. (2004). Category number impacts rule-based but not information-integration category learning: Further evidence for dissociable category-learning systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30, 227–245. <http://dx.doi.org/10.1037/0278-7393.30.1.227>**.
- Maddox, W. T., Filoteo, J. V., & Lauritzen, J. S. (2007). Within-category discontinuity interacts with verbal rule complexity in perceptual category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33, 197–218. <http://dx.doi.org/10.1037/0278-7393.33.1.197>**.
- Maddox, W. T., Filoteo, J. V., Lauritzen, J. S., Connolly, E., & Hejl, K. D. (2005). Discontinuous categories affect information-integration but not rule-based category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31, 654–669. <http://dx.doi.org/10.1037/0278-7393.31.4.654>**.
- Maddox, W. T., Glass, B. D., O'Brien, J. B., Filoteo, J. V., & Ashby, F. G. (2010). Category label and response location shifts in category learning. *Psychological Research*, 74, 219–236. <http://dx.doi.org/10.1007/s00426-009-0245-z>**.
- Maddox, W. T., & Ing, A. D. (2005). Delayed feedback disrupts the procedural-learning system but not the hypothesis-testing system in perceptual category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31, 100–107. <http://dx.doi.org/10.1037/0278-7393.31.1.100>**.
- Maddox, W. T., Lauritzen, J. S., & Ing, A. D. (2007). Cognitive complexity effects in perceptual classification are dissociable. *Memory & Cognition*, 35, 885–894. <http://dx.doi.org/10.3758/BF03193463>**.
- Maddox, W. T., Love, B. C., Glass, B. D., & Filoteo, J. V. (2008). When more is less: Feedback effects in perceptual category learning. *Cognition*, 108, 578–589. <http://dx.doi.org/10.1016/j.cognition.2008.03.010>**.
- Markman, A. B., Maddox, W. T., & Worthy, D. A. (2006). Choking and excelling under pressure. *Psychological Science*, 17, 944–948. <http://dx.doi.org/10.1111/j.1467-9280.2006.01809.x>**.
- Matzke, D., Nieuwenhuis, S., van Rijn, H., Slagter, H. A., van der Molen, M. W., & Wagenmakers, E.-J. (2015). The effect of horizontal eye movements on free recall: A preregistered adversarial collaboration. *Journal of Experimental Psychology: General*, 144, e1–e15. <http://dx.doi.org/10.1037/xge0000038>.
- McLaren, I. P. L., Forrest, C. L. D., McLaren, R. P., Jones, F. W., Aitken, M. R. F., & Mackintosh, N. J. (2014). Associations and propositions: The case for a dual-process account of learning in humans. *Neurobiology of Learning and Memory*, 108, 185–195. <http://dx.doi.org/10.1016/j.nlm.2013.09.014>.
- Melnikoff, D. E., & Bargh, J. A. (2018). The mythical number two. *Trends in Cognitive Sciences*, 22, 280–293. <http://dx.doi.org/10.1016/j.tics.2018.02.001>.

- Minda, J. P., & Rabi, R. (2015). Ego depletion interferes with rule-defined category learning but not non-rule-defined category learning. *Frontiers in Psychology*, 6(35), 1–9. <http://dx.doi.org/10.3389/fpsyg.2015.00035>**.
- Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., et al. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, 1(0021), <http://dx.doi.org/10.1038/s41562-016-0021>.
- Nadler, R. T., Rabi, R., & Minda, J. P. (2010). Better mood and better performance: learning rule-described categories is enhanced by positive mood. *Psychological Science*, 21, 1770–1776. <http://dx.doi.org/10.1177/0956797610387441>**.
- Newell, B. R., & Dunn, J. C. (2008). Dimensions in data: testing psychological models using state-trace analysis. *Trends in Cognitive Sciences*, 12, 285–290. <http://dx.doi.org/10.1016/j.tics.2008.04.009>.
- Newell, B. R., Dunn, J. C., & Kalish, M. (2010). The dimensionality of perceptual category learning: A state-trace analysis. *Memory & Cognition*, 38, 563–581. <http://dx.doi.org/10.3758/MC.38.5.563>.
- Newell, B. R., Dunn, J. C., & Kalish, M. (2011). Systems of category learning: Fact or fantasy? In B. H. Ross (Ed.), *Psychology of learning and motivation* (Vol. 54) (pp. 167–215). San Diego, CA: Academic Press.
- Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology*, 45, 137–141. <http://dx.doi.org/10.1027/1864-9335/a000192>.
- Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, 115, 39–57. <http://dx.doi.org/10.1037/0096-3445.115.1.39>.
- Open Science Collaboration (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, 7, 657–660. <http://dx.doi.org/10.1177/1745691612462588>.
- Osherson, D., Perani, D., Cappa, S., Schnur, T., Grassi, F., & Fazio, F. (1998). Distinct brain loci in deductive versus probabilistic reasoning. *Neuropsychologia*, 36, 369–376. [http://dx.doi.org/10.1016/S0028-3932\(97\)00099-7](http://dx.doi.org/10.1016/S0028-3932(97)00099-7).
- Osman, M. (2004). An evaluation of dual-process theories of reasoning. *Psychonomic Bulletin & Review*, 11, 988–1010. <http://dx.doi.org/10.3758/BF03196730>.
- Osman, M. (2013). A case study: Dual-process theories of higher cognition—Commentary on Evans and Stanovich (2013). *Perspectives on Psychological Science*, 8, 248–252. <http://dx.doi.org/10.1177/1745691613483475>.
- Pashler, H., & Wagenmakers, E.-J. (2012). Editors introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, 7, 528–530. <http://dx.doi.org/10.1177/1745691612465253>.
- Paxton, J. M., & Greene, J. D. (2010). Moral reasoning: Hints and allegations. *Topics in Cognitive Science*, 2, 511–527. <http://dx.doi.org/10.1111/j.1756-8765.2010.01096.x>.
- Prince, M., Brown, S., & Heathcote, A. (2012). The design and analysis of state-trace experiments. *Psychological Methods*, 17, 78–99. <http://dx.doi.org/10.1037/a0025809>.
- Prince, M., Hawkins, G., Love, J., & Heathcote, A. (2012). An R package for state-trace analysis. *Behavior Research Methods*, 44, 644–655. <http://dx.doi.org/10.3758/s13428-012-0232-y>.
- Rips, L. J. (2001). Two kinds of reasoning. *Psychological Science*, 12, 129–134. <http://dx.doi.org/10.1111/1467-9280.00322>.
- Roberts, M. J., & Newton, E. J. (2001). Inspection times, the change task, and the rapid-response selection task. *Quarterly Journal of Experimental Psychology*, 54, 1031–1048. <http://dx.doi.org/10.1080/713756016>**.
- Rotello, C. M., & Heit, E. (2009). Modeling the effects of argument length and validity on inductive and deductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35, 1317–1330. <http://dx.doi.org/10.1037/a0016648>.
- Rotello, C. M., Heit, E., & Dubé, C. (2015). When more data steer us wrong: replications with the wrong dependent measure perpetuate erroneous conclusions. *Psychonomic Bulletin & Review*, 22, 944–954. <http://dx.doi.org/10.3758/s13423-014-0759-2>.
- Schacter, D. L., & Tulving, E. (Eds.), (1994). *Memory systems*. Cambridge, MA: MIT Press.
- Shanks, D. R., Newell, B. R., Lee, E. H., Balakrishnan, D., Ekelund, L., Cenac, Z., et al. (2013). Priming intelligent behavior: An elusive phenomenon. *PLoS One*, 8, e56515. <http://dx.doi.org/10.1371/journal.pone.0056515>.
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22, 1359–1366. <http://dx.doi.org/10.1177/0956797611417632>.
- Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, 119, 3–22.
- Sloman, S. A. (2014). Two systems of reasoning, An update. In J. W. Sherman, B. Gawronski, & Y. Trope (Eds.), *Dual process theories of the social mind*. New York, NY: Guilford Press.
- Smith, J. D., Johnston, J. J. R., Musgrave, R. D., Zakrzewski, A. C., Boomer, J., Church, B. A., et al. (2014). Cross-modal information integration in category learning. *Attention, Perception, & Psychophysics*, 76, 1473–1484. <http://dx.doi.org/10.3758/s13414-014-0659-6>**.
- Spiering, B. J., & Ashby, F. G. (2008). Initial training with difficult items facilitates information integration, but not rule-based category learning. *Psychological Science*, 19, 1169–1177. <http://dx.doi.org/10.1111/j.1467-9280.2008.02219.x>**.
- Squire, L. R. (1992). Declarative and nondeclarative memory: Multiple brain systems supporting learning and memory. *Journal of Cognitive Neuroscience*, 4, 232–243. <http://dx.doi.org/10.1162/jocn.1992.4.3.232>.
- Stanovich, K. E., West, R. F., & Toplak, M. E. (2014). Rationality, intelligence, and the defining features of Type 1 and Type 2 processing. In J. W. Sherman, B. Gawronski, & Y. Trope (Eds.), *Dual-process theories of the social mind* (pp. 80–91). New York, NY: Guilford Press.
- Stanton, R. D., & Nosofsky, R. M. (2007). Feedback interference and dissociations of classification: Evidence against the multiple-learning systems hypothesis. *Memory & Cognition*, 35, 1747–1758. <http://dx.doi.org/10.3758/BF03193507>.
- Staresina, B. P., Fell, J., Dunn, J. C., Axmacher, N., & Henson, R. N. (2013). Using state-trace analysis to dissociate the functions of the human hippocampus and perirhinal cortex in recognition memory. *Proceedings of the National Academy of Sciences*, 110, 3119–3124. <http://dx.doi.org/10.1073/pnas.1215710110>.
- Stephens, R. G., Dunn, J. C., & Hayes, B. K. (2018a). Are there two processes in reasoning? The dimensionality of inductive and deductive inferences. *Psychological Review*, 125, 218–244. <http://dx.doi.org/10.1037/rev0000088>.
- Stephens, R. G., Dunn, J. C., & Hayes, B. K. (2018b). Belief bias is response bias: Evidence from a two-step signal detection model. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, Advance online publication, <http://dx.doi.org/10.1037/xlm0000587>.
- Stephens, R. G., & Kalish, M. L. (2018). The effect of feedback delay on perceptual category learning and item memory: Further limits of multiple systems. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, <http://dx.doi.org/10.1037/xlm0000528>, published online ahead of print.
- Tanaka, J. W., & Gordon, I. (2011). Features, configuration, and holistic face processing. In A. J. Calder, G. Rhodes, M. H. Johnson, & J. V. Haxby (Eds.), *The Oxford handbook of face perception* (pp. 177–194). New York, NY: Oxford University Press, <http://dx.doi.org/10.1093/oxfordhb/9780199559053.013.0010>.
- Tharp, I. J., & Pickering, A. D. (2009). A note on DeCaro, Thomas, and Beilock (2008): Further data demonstrate complexities in the assessment of information-integration category learning. *Cognition*, 111, 410–414. <http://dx.doi.org/10.1016/j.cognition.2008.10.003>.
- Vandekerckhove, J., Matzke, D., & Wagenmakers, E.-J. (2015). Model comparison and the principle of parsimony. In J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels (Eds.), *Oxford handbook of computational and mathematical psychology* (pp. 300–319). Oxford, UK: Oxford University Press.
- Wagenmakers, E.-J., Krypotos, A.-M., Criss, A. H., & Iverson, G. (2012). On the interpretation of removable interactions: a survey of the field 33 years after Loftus. *Memory & Cognition*, 40, 145–160. <http://dx.doi.org/10.3758/s13421-011-0158-0>.
- Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, 7, 632–638. <http://dx.doi.org/10.1177/1745691612463078>.
- Zeithamova, D., & Maddox, W. T. (2006). Dual-task interference in perceptual category learning. *Memory & Cognition*, 34, 387–398. <http://dx.doi.org/10.3758/BF03193416>**.
- Zeithamova, D., & Maddox, W. T. (2007). The role of visuospatial and verbal working memory in perceptual category learning. *Memory & Cognition*, 35, 1380–1398. <http://dx.doi.org/10.3758/BF03193609>**.