Metacognitive Control in Single- vs. Dual-Process Theory[1]

Aliya R. Dewey

Department of Philosophy, University of Arizona

Recent work in cognitive modelling has found that most of the data that has been cited as evidence for the dual-process theory (DPT) of reasoning is best explained by non-linear, "monotonic" one-process models (Stephens et al., 2018, 2019). In this paper, I consider an important caveat of this research: it uses models that are committed to unrealistic assumptions about how effectively task conditions can isolate Type-1 and Type-2 reasoning. To avoid this caveat, I develop a coordinated theoretical, experimental, and modelling strategy to better test DPT. First, I propose that Type-1 and Type-2 reasoning are defined as reasoning that precedes and follows metacognitive control, respectively. Second, I argue that reasoning that precedes and follows metacognitive control can be effectively isolated using debiasing paradigms that manipulate metacognitive heuristics (e.g., processing fluency) to prevent or trigger metacognitive control, respectively. Third, I argue that monotonic modelling can allow us to decisively test DPT only when we use them to analyse data from this particular kind of debiasing paradigm.

Keywords: dual-process theory, single-process theory, meta-reasoning, metacognitive control, metacognitive heuristics, debiasing

The psychology of reasoning has been divided along theoretical lines for decades. According to dual-process theory (DPT), there are two *qualitatively* different types of reasoning (e.g., Sloman, 1996; Smith & DeCoster, 1999; Kahneman, 2011; Evans & Stanovich, 2013; Pennycook et al., 2015; De Neys & Pennycook, 2019; Bago & De Neys, 2020). They have been classified as intuition and deliberation, fast and slow reasoning, automatic and controlled reasoning, etc. but are now often classified in neutral terms as Type-1 (T1) reasoning and Type-2 (T2) reasoning (Evans & Stanovich, 2013). According to single-process theory (SPT), though, there is only a single type of reasoning, which continuously varies from intuitive to deliberative, fast to slow, automatic to controlled, etc. Thus, the difference between whatever gets classified as T1 and T2 reasoning by DPT is only *quantitative* (Osman, 2004; Keren & Schul, 2009; Kruglanski & Gigerenzer, 2011).

Dual-process theorists from the reasoning literature have reported a host of data that are better explained by linear two-process models than by linear one-process models (Evans, 2008, 2010; Stanovich, 2011; Evans & Stanovich, 2013). After tacitly assuming that linear models are better than non-linear models (by using linear statistics), these dual-process theorists have concluded that these results confirm DPT. But single- and dual-process theorists from the cognitive modelling literature both agree that *monotonic models*, a certain kind of non-linear model, provide a more

[1] Penultimate draft of article at *Thinking & Reasoning*. Please cite the published article, not this draft.

realistic explanation of reasoning (Rips, 2001; Rotello & Heit, 2009; Singmann & Klauer, 2011). And recent advances in monotonic analysis have allowed single-process theorists to show that the data cited as evidence for DPT are better explained by monotonic one-process models than by monotonic two-process models (Stephens et al., 2018, 2019).

Do these modelling results settle the debate between SPT and DPT in favour of SPT, and confirm that any differences in reasoning are merely quantitative? No, unfortunately, because there are a few caveats with interpreting these models. The basic problem is that monotonic analysis can tell us whether two sets of responses were created by the same type of process or two different types of processes, but it can't tell us whether those processes have the defining properties of the single or dual types of reasoning posited by SPT and DPT. For example, Stephens et al. (2018) used their monotonic analysis to show that inductive and deductive judgments were generated by the same type of processing. But this is irrelevant to the debate about SPT and DPT since the distinction between intuitive and deductive reasoning (regardless of whether it is qualitative or quantitative) is different from the distinction between T1 and T2 reasoning (i.e., intuition and deliberation, fast and slow reasoning, automatic and controlled reasoning, etc.).[2]

In this paper, my goal is to develop an integrated theoretical, experimental, and modelling strategy to ensure that any type-difference that monotonic analysis finds in the data can be best interpreted as the difference between T1 and T2 reasoning (vs. some other type-difference). In §1, I'll explain why monotonic one-process models provide better explanations of the data than linear two-process models. In §2, I'll review the recent advancements in monotonic analysis. In §3, I'll develop a novel formulation of DPT that avoids the problems with popular formulations of DPT: T1 and T2 reasoning are each defined as preceding and following metacognitive control, respectively. In §4, I'll draw on evidence about metacognitive control to consider the theoretical implications of DPT. In §5, I'll translate those theoretical implications into experimental predictions for DPT and explain how to use monotonic analysis to evaluate those predictions and hence, decisively test DPT.

§1. State-Trace Analysis

A peculiar, confusing feature of the debate between SPT and DPT is that both sides have insisted that the same data is best explained by and hence, evidence for their respective theories (Osman, 2004; Keren & Schul, 2009; Kahneman, 2011; Kruglanski & Gigerenzer, 2011; Keren, 2013; Evans & Stanovich, 2013; Pennycook et al., 2018). In this section, I'll argue that this confusion is easily resolved once we formalize the modelling assumptions that are implicit in these arguments: the available data are evidence for two-process models under the assumption that linear models are appropriate, but they are evidence for one-process models under the assumption that non-linear, "monotonic" (to be defined below) models are appropriate. I'll also argue that the latter assumption is correct: non-linear, monotonic models are appropriate for explaining reasoning data.

To focus on the differences between linear and monotonic models of the same evidence, I'll avoid reviewing all the data that is normally cited as evidence for SPT and/or DPT. Instead, I'll take the

[2] After all, inductive reasoning can be slow, controlled, intervening, etc. (e.g., the evaluation of difficult scientific arguments) and deductive reasoning can be fast, automatic, default, etc. (e.g., the evaluation of simple logical rules).

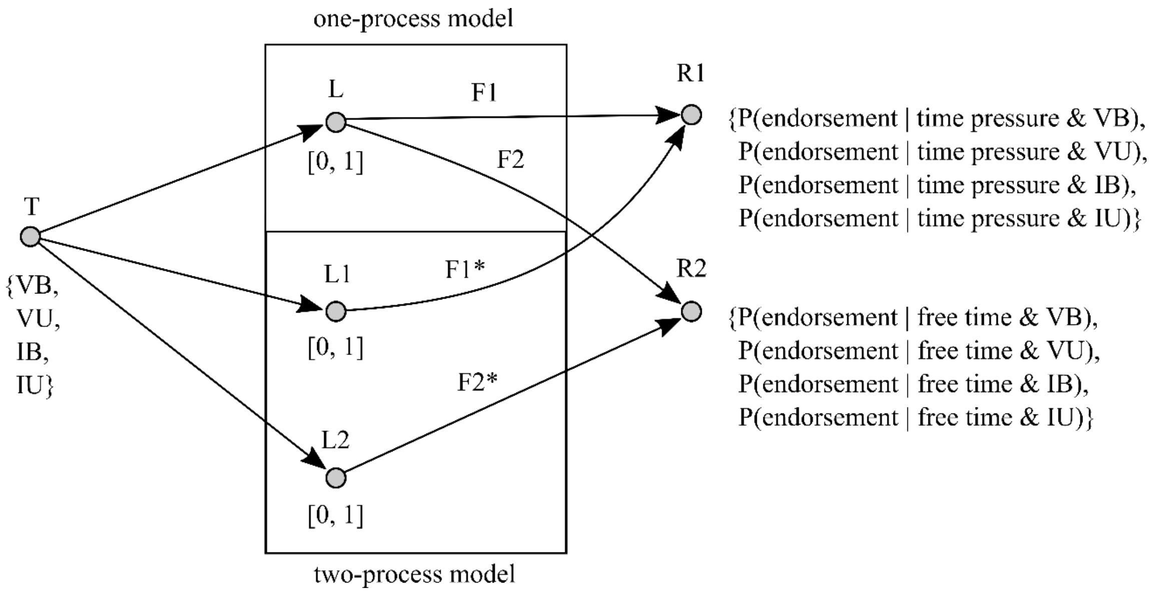
one-process model
L
F1
F2
[0, 1]
T
{VB,
VU,
IB,
IU}
L1
F1*
[0, 1]
F2*
L2
[0, 1]
two-process model
R1
{P(endorsement | time pressure & VB),
P(endorsement | time pressure & VU),
P(endorsement | time pressure & IB),
P(endorsement | time pressure & IU)}
R2
{P(endorsement | free time & VB),
P(endorsement | free time & VU),
P(endorsement | free time & IB),
P(endorsement | free time & IU)}

Figure 1. A diagram depicting generic one- and two-process models equivalent to those used by Stephens et al. (2019). T is the task variable, which can take on four states. R_1 and R_2 are the response variables, which can each take on four states (the endorsement rates for all four types of syllogisms under either time pressure or free time). L is the latent variable for the one-process model, and it can take on a range of values from 0 to 1. L explains both R_1 and R_2 via the functions F_1 and F_2 (which are linear in linear models and monotonic in monotonic models). L_1 and L_2 are the latent variables for the two-process model, and each can take on a range of values from 0 to 1. L_1 explains R_1 via F_1* and L_2 explains R_2 via F_2*. Like F_1 and F_2, F_1* and F_2* are linear in linear models and monotonic in monotonic models.

belief-bias study by Evans & Curtis-Holmes (2005) as a simple yet representative case study and compare the linear and monotonic models of the data that they report at length (Fig. 1). In that belief-bias study, Evans & Curtis-Holmes instructed subjects to judge whether "the conclusions necessarily followed from the premises" in syllogisms that varied in their validity (task-relevant) and the believability of their conclusions (task-irrelevant). They found that time pressure (a) increased mean rate of endorsement for syllogisms with believable conclusions and (b) decreased mean rate of endorsement for valid syllogisms (Fig. 2A).

They skipped the modelling step and inferred that (a) there are two types of reasoning and (b) time pressure must be promoting T1 reasoning (increasing the efficacy of its preference for believable conclusions) and inhibiting T2 reasoning (decreasing the efficacy of its preference for validity). By skipping the modelling step, Evans & Curtis-Holmes (2005) elided the critical issue of whether the relations among the variables in their explanation are linear, monotonic, or something else. To reveal the importance of this distinction, let's revisit this modelling step and identify the constraints on the theoretical inferences that we can draw from the data that they've reported.

We'll consider four models here: (a) one-process monotonic, (b) two-process monotonic, (c) one-process linear, and (d) two-process linear. Despite their differences, all four models share the same basic structure. First, they use the same task variable, which defines four kinds of syllogisms in the task: valid-believable (VB), valid-unbelievable (VU), invalid-believable (IB), and invalid-unbelievable (IU) (Fig. 1).[3] Second, all four models share the same two response variables, which describe the behaviours that both models are supposed to predict: the mean rate of endorsement

[3] Alternatively, we could represent this as the Cartesian product of two task variables, validity and believability.

with and without time pressure (Stephens et al., 2019). These two variables can take on four values each: the mean rates of endorsement for all four kinds of syllogisms with and without time pressure (Fig. 1). These are the eight data points that each model must fit. Third, all four models introduce latent variables, which are hypothetical, unobservable variables that map from the task variables onto the response variables.

The goal of cognitive modelling is to sufficiently explain the data using the least number of latent variables that are best interpreted as real cognitive entities. Now, single-process theorists claim that there is only one type of reasoning, which integrates considerations of syllogism validity and conclusion believability into a single measure of argument strength. Thus, a one-process model may use a single latent variable to represent the single measure of argument strength computed by the single type of reasoning (Fig. 1). Likewise, dual-process theorists claim that there are two types of reasoning, which use different rules to integrate considerations of validity and believability into two measures of argument strength. Thus, a two-process model may use two latent variables to represent the two measures of argument strength computed by each type of reasoning (Fig. 1). Then each latent variable is mapped onto only one response variable: here, the latent variables for T1 and T2 reasoning are mapped onto the response variables for the mean rate of endorsement under time pressure and free time conditions, respectively (Fig. 1).

However, constraints must be added to the latent variables so that they fulfill their cognitive interpretations. Rips (2001) argues that there is only one plausible constraint: if the latent variables are to represent subjective argument strength, the response function must not map an increase on the latent variables onto a decrease in the response variables (mean endorsement rates). After all, an increase in a syllogism's subjective argument strength couldn't cause a decrease in its mean rate of endorsement. Or the latent variables could represent subjective argument weakness (instead of strength), such that the response function must not map an increase on the latent variables onto an increase in the response variables. In general, then, the response function that maps the latent variables onto the response variables must be *monotonic*: the response variables must be non-decreasing or non-increasing over every value of the latent variable.

Note that we cannot justifiably assume that the response function is linear (a kind of monotonicity). After all, we have no reason at all to believe that any increase in subjective argument strength will lead to a fixed increase in mean endorsement rates regardless of the level of subjective argument strength. In fact, we have reason to reject this. Changes in subjective argument strength at the highest and lowest levels of subjective argument strength will create much smaller differences in mean endorsement rates than changes at intermediate levels of subjective argument strength. After all, subjects are only prepared to change their decisions about endorsing syllogisms in the middle of their subjective scales for argument strength. Hence, response functions are closer to *sigmoidal* (another kind of monotonicity) than linear, but assuming that response functions are sigmoidal may still be too strong, so the weaker monotonicity assumption is still generally preferred in the cognitive modelling literature (Loftus, 1978; Wagenmakers et al., 2012; Stephens et al., 2019).

So, we have good reason to believe that monotonic models are appropriate for explaining reasoning data and linear models aren't. Next, let's compare one- and two-process models, starting with monotonic models before proceeding to linear models. Dunn & Kirsner (1988) note that one-process monotonic models make a unique prediction that two-process monotonic models don't

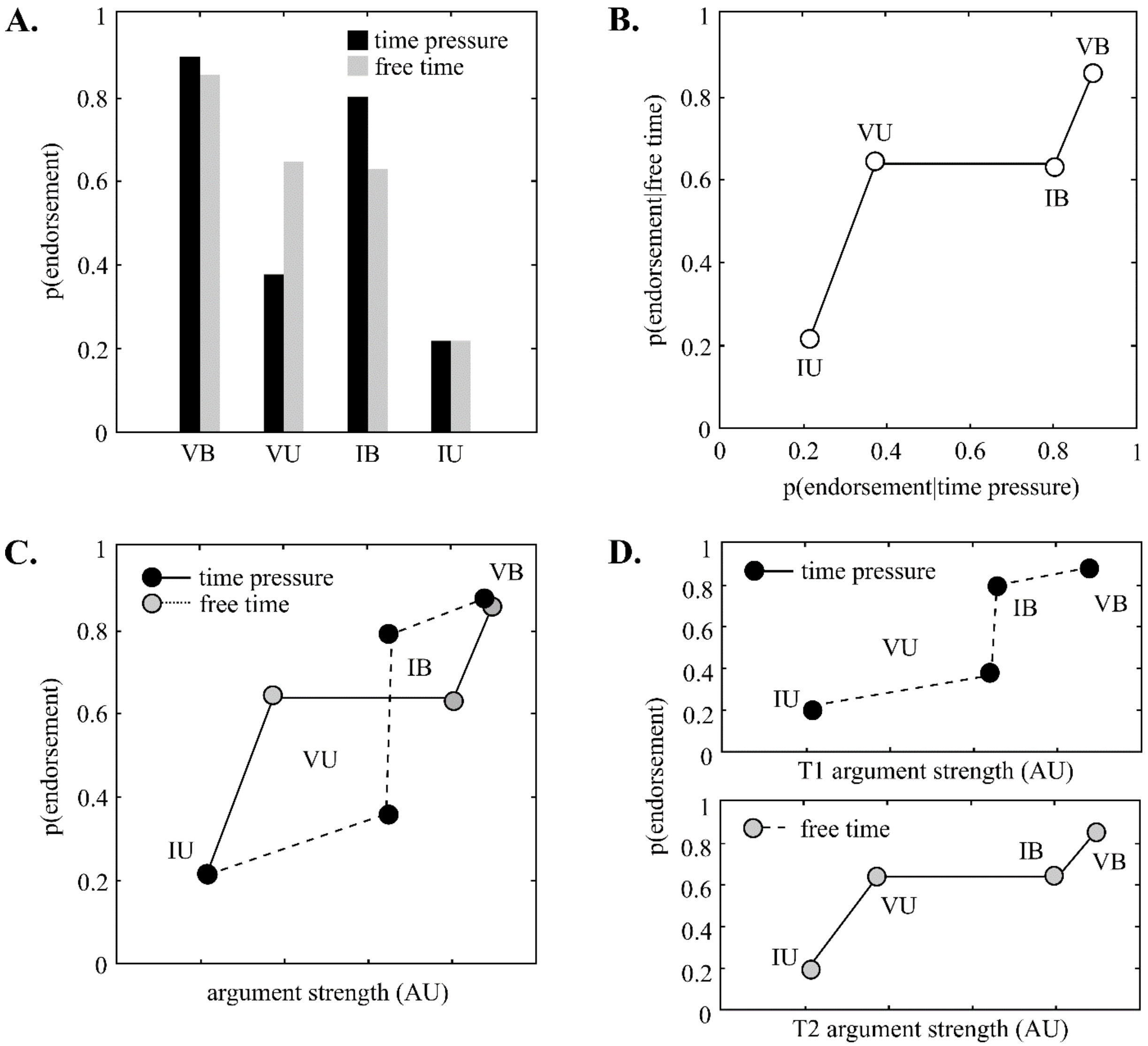

Figure 2. Data reported by Evans & Curtis-Holmes (2005) on the probability of endorsement for valid-believable (VB), valid-unbelievable (VU), invalid-believable (IB), and invalid-unbelievable (IU) syllogisms with and without time pressure. (B) State-trace plot of the data reported in A: the probability of endorsement increases in the same order with and without time pressure: IU < VU < IB < VB. (C) A monotonic one-process model that explains the probability of endorsement as a function of a single subjective metric for argument strength at two levels of time pressure (some and none), which is computed by a single type of reasoning. (D) A monotonic two-process model that explains the probability of endorsement for each level of time pressure as a function of one subjective metric for argument strength, which is computed by one type of reasoning (T1 reasoning for time pressure and T2 reasoning for free time).

make: the response variables (endorsement rates with and without time pressure) can be mapped via monotonic functions onto each other.[4] To graphically test this prediction, we can use a *state-*

[4] Let P_{TP} and P_{FT} be endorsement rates with and without time pressure, let L be the single latent variable, and let F_1 and F_2 be functions of L, such that $P_{TP} = F_1(L)$ and $P_{FT} = F_2(L)$. Then $L = F_1^{-1}(P_{TP})$ and $L = F_2^{-1}(P_{FT})$. F_1^{-1} and F_2^{-1} are functions (and monotonic functions, in particular) if and only if F_1 and F_2 are monotonic functions. Since one-process models are distinct from two-process models in claiming that L is a function of both P_{TP} and P_{FT}, they permit us to substitute L: $P_{TP} = F_1(F_2^{-1}(P_{FT}))$ and $P_{FT} = F_2(F_1^{-1}(P_{TP}))$. $F_1(F_2^{-1})$ and $F_2(F_1^{-1})$ are functions (and monotonic functions, in particular) if and only if F_1 and F_2 are monotonic functions. Therefore, if F_1 and F_2 are monotonic functions over a single latent variable, then P_{TP} and P_{FT} are monotonic functions of each other.

trace plot (Bamber, 1979; Dunn, 2008), which plots the response variables as functions of each other. If we can draw a monotonic curve through the data without significant error, the prediction is correct, and the data are best explained by monotonic models.

A state-trace plot visually confirms that we can draw a monotonic curve through the data with very little error (Fig. 2B). Since this monotonic one-process model explains the data very well without assuming linearity, there is no need to introduce a second latent variable and hence, there is no need to use a two-process model. We can visualize this one-process model even better by plotting the data points on the single latent variable, instead of plotting them against each response variable. I've chosen one possible way to do this for the sake of illustration: I copied the data points for the free time condition and reflected the data points for the time pressure condition over the diagonal (Fig. 2C). This suggests that differences in mean endorsement rates are proportionate to differences in subjective argument strength.[5] Of course, I could have illustrated it many other ways, since the only constraint is monotonicity: the data points must have the same orders over the X and Y axes.

Now, let's consider the interpretation of this one-process model. Given monotonicity, it predicts that the ordering of preferences will be identical across conditions: with and without time pressure, the single type of reasoning computes argument strength by (a) prioritizing validity (IU to VU), (b) prioritizing believability (VU to IB), and (c) again prioritizing validity (IB to VB). The only difference between time pressure and free time is the relative strengths of the preferences: time pressure decreases its preference for validity (when it prefers validity) and increases its preference for believable conclusions (when it prefers them). This is a quantitative difference, as the single-process theorist would say, not a qualitative one.

The monotonic one-process model fits the data so well that visual inspection confirms that it's unnecessary for the two-process model to use a second latent variable to explain the data (Fig. 2D). Stephens et al. (2019) used state-trace analysis with *conjoint monotonic regression* (Kalish et al., 2016) to formally confirm that the monotonic one-process model explains the response rates better than the monotonic two-process model does. In fact, they even showed that *all* of the data that Evans & Stanovich (2013) cite as evidence for DPT is actually best explained by one-process monotonic models, except for three studies under extremely optimistic assumptions about variance (i.e., Klauer et al., 2000; De Neys et al., 2005; Roberts & Newton, 2001).

Since monotonic models are appropriate for explaining endorsement rates on reasoning tasks and linear models aren't, this is sufficient reason to conclude that the data we've considered is actually evidence for SPT, not for DPT. But it's important to ask: why did Evans & Curtis-Holmes (2005) conclude that their data indicate a two-process model? Well, they didn't build or formally compare these models. Instead, they calculated a belief index (VB + IB – VU – IU) and a logic index (VB + VU – IB – IU) and used t-tests to show that the belief index was significantly larger with time pressure and the logic index was significantly larger without time pressure. This was supposed to demonstrate that T1 reasoning (which prefers believability) can be dissociated from T2 reasoning (which prefers validity).

[5] Note that the reflection of the data points is a linear function. Does this undermine our monotonic analysis? No, Figure 1C arbitrarily stipulates that there is a linear relation between the two performance variables just for the sake of illustration, but it doesn't assume that the response functions themselves are linear.

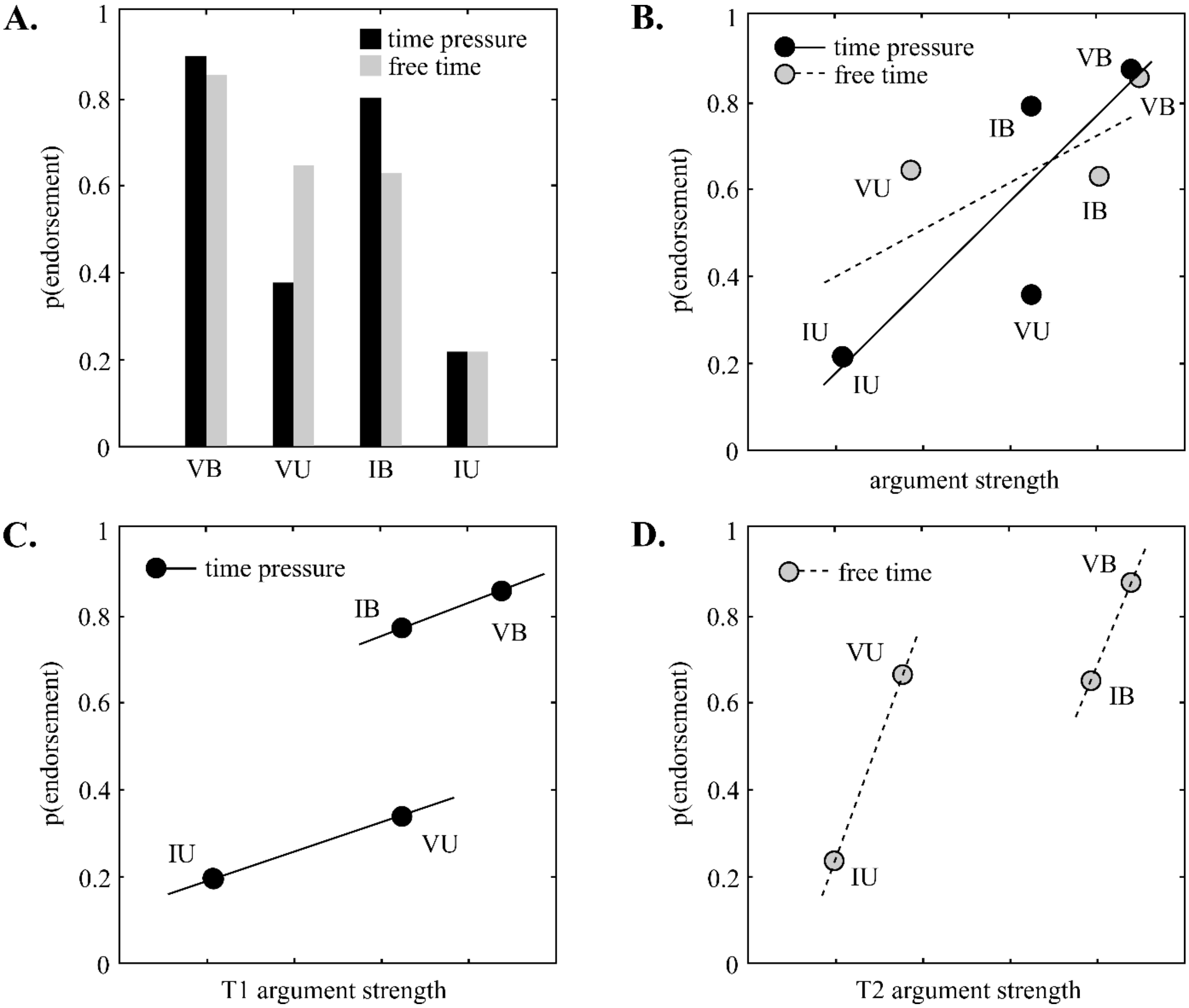

Figure 3. (A) Data reported by Evans & Curtis-Holmes (2005), reproduced for convenience. (B) A linear one-process model that explains endorsement rates under two levels of time pressure (none and some) as a function of a single underlying measure of argument strength (computed by one type of reasoning). Note the low fit. (C) The response function for responses under time pressure in a linear two-process model, represented at two levels of believability. Note the perfect fit. (D) The response function for responses under free time in a linear one-process model, represented as two levels of believability. Note the perfect fit again. Also note the much higher slopes for the lines in Panel D compared to the lines in Panel C: T2 reasoning has a much stronger preference for validity than T1 reasoning has.

However, their use of t-tests presumes that the underlying response functions are linear (Stephens et al., 2019). Implicitly, then, Evans & Curtis-Holmes (2005) have assumed that the response functions from each latent variable to each performance variable must be linear, despite evidence to the contrary. It turns out that when we do assume that response functions are linear, two-process models predict the data significantly better than one-process models. In Fig. 3, I've plotted the data on the latent variables in the same way that I did in Fig. 2: I've preserved the order and distances between the mean endorsement rates for each syllogism in the order and distances between their subjective argument strength. Again, though, any way of plotting them would be acceptable, so long as the order of data points on the X and Y axes are the same (monotonicity).

The differences among the three other panels concern the linear models (the straight lines) that are fitted to the data. Fig. 3B represents a *linear* one-process model: it explains the mean endorsement rates with and without time pressure as two *linear* functions of the same latent variable. Note how poorly it fits the data. Fig. 3C and Fig. 3D represent the response functions for the same *linear* two-process model. Fig. 3C explains the mean endorsement rates with time pressure for different levels of believability as functions of one latent variable, which represents the subjective argument strength computed by T1 reasoning. Fig. 3D explains the mean endorsement rates without time pressure for different levels of believability as functions of one latent variable, which represents the subjective argument strength computed by T2 reasoning. Note how perfectly it fits the data.[6]

The linear two-process model fits the data so much better than the linear one-process model that visual inspection confirms that it's necessary for linear models to use a second latent variable to explain the data. And the t-test by Evans & Curtis-Holmes (2005) formally confirms this. This explains why dual-process theorists have cited these data as evidence for DPT: they use linear statistics (e.g., t-tests) without questioning or justifying their assumption that linear models are appropriate for explaining responses to reasoning tasks and therefore, they find that two latent variables are required to fit the data. This error illustrates the importance of cognitive modelling: it forces us to choose our assumptions more carefully.

§2. Signed Difference Analysis

Stephens et al. (2019) have shown that none of the evidence that Evans & Stanovich (2013) cite confirms DPT or falsifies SPT. However, this doesn't confirm SPT or falsify DPT because their state-trace analysis has a few caveats. One is that most of the studies that Stephens et al. (2019) analyzed were underpowered for state-trace analysis: e.g., Evans & Curtis-Holmes (2005) used four kinds of tasks, so they only had four data points for their state-trace plot. Studies that are powered for monotonic analysis might find more success in finding sufficient evidence to reject one-process models.

For example, Singmann & Klauer (2011) ran a high-powered study that did find evidence to falsify one-process models and confirm two-process models (see also: Rips, 2001; Oberauer, 2006; Klauer et al., 2010).[7] They used a different pair of conditions than Evans & Curtis-Holmes (2005) to dissociate T1 and T2 reasoning: they had subjects evaluate the same syllogisms under *induction instructions* (i.e., "How likely is the conclusion to be true?") and under *deduction instructions* (i.e., "How valid is the conclusion?"). Then they used a semi-formal kind of monotonic analysis to show that different types of reasoning are involved in evaluating the syllogisms under the induction and deduction instructions. They classified the first as T1 reasoning and the second as T2 reasoning.

Another, deeper caveat is that both the one- and two-process models that Stephens et al. (2019) and Singmann & Klauer (2011) used fail to distinguish between two distinct kinds of parameters: (a) sensitivity and (b) response bias (Stephens et al., 2018). Sensitivity is the ability to discriminate between *targets* and *lures* (stimuli that should be endorsed and rejected, respectively) and response

[6] In this example, the fit is perfect because the model is *saturated*: two linear functions can use two latent variables to perfectly fit four mean responses. But the same point holds for results with more than four mean responses: two linear functions on two latent variables can always fit response data better than two linear functions on one latent variable.
[7] Stephens et al. (2018) later confirmed with Kalish et al.'s (2016) conjoint monotonic regression that this evidence was statistically significant.

bias is the disposition to endorse, regardless of whether the stimulus is a target or lure.[8] It is known from *signal detection theory* that the best explanation of performance on identification tasks (e.g., endorsement tasks) requires this distinction (Macmillan & Creelman, 2005). Hence, the failure of one-process models to explain the evidence reported by Singmann & Klauer (2011) may only be the result of their using one latent variable, when they should have used two distinct parameters.

To rectify this caveat, Stephens et al. (2018) developed multi-parameter models for SPT and DPT to explain the data reported by Singmann & Klauer (2011). Like the models that we considered in §1, these models start from latent variables that map from the task variables and are best interpreted as subjective measures of argument strength computed by one or two types of reasoning. Unlike the models in §1, though, both valid and invalid syllogisms form different distributions over each latent variable: the former at higher values and the latter at lower values to the extent that subjective argument strength is a reliable proxy for objective argument strength (Fig. 3A). *Sensitivity* is a parameter that measures the distance between the means of these distributions: the further apart they are, the greater the agent's power to discriminate between them.[9]

Since one-process models use one latent variable to represent the argument strength computed by a single type of reasoning, they use one parameter (x) to represent the sensitivity of that single type of reasoning to the validity and invalidity of syllogisms. Since two-process models use two latent variables to represent the argument strengths computed by T1 and T2 reasoning, they use two parameters (x_1 and x_2) to separately represent the sensitivity of T1 and T2 reasoning to the validity and invalidity of syllogisms. Every formulation of DPT would agree that the sensitivity of the argument strength that's computed by T2 reasoning is greater than the sensitivity of the argument strength that's computed by T1 reasoning: i.e., $x_1 < x_2$.

For the agent to decide whether to endorse any syllogism, they require a response criterion, such that they produce an endorsement response only when the latent variable takes on a value greater than the criterion for the syllogism. Low positions represent a response bias to endorse, and high positions represent a response bias to reject. In a one-process model, the agent has one subjective measure of argument strength, so they decide whether to endorse any syllogism by considering its value for the same latent variable and then comparing it to either its induction criterion (c_I) or its deduction criterion (c_D). Since logical necessity is a much more demanding standard of argument strength than inductive plausibility, the criterion for endorsement under deduction instructions is likely set higher than the criterion for endorsement under induction instructions (Rips, 2001). So, deductive responses involve a greater bias to reject, and inductive responses involve a greater bias to endorse. Like sensitivity, the deduction and induction criteria both count as parameters.

In a two-process model, though, the agent has two subjective measures of argument strength: one that measures inductive argument strength with sensitivity x_1 and is computed by T1 reasoning, and one that measures the deductive argument strength with sensitivity x_2 and is computed by T2 reasoning. So, the agent can decide whether to endorse any syllogism by using only the subjective measure of argument strength that is relevant to the task and then applying the relevant criterion

[8] Rotello & Heit (2009) and Heit & Rotello (2010) were the first to apply insights from signal detection theory to the debate between SPT and DPT.

[9] Although the argument strength of any given syllogism counts as a latent variable in this model, sensitivity counts as a parameter—it is treated as a fixed value for all agents.

to that subjective measure. That is, the agent can evaluate a syllogism under induction or deduction instructions by looking up the value that it takes for inductive or deductive argument strength and then comparing that value to its induction (c_I) or deduction (c_D) criterion, respectively. Therefore, this two-process model assumes that T1 reasoning is *exclusively* responsible for endorsement under induction instructions and T2 reasoning is *exclusively* responsible for endorsement under deduction instructions. I'll criticize this assumption at length in §3.

Since variables are much more difficult to model than parameters, Stephens et al. (2018) exclude the independent and latent variables from their one-process model and only include the parameters: x, c_I, and c_D for the one-process model and x_1, x_2, c_I, and c_D for the two-process model. Next, they define linear *structural functions* (u) over these parameters and then use monotonic *measurement functions* (F) to map those structural variables onto the endorsement rates under the four conditions (P). The structural functions are different for one- and two-process models since they have to map different sets of parameters for each model onto the same structural variables (Table 1; Fig. 3A–C). But the measurement functions are the same for both models (Table 1; Fig. 3D).

Structural functions (1-process)	Structural functions (2-process)	Measurement functions	Response rates
$u_{vD} = x - c_D$	$u_{vD} = x_1 - c_D$	F_{vD}	P (endorsement | valid, deduction) = $F_{vD} (u_{vD})$
$u_{iD} = - c_D$	$u_{iD} = - c_D$	F_{iD}	P (endorsement | invalid, deduction) = $F_{iD} (u_{iD})$
$u_{vI} = x - c_I$	$u_{vI} = x_2 - c_I$	F_{vI}	P (endorsement | valid, induction) = $F_{vI} (u_{vI})$
$u_{iI} = - c_I$	$u_{iI} = - c_I$	F_{iI}	P (endorsement | invalid, induction) = $F_{iI} (u_{iI})$

Table 1. Stephens et al.'s (2018) *one- and* two-process models, which explain the mean endorsement rates of valid and invalid syllogisms under deduction and induction instructions.

So, the endorsement rate for invalid syllogisms under deduction instructions (P_{iD}) is explained by a monotonic measurement function (F_{iD}) over the structural variable (u_{iD}) that takes the distance between the response criterion (c_D) and the center of the distribution of invalid syllogisms (which is set at 0 in both models): $0 + c_D$. But the models take the negative of the deduction criterion ($-c_D$), so that the lower the deduction criterion, the shorter this distance, the larger this value, and the higher the endorsement rate (P_{iD}): $u_{iD} = - c_D$. Likewise, the endorsement rate for valid syllogisms under deduction instructions (P_{vD}) is explained by a measurement function (F_{vD}) over the structural function (u_{vD}) that takes the distance between the center of the distribution of valid syllogisms ($x - 0$ or $x_2 - 0$)[10] and the deduction criterion: $x - c_D$ or $x_2 - c_D$. The lower the deduction criterion, the longer this distance and the higher the endorsement rate (P_{vD}): $u_{vD} = x - c_D$ (in the one-process model) or $u_{vD} = x_2 - c_D$ (in the two-process model). The same explanation is extended to responses under induction instructions.

[10] Recall that the center of the distribution of invalid syllogisms is set at 0 and the distance between the centers of the distributions is x or x_2 in one- and two-process models, respectively.

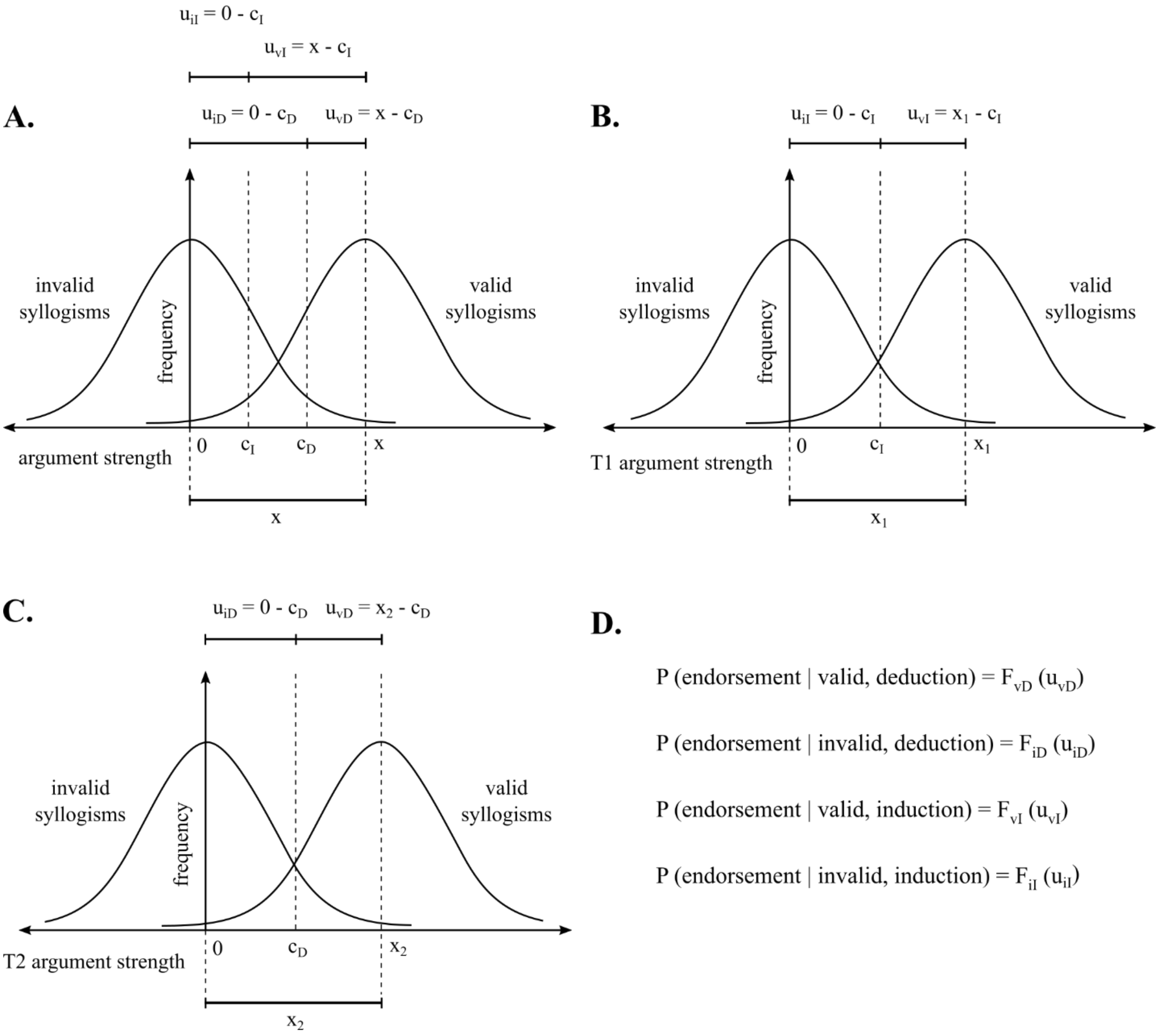

Figure 4. (A) A one-process model with all structural functions defined over one sensitivity parameter (x) and two response criteria (c_I and c_D). (B) Half of a two-process model with two structural functions (for induction instructions only) defined over one sensitivity parameter (x_1) and one response criterion (c_I). (C) The other half of the same two-process model with two structural functions (for deduction instructions only) defined over one sensitivity parameter (x_2) and one response criterion (c_D). (D) The measurement functions that map the structural functions onto response rates in four conditions, which are shared by both the one-process and two-process models.

State-trace analysis cannot adjudicate between these one- and two-process models because it can only test for monotonicity between two response variables and these models have four response variables. So, Stephens et al. (2018) instead used *signed difference analysis* (SDA), which is a generalized form of state-trace analysis (Dunn & James, 2003; Dunn & Anderson, 2017). SDA asks whether one or two sensitivity parameters are needed to explain how changes in contingent task conditions (which are independent of response functions) cause changes in the four response variables via the structural variables, assuming that the measurement functions (F) are monotonic. For example, SDA can ask whether one or two sensitivity parameters are needed to explain—given that F is monotonic—how changing syllogisms from *counter-logical* (syllogisms that are valid if and only if their content is unbelievable) to *pro-logical* (syllogisms that are valid if and only if their content is believable)—both contingent task conditions—causes the observed changes in the four response variables.

These changes in the structural and response variables can be represented by *sign vectors*. So, for example, (+ + + +) is the sign vector that represents a change from one contingent task condition to another that has the same effect on all four response variables.[11] Singmann & Klauer (2011) haven't reported any statistically significant occurrences of this sign vector (Stephens et al., 2018) but Evans & Curtis-Holmes (2005) appear to have. They found that a change from syllogisms with unbelievable conclusions to those with believable conclusions had the same effect on all four response functions: it increased endorsement rates regardless of correctness and time pressure (Fig. 2A; see Footnote 10). Hence, a one-process model can easily explain this change: increasing believability simply increases the response bias to endorse, regardless of correctness and time pressure. As expected, then, SDA confirms the conclusion of our state-trace analysis: Evans & Curtis-Holmes (2005) report data that are consistent with one-process models.

Some sign vectors won't be possible, though, given a particular model. Stephens et al. (2018) note that their one-process model forbids only one sign vector: (+ – – +). After all, an increase in the distance between the deduction criterion and the center of the valid syllogisms distribution ($u_{vD} = x - c_D$) could be the result of an increase in sensitivity (x) or a decrease in the deduction criterion (c_D). But the second function decreases ($u_{iD} = - c_D$), such that the deduction criterion (c_D) must be increasing. Therefore, the sensitivity (x) must be increasing. Simultaneously, though, the distance between the induction criterion and the center of the valid targets distribution ($u_{vI} = x - c_I$) has decreased, which could be the result of a decrease in sensitivity (x) or an increase in induction criterion (c_I). Since we've already concluded that sensitivity (x) must be increasing, we must infer that the induction criterion (c_I) is increasing. (We aren't forced to infer this if there are two sensitivity parameters.) However, the fourth function is also increasing ($u_{iI} = - c_I$), such that the induction criterion (c_I) must be decreasing. This is a contradiction, so (+ – – +) contradicts the one-process model. We can resolve this contradiction, though, if we add a second sensitivity parameter and allow that T1 sensitivity (x_1) is decreasing and T2 sensitivity (x_2) is increasing.[12]

To refute a one-process model and confirm a two-process model, we'd need to find sufficiently many contingent task conditions that are such that changing between them produces sufficiently large changes in the response variables that are represented by the forbidden sign vector (+ – – +). Stephens et al. (2018) used an approximation of conjoint monotonic regression (Kalish et al., 2016) to develop several monotonic models through the four-dimensional state-trace plot and then approximated the fit. On all three metrics—frequency, minimum deviation, and goodness-of-fit—Stephens et al. found that the model that best explained all the evidence was the one-process model

[11] The sign vector is calculated by subtracting the vector of response variables for one contingent task condition from the vector of response variables for the other contingent task condition, discarding the values, and keeping only their signs. For example, take the response vectors reported by Evans & Curtis-Holmes (2005), where each has the form (P(E | V, free time) P(E | I, free time) P(E | V, time pressure) P(E | I, time pressure)). Figure 2A suggests that the response vectors for believable and unbelievable syllogisms are approximately (0.85 0.65 0.9 0.8) and (0.4 0.63 0.2 0.2), respectively, so the difference vector is (0.85 – 0.4 0.65 – 0.63 0.9 – 0.2 0.8 – 0.2) = (+0.35 +0.02 +0.7 +0.6) and the sign vector is (+ + + +).

[12] Since the values of these linear functions over the parameters are mapped onto endorsement rates by monotonic functions (F_{vD}, F_{iD}, F_{vI}, and F_{iI}), the permitted and forbidden sign vectors should be preserved in the response data.

that I've been discussing. In fact, they found no significant evidence that there were sufficiently large or frequent instances of the forbidden sign vector to reject it. From this impressive result, they concluded that the success of this one-process model provisionally confirms SPT and refutes DPT.

§3. Defining Properties

Stephens et al. (2018) have confirmed that even high-powered studies specifically designed for monotonic analysis fail to produce evidence that confirms DPT or falsifies SPT—once we allow one-process models to use multiple parameters in order to account for the core insights of signal detection theory. But this still doesn't confirm SPT or falsify DPT because important caveats still remain with their state-trace analysis of all the data cited by Evans & Stanovich (2013) and their SDA of Singmann & Klauer's (2011) data. Unlike the previous caveats, though, these caveats are theoretical and experimental more than modelling, so the nature of our analysis will shift into more familiar territory for readers from the reasoning literature.

One caveat of both the state-trace analysis (Stephens et al., 2019) and SDA (Stephens et al., 2018) is that their two-process models assume that each response variable is only explained by one latent variable. Their two-process models used (a) only the single latent variable that represents argument strength computed by T1 reasoning to explain mean endorsement rates under time pressure (for Evans & Curtis-Holmes, 2005) and induction instructions (for Singmann & Klauer, 2011) and (b) only the single latent variable that represents argument strength computed by T2 reasoning to explain mean endorsement rates under free time and deduction instructions. Thus, they assume that each condition effectively isolates T1 or T2 reasoning. But that assumption is quite unrealistic: any dual-process theorist would insist that each condition elicits both T1 and T2 reasoning, just in different proportions.

Another caveat of both analyses is that even if they did find evidence to reject one-process models, it's unclear whether this confirms DPT because it's unclear whether the two types of processing that are represented by the two-process models are the same things as the two types of reasoning that are purported to exist by DPT. Suppose that Singmann & Klauer's (2011) data did confirm a two-process model: that would confirm that inductive and deductive reasoning are qualitatively different, not that T1 and T2 reasoning are qualitatively different. Likewise, suppose that Evans & Curtis-Holmes' (2005) data did confirm a two-process model: that would confirm that rushed and unrushed reasoning are qualitatively different, but that would confirm that T1 and T2 reasoning are qualitatively different only if rushed and unrushed reasoning just are T1 and T2 reasoning.

Both caveats arise from the same problem: there isn't a necessary relation between the dissociating conditions and the T1/T2 reasoning distinction. Thus, the conditions fail to selectively dissociate T1 and T2 reasoning. As a result, DPT can be true even if the evidence confirms a one-process model (as in the first caveat) and DPT can be false even if the evidence confirms a two-process model (as in the second caveat). This problem indicates the limits of modelling: SDA can tell us whether two types of processing respond to tasks under two different conditions, but it can't tell us whether those two types of processing count as T1 and T2 reasoning—as opposed to some other pair of processing types. To ensure that any type-difference that SDA finds is the T1/T2 difference, we need to design experiments with conditions that succeed in effectively isolating responses with the defining properties of T1 and T2 reasoning.

Before we can do that, we must develop a plausible formulation of DPT that specifies the defining properties of T1 and T2 reasoning. One constraint is that the defining properties of T1 and T2 reasoning must be duals (or opposites), where X is a dual of Y if and only if not-X is just Y and not-Y is just X. Most formulations of DPT aim to satisfy this condition when they define T1 vs. T2 reasoning as being automatic vs. effortful (Kahneman, 2011), autonomous vs. controlled (Evans & Stanovich, 2013), or intuitive vs. deliberative (De Neys, 2021). Evans & Curtis-Holmes (2005) used conditions that did dissociate dual properties: time pressure and free time are dual properties because not having time pressure is just having free time and not having free time is just having time pressure. By comparison, Singmann & Klauer (2011) used conditions that dissociated non-dual properties: being inductive and deductive aren't dual properties because it is possible to be neither inductive nor deductive, as in the case of abductive reasoning.

By this standard, then, Evans & Curtis-Holmes' (2005) study is more relevant to testing DPT than Singmann & Klauer's (2011) study. Moreover, Evans has defended a formulation of DPT, which claims that the defining properties of T1 and T2 reasoning include lacking and having access to working memory (Evans & Stanovich, 2013; c.f. Evans, 2017). These properties satisfy the duality condition too: to not lack access to working memory is just to have access to it and to not have access to working memory is just to lack access to it. So, if SDA did find that there are two types of processing involved in responses to their experiment, the best interpretation of that result would be that time pressure had impaired access to one type of reasoning that lacked access to working memory (maybe T1 reasoning) and inhibiting the other type of reasoning (maybe T2 reasoning), which had access to working memory.

Another constraint on any plausible formulation of DPT is that the defining properties of T1 and T2 reasoning must be discrete, not continuous. After all, if the defining properties of T1 and T2 reasoning are continuous, then T1 and T2 reasoning would only be *quantitatively* different (per SPT), not *qualitatively* different (per DPT). This objection has often been raised against popular formulations of DPT: while the difference between lacking vs. having access to working memory, being autonomous vs. controlled, etc. may be discrete, no reasoning processes entirely lack access to working memory, entirely have autonomy, etc., so T1 and T2 reasoning would only differ in the amounts of access that they have to working memory, the amount of autonomy that they have, etc. (Newstead, 2000; Keren & Schul, 2009; Kruglanski & Gigerenzer, 2011; De Neys, 2021).

Some dual-process theorists have defended popular formulations of DPT by suggesting that sufficiently large quantitative differences can count as qualitative differences (Samuels, 2009). De Neys (2021) has argued that this view is intractable: no amount of data or theory can tell us whether the difference between the defining properties of T1 and T2 reasoning is *sufficiently* small to count as quantitatively different or *sufficiently* large to count as qualitatively different (De Neys, 2021; see also Newstead, 2000; McCallum et al., 2002; Melnikoff & Bargh, 2018). The standard for sufficiency is arbitrary: e.g., we'd have to stipulate what amount of difference in access to working memory for T1 and T2 reasoning is sufficient for them to count as qualitatively different.

If we formulate DPT as claiming that the defining properties of T1 and T2 reasoning are duals of a *discrete* variable, though, we sidestep these murky theoretical and empirical issues. This gives discrete formulations of DPT a significant advantage. Now, at least one pair of dual, discrete properties is often attributed to T1 and T2 reasoning: the default vs. intervening distinction. These are duals since not being default is just being intervening and not being intervening is just being default. And they're discrete, since being default and being intervening lie in an ordinal sequence: default reasoning occurs *first* and under certain conditions, intervening reasoning occurs *second*. Since the default vs. intervening distinction satisfies both of our theoretical constraints, it provides a promising formulation of DPT: T1 reasoning is defined (at least, partly) as *default* reasoning and T2 processing is defined (at least, partly) as *intervening* reasoning.[13] [14]

Before we continue down this theoretical line, let's recall our purpose for developing a plausible formulation of DPT. Our question is whether DPT is correct in claiming that reasoning is divided into qualitatively different, dual types. We found in §1 and §2 that the available data indicates SPT under monotonic analysis. In the beginning of this section, though, I argued that these modelling results are biased against DPT because the models assume that the task conditions in experiments *exclusively* elicit T1 and T2 reasoning, which most dual-process theorists would reasonably deny. To design experiments with conditions that succeed in effectively isolating T1 and T2 reasoning, though, we must select a plausible formulation of DPT. I've argued that one popular formulation of DPT is theoretically plausible: T1 and T2 reasoning have the defining properties of being default and intervening, respectively.

§4. Metacognitive Control

The defaulting/intervening distinction is a coherent definition of the T1/T2 distinction, but it isn't obviously interesting or relevant: why does it matter whether the default/intervening distinction is qualitative or quantitative? How would knowing whether it's qualitative or quantitative affect our understanding of reasoning? How does it help us in designing task conditions that can effectively isolate T1 and T2 reasoning? These are questions that we must answer if our theoretical discussion is to be experimentally productive (De Neys, 2021). The goal of this section is to develop the default/intervening definition of the T1/T2 distinction into a substantive distinction that is relevant for understanding reasoning. Then we'll use it to develop an integrated experimental strategy in §5.

What makes the distinction between default and intervening reasoning relevant to understanding reasoning is they stand in different relations to metacognitive control, which decides whether to engage intervening reasoning in response to default reasoning (e.g., Thompson, 2009; Pennycook et al., 2015; Ackerman & Thompson, 2017). I propose that default reasoning is just reasoning that *precedes* metacognitive control and intervening reasoning is just reasoning that *follows*

[13] I've stated and briefly defended this view before in a commentary on De Neys 2021 (Dewey, 2021). This paper provides a full elaboration (and, in few places, correction) of the ideas mentioned in that commentary.

[14] While most dual-process theorists accept that T1 and T2 reasoning are default and intervening, respectively, it hasn't been proposed they are defining properties (to my knowledge). That said, most dual-process theorists now reject the exact formulations of default-interventionism in favour of a hybrid theory, once dubbed "Dual-Process Theory 2.0" by De Neys (2017). So, I should clarify: my proposal that the T1/T2 distinction is (partly) defined by the default/ intervening distinction is consistent with this hybrid theory.

metacognitive control. Other familiar distinctions can be reinterpreted relative to metacognitive control too: e.g., the distinction between T1 reasoning being autonomous and T2 reasoning being controlled, if autonomy is understood not as just being completely uncontrolled (Evans & Stanovich, 2013) but as preceding metacognitive control. This refines our formulation of DPT from §3.

To better understand the importance of this formulation of DPT for our understanding of reasoning, let's review the causal relations between reasoning and metacognitive control. Ackerman & Thompson (2017) offer a useful framework that maps out several possible relations between reasoning and metacognitive control. They suggest that reasoning consists of several things (e.g., task identification, judgments of solvability), but we'll consider just the steps that include and mediate between default and intervening reasoning, which forms and selects the judgment responses.

First, default reasoning produces several cues while it develops an initial judgment about a task. One reasoning cue is *processing conflict* (i.e., the perceived contradiction among two or more processes), which is correlated with actual contradiction among the contents of multiple reasoning responses. As such, processing conflict can be a heuristic for error (De Neys, 2012, 2014). Another example is *processing fluency* (i.e., the perceived ease of responding), which is anti-correlated with processing conflict (Bonner & Newell, 2010; Thompson et al., 2011; Gangemi et al., 2015) and it's correlated with past processing (which is itself correlated with correct processing probably because past processing tends to be reinforced by positive feedback and diminished by negative feedback). As such, processing fluency can be a heuristic for correctness.

Second, *metacognitive monitoring*, the first stage of meta-reasoning, collects these reasoning cues, uses them as *metacognitive heuristics* for error, and integrates them into feelings of rightness and error that evaluate the correctness of any judgment (Ackerman & Thompson, 2017). Of course, the reasoning cues aren't perfectly correlated with the correctness conditions of reasoning, so they are imperfect heuristics for metacognitive monitoring (Shynkaruk & Thompson, 2006). Critically, then, metacognitive monitoring can be manipulated by changes to the reliability of reasoning cues: e.g., processing conflict (De Neys, 2012, 2014; Handley & Trippas, 2015; Pennycook et al., 2015; Bago & De Neys, 2017, 2019) and processing fluency (Thompson et al., 2011, 2013; Thompson & Morsanyi, 2012; Thompson & Johnson, 2014). We'll exploit this fact in §5.

Third, *metacognitive control*, the second stage of meta-reasoning, evaluates these feelings of rightness and error against thresholds to decide its response (Ackerman & Thompson, 2017): e.g., if some initial judgment by default reasoning causes feelings of error that exceed the threshold, metacognitive control may decide to intervene. Little is known about the specific effects of this intervention: some speculate that it increases working memory, time, and other resources that are available to reasoning (Thompson, 2009; Pennycook et al., 2015). Most agree, though, that this metacognitive intervention often—but not always (Gigerenzer & Brighton, 2009; Elqayam & Evans, 2011; Evans & Stanovich, 2013)—improves the performance of reasoning, especially during the performance of conflict tasks (where the heuristic response is incorrect). Otherwise, it would be difficult to explain why the reasoning system uses a costly intervening process at all.

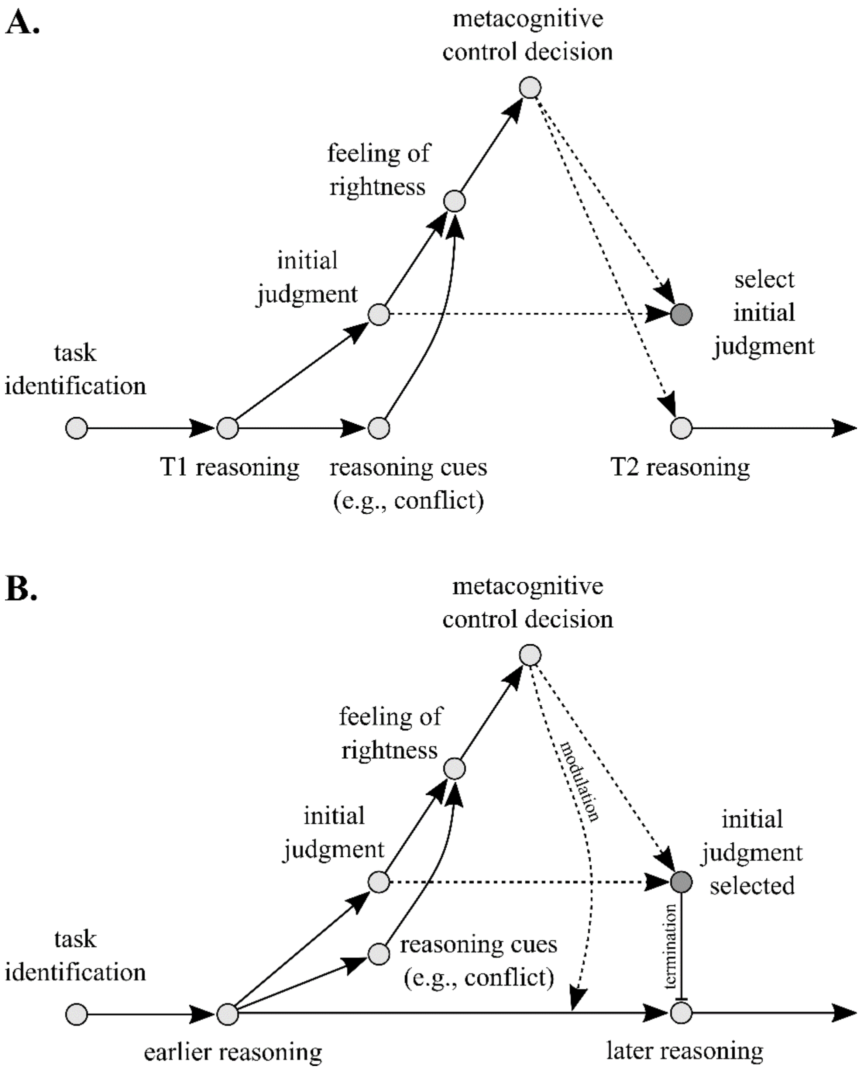

Figure 5. (A) A simplified diagram of DPT's predictions about the causal interactions between reasoning and metacognitive control. Note that the solid arrows represent causation whereas the dashed arrows represent mutually-exclusive instances of causation: e.g., the metacognitive control decision either results in the selection of the initial judgment or the initiation of T2 reasoning. (B) A simplified diagram of SPT's predictions about the causal interactions between reasoning and metacognitive control. Note that the inhibition arrow represents termination: i.e., the selection of an initial judgment terminates any further reasoning about the task.

Ackerman & Thompson's (2017) meta-reasoning framework doesn't make a distinction between T1 and T2 reasoning, but it explains why our formulation of DPT is relevant to our understanding of reasoning: they imply different models of reasoning and metacognitive control. If metacognitive control is an ongoing process that continuously modulates reasoning, per my formulation of SPT, then the difference between reasoning that precedes and follows metacognitive control would be quantitative: any reasoning process would occur both before and after different quantities of modulatory input from metacognitive control (Fig. 5B). This would suggest that meta-reasoning has limited, marginal control over reasoning, or its access to computational resources (time, working memory, etc.). Since its decisions are so limited, it probably has access to relatively little information—perhaps just from reading the reasoning cues, as in Ackerman & Thompson's (2017) meta-reasoning framework.

But if T1 and T2 reasoning is just reasoning that precedes and follows metacognitive control and the difference between T1 and T2 reasoning is qualitative, per my formulation of DPT, then there must be a qualitative difference between reasoning that precedes and follows metacognitive control. For this difference to be qualitative, metacognitive control must be a discrete event that mediates between default and intervening reasoning (Fig. 5A). What constitutes this discrete event? The simplest explanation is that it involves the complete suspension of default reasoning and the initiation of intervening reasoning. But that is unlikely, given the evidence that intuitive, unconscious reasoning continues for long incubation periods after initial exposure to a problem (Sio & Ormerod, 2009). A better kind of explanation is that the discrete event of metacognitive control causes the reasoning outcome in short time scales to become almost exclusively controlled by intervening reasoning.

One version of this explanation is that metacognitive control changes which neural networks implement reasoning: it responds to reasoning cues in the default network that implements T1 reasoning by initiating a response in the intervening network, which then implements T2 reasoning. Once the intervening network is activated to respond to the task, the default network may stop competing with the intervening network by disactivating or settling into an incubation mode, where it may continue to process information, but not in a way that affects short-term judgment-making. While some dual-process theorists have argued that neural imaging evidence indicates this explanation (Evans & Stanovich, 2013), most evidence is better interpreted as dissociating the neural substrates of meta-reasoning from reasoning (e.g., De Neys et al., 2008; Simon et al., 2015; Vartanian et al., 2016; Mevel et al., 2018).

Another version of this explanation is that metacognitive control changes the parameters with which the same neural network implements reasoning: it responds to reasoning cues in the default state of the network that implements T1 reasoning by re-parametrizing the network so that it implements T2 reasoning. This re-parametrization would be a discrete event that qualitatively separates T1 and T2 reasoning. There is evidence that the same network implements T1 and T2 reasoning: the right inferior prefrontal cortex has been implicated in metacognitive control (De Neys et al., 2008; Oldrati et al., 2016; Vartanian et al., 2018; Mevel et al., 2019; Andersson et al., 2020) and it is known to drive inhibitory feedback through subcortical regions back to the same cortical networks that initially trigger metacognitive control (Aron et al., 2004, 2014).[15] However, there is still no evidence whether metacognitive control causes a discrete re-parametrization of the same network (per DPT) or continuously modulates the same network (per SPT).

Both explanations agree that metacognitive control makes an intelligent decision that determines how to implement T2 reasoning and thus establishes a significant, qualitative difference between T1 and T2 reasoning. This suggests that metacognitive control is more intelligent than our formulation of SPT alleged. In fact, it could be so intelligent that it might require access to the semantic content of reasoning, not just the reasoning cues. For example, if the first explanation is right that metacognitive control selects a new network to implement T2 reasoning, it will probably require a lot more information than processing fluency and conflict to make an intelligent decision about which network should take over from the default network, which implements T1 reasoning.

[15] This would also explain why no evidence has been found that different neural networks implement default and intervening reasoning (e.g., De Neys et al., 2008; Simon et al., 2015; Vartanian et al., 2016; Mevel et al., 2018).

Hence, our formulations of DPT and SPT imply different yet novel, relevant, and plausible claims about the architecture of reasoning and meta-reasoning.

§5. Experimental Strategy

Recall from §3 that our reason for identifying the defining properties of T1 and T2 reasoning is to determine what conditions could effectively isolate those defining properties and hence, dissociate T1 and T2 reasoning. Now, if T1 and T2 reasoning are defined by their properties of preceding and following metacognitive control, as I argued in §4, then we should find conditions that effectively isolate T1 and T2 reasoning by effectively preventing and causing metacognitive control. That is, we should find (a) conditions that prevent metacognitive control, selectively elicit T1 reasoning, and so impair performance, and (b) conditions that ensure metacognitive control, selectively elicit T2 reasoning, and so improve performance.[16] Each is a specific kind of biasing and debiasing condition, respectively, so we need to develop a specific kind of debiasing paradigm to decisively test DPT.

In §4, we considered just two metacognitive heuristics: processing conflict and processing fluency. Let's consider each in turn. Processing conflict is positively related with metacognitive control: increasing processing conflict decreases feelings of rightness in metacognitive monitoring, which increases the likelihood that metacognitive control decides to intervene in reasoning. As far as I know, though, there is no direct way to manipulate conflict without changing the task. Changing tasks creates a serious confound, though. If SDA indicates that there is a type-difference between these biasing and debiasing conditions, then the type-difference could be (a) between T1 and T2 reasoning or (b) the two types of default T1 responses that respond to the two different tasks. The only way to rule out the second, confounding explanation is to hold the task fixed and ensure that the biasing and debiasing conditions elicit the same default (T1) responses.[17]

For example, Mastrogiorgio & Petracca (2014) found that a simple modification to the bat-and-ball task ("A bat and a ball cost $1.10. The bat costs $1.01 [vs. $1.00, as in the standard version] more than the ball. How much does the ball cost?") significantly improves performance. One explanation is that the modified task usually causes T2 intervention whereas the standard task usually causes only T1 reasoning. However, another, equally good explanation is that the standard task usually cues the substitution heuristic in T1 reasoning (Kahneman & Frederick, 2005) and the

[16] Recall that this claim only assumes that T1 and T2 reasoning are correlated with lower and higher performance, respectively, on conflict tasks in particular. Hence, this assumption is perfectly consistent with the possibility that there are tasks in which T1 reasoning is correlated with higher performance and T2 reasoning is correlated with lower performance (see Gigerenzer & Brighton, 2009). As a result, this assumption does not imply the problematic claim that T1 and T2 reasoning are defined as less and more reliable reasoning, respectively—a claim criticized at length by Elqayam & Evans (2011).

[17] For certain pairs of tasks, of course, there could be further evidence that indicates that any type-difference found between responses to the two tasks by signed difference analysis is better understood as a difference between T1 and T2 responses than as a difference between two types of default T1 responses that respond to the two different tasks. In such cases, a type-difference found in responses to the two tasks would count as legitimate evidence for DPT. For example, base rate tasks might satisfy this condition: differences in base rate between tasks might elicit the same stereotypical response and elicit base-rate responses that are only quantitatively different. If this is true, we could rule out that any type-difference between responses to different conflict and no-conflict base rate tasks is a type-difference between T1 responses. Then we could infer that it's a type-difference between T2 responses and hence, evidence for DPT. I'm skeptical that we currently have enough evidence to rule out this possibility, though.

modified task totally fails to cue that heuristic and instead cues “logical intuitions” (De Neys, 2012) in T1 reasoning that compute the correct response. In this case, the two tasks dissociate two T1 responses—not T1 reasoning from T2 reasoning. Hence, we cannot use this kind of evidence to test SPT or DPT.[18]

Fortunately, processing fluency is easier to manipulate while controlling for the identity of the task and even task instructions. For the sake of thoroughness, I’ll review four manipulations that define four types of debiasing paradigms. The first type uses task formatting to manipulate processing fluency. For example, Hoover & Healy (2017) found significant increases in performance when the bat-and-ball task was written in algebraic notation (“X + Y = $1.10; X – Y = $1.00; solve for Y.”) rather than the standard, English notation (“A bat and a ball cost $1.10. The bat costs a dollar more than the ball. How much does the ball cost?”) (see also: Mata, 2020). For the bat-and-ball task, we already have evidence to believe that the bat-and-ball task in the English notation elicits two conflicting intuitive responses: a heuristic one and an algebraic one (see De Neys, 2012, 2014 for review). We can expect algebraic notation to elicit the same response. So, it would appear that algebraic notation somehow promotes the selection of the algebraic response by promoting its relative fluency. After all, it’s easier, faster, and more intuitive to interpret “the bat is $1.00 more than the ball” as “the bat is $1.00” (Kahneman & Frederick, 2005; Kahneman, 2011) than to interpret “X = $1.00 + Y” as “X = $1.00”.

There are two explanations of this effect.[19] DPT explains that the algebraic notation promotes the fluency of the algebraic response, causing it to compete more effectively against the heuristic response, resulting in more conflict between T1 responses, which promotes metacognitive control, resulting in T2 reasoning, which selects the algebraic response (Fig. 6A; Pennycook et al., 2015). If this is true, then the bat-and-ball task under English notation should (effectively but imperfectly) isolate T1 reasoning and the bat-and-ball task under algebraic notation should (effectively but imperfectly) isolate T2 reasoning. An SDA could confirm this: a one-process model cannot explain the responses under both notations. By comparison, SPT explains that the improved-fluency algebraic response competes against the impaired-fluency heuristic response, resulting in more conflict in T1 reasoning, which promotes marginal metacognitive modulation until the single type of reasoning sufficiently prefers the algebraic response to select it (Fig. 6B). If this is true, then the bat-and-ball task under English notation should elicit only quantitatively less of the same type of reasoning as the bat-and-ball task under algebraic notation. An SDA could confirm this too.

[18] The same problem extends to other paradigms that use different instructions for biasing and debiasing conditions: e.g., instructions to respond quickly and then respond after further thought (Thompson et al., 2011), instructions to consider alternatives to the initial intuitive response (Hirt & Markman, 1995), instructions to evaluate the inductive vs. deductive correctness of an argument (Rips, 2001), and training on statistical and decision-making principles (Nisbett et al., 1987; Larrick et al., 1990) significantly increase performance. Again, any type-difference that SDA finds could be interpreted as a type-difference in the default responses to different instructions for the same task—rather than as a type-difference between default and intervening responses. Of course, that doesn’t mean that these other debiasing paradigms are useful for studying default and intervening reasoning—just that they aren’t useful for determining whether they are qualitatively or quantitatively different.

[19] A third explanation is consistent with both SPT and DPT: the algebraic notation might sufficiently promote the fluency of the algebraic response that it outcompetes the heuristic response, resulting in its selection as the initial response—without facilitation by metacognitive control. If this were true, then the bat-and-ball task with algebraic notation would fail to selectively elicit putative T2 reasoning. This would make the task unsuitable for settling the debate between SPT and DPT with SDA. Fortunately, this explanation seems unlikely: performance on the bat-and-ball task with algebraic notation is still quite low, which makes it improbable that metacognitive control isn’t engaged.

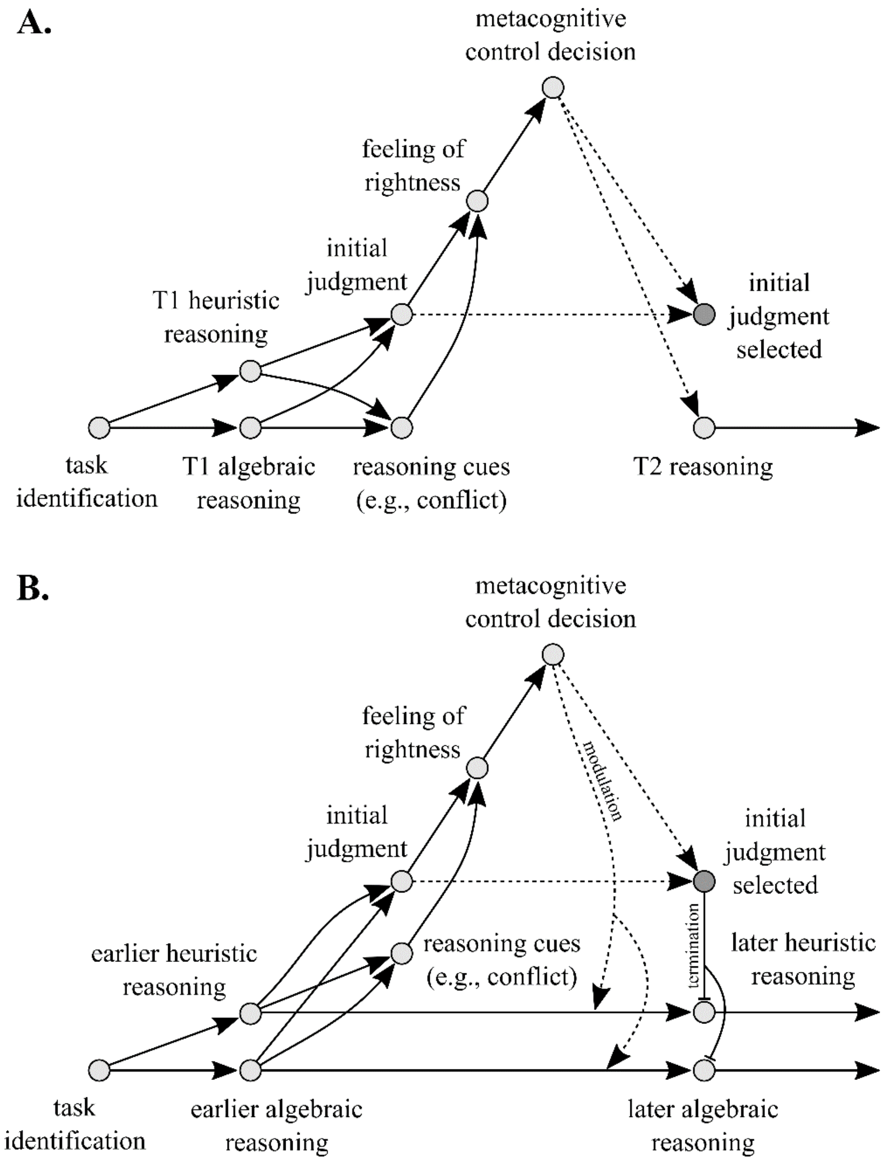
A.
metacognitive control decision
feeling of rightness
initial judgment
T1 heuristic reasoning
initial judgment selected
task identification
T1 algebraic reasoning
reasoning cues (e.g., conflict)
T2 reasoning
B.
metacognitive control decision
feeling of rightness
modulation
initial judgment
initial judgment selected
termination
earlier heuristic reasoning
reasoning cues (e.g., conflict)
later heuristic reasoning
task identification
earlier algebraic reasoning
later algebraic reasoning

Figure 6. The distinction between two types of T1 processes (T1 heuristic and T1 algebraic reasoning) can be confused with the distinction between the two broader types of reasoning processes (T1 and T2 reasoning). However, these are distinct type-differences. (A) A diagram that adds the heuristic/algebraic type-distinction to the diagram from Figure 4A for DPT. (B) A diagram that adds the heuristic/algebraic type-distinction to the diagram from Figure 4B for SPT. Note that the branching arrows indicate that the source node can't differentiate its causal influence on the target notes: e.g., the metacognitive control decision can't differentiate its modulation of heuristic and algebraic reasoning.

The same explanation applies to the other types of debiasing paradigms. The second type uses priming conditions to manipulate processing fluency. For example, Hoover & Healy (2017) found significant and substantial increases in performance when performance on the bat-and-ball task was primed with a similar problem in algebraic notation (e.g., "X + 3Y = 13; Y = 6 – 2X; solve for X") (see also Hoover & Healy, 2021). Here, priming with an algebraic problem appears to temporarily increase the fluency of algebraic reasoning, which allows the intuitive algebraic response to the bat-and-ball task to more effectively compete with the intuitive heuristic response (Pennycook, 2017). Again, there are two explanations why this tends to improve performance: (a) it increases the probability of metacognitive control, which causes T2 intervention (per DPT) or (b) it sufficiently increases the metacognitive modulation of a single type of reasoning (per SPT). SDA can test these explanations by checking whether a one- or two-process model suffices to explain these responses with and without algebraic priming.

The third type of debiasing paradigm uses conditions that promote the salience of processing conflict between the heuristic response and the task. For example, Bourgeois-Gironde & Van der Henst (2009) found significant increases in performance when participants had to evaluate answers that specified the costs of *both* items in the bat-and-ball task and related tasks (e.g. "the ball is $0.05 and the bat is $1.05" vs. "the ball is $0.10 and the bat is $1.10") rather than when they had to evaluate answers that specified the cost of only *one* item (e.g. "the ball is $0.05" vs. "the ball is $0.10"). One explanation is that explicitly representing the comparison between the costs of both items makes it more salient that the heuristic response contradicts the task. That is, it increases the processing fluency of the T1 response that registers the conflict between the favoured response and the task representation (e.g., $1.05 + $0.05 = $1.10 vs. $1.00 + $0.10 = $1.20). This increases processing conflict, which tends to improve performance by either (a) increasing the probability of metacognitive control, causing the intervention of T2 reasoning (per DPT) or (b) increasing the metacognitive modulation of a single type of reasoning (per SPT). SDA can determine which.

Finally, the fourth type of debiasing paradigm includes conditions that indirectly affect processing fluency and processing conflict. For example, time pressure might increase processing fluency (Ackerman & Thompson, 2017) or decrease processing conflict by decreasing the amount of time available for registering conflicts between the most fluent, heuristic response and the less fluent, true response. Thus, time pressure would decrease processing conflict and hence, the probability of metacognitive control, which decreases performance. A similar explanation could be given for cognitive loading too. So, time pressure and cognitive loads might effectively isolate T1 reasoning, but neither effectively isolates T2 reasoning: performance in debiasing conditions is still low, which suggests that analytic engagement often fails to occur. This might explain why Stephens et al. (2019) didn't find enough evidence under these conditions to reject one-process models. To counter this, though, time pressure and cognitive loading could be crossed with debiasing conditions to more effectively isolate T2 reasoning.

Once we've used any one or combination of these four types of debiasing paradigms to effectively isolate default and intervening reasoning on the same tasks, we can modify Stephens et al.'s (2018) one- and two-process models to explain the responses. The one-process model will need just three parameters: (a) sensitivity (x), which is the difference between the means of the distributions of correct and incorrect responses over a latent variable and which we can interpret as subjective argument strength, (b) the default response criterion (c_1), and (c) the intervening response criterion (c_2) (Table 2). And the two-process model will need four parameters: (a) the sensitivity of T1 reasoning (x_1), (b) the sensitivity of T2 reasoning (x_2), (c) the response criterion of T1 reasoning (c_1), and (d) the response criterion of T2 reasoning (c_2) (Table 2). Then the structural and measurement functions can be defined in the same way to explain endorsement rates for correct and incorrect solutions to tasks under biasing and debiasing conditions.

Structural functions (1-process)	Structural functions (2-process)	Measurement functions	Response rates
$u_{CD} = x - c_2$	$u_{CD} = x_2 - c_2$	F_{CD}	P (endorsement | correct, debiasing) = F_{CD} (u_{CD})
$u_{ID} = - c_2$	$u_{ID} = - c_2$	F_{ID}	P (endorsement | incorrect, debiasing) = F_{ID} (u_{ID})
$u_{CB} = x - c_1$	$u_{CB} = x_1 - c_1$	F_{CB}	P (endorsement | correct, biasing) = F_{CB} (u_{CB})
$u_{IB} = - c_1$	$u_{IB} = - c_1$	F_{IB}	P (endorsement | incorrect, biasing) = F_{IB} (u_{IB})

Table 2. An adaptation of Stephens et al.'s (2018) one- and two-process models, which uses three and four parameters, respectively, to explain the mean endorsement rates of correct and incorrect solutions to tasks under suitable biasing and debiasing conditions, which manipulate reasoning cues (metacognitive heuristics) while holding the tasks fixed.

Finally, we need to create variance to challenge the one- and two-process models. In particular, we need to look for further contingent task conditions that produce sign vectors that are forbidden by the one-process model. Recall from §2 that (+ – – +) was the forbidden sign vector for Stephens et al.'s (2018) one-process model. Since our one-process model is isomorphic to theirs, (u_{CD} u_{ID} u_{CB} u_{IB}) = (+ – – +) is the only forbidden vector for our one-process model too. The goal of our experimental strategy should be to find this vector: i.e., find sufficiently many contingent task conditions (defined independently of correctness and debiasing) that—for any given task—can sufficiently (a) increase the rate of endorsement of a correct solution under debiasing conditions, (b) decrease the rate of endorsement of an incorrect solution under debiasing conditions, (c) decrease the rate of endorsement of a correct solution under biasing conditions, and (d) increase the rate of endorsement of an incorrect solution under biasing conditions.

What kind of contingent task conditions might produce this forbidden sign vector? Well, recall from §2 that (+ – – +) would require a contingent task condition to have the following effects on the underlying structural parameters: (a) an increase in the sensitivity of intervening reasoning (x_2), (b) an increase in the intervening response criterion (c_2), (c) a decrease in the sensitivity of default reasoning (x_1), and (d) a decrease in the default response criterion (c_1). In other words, the onus is on dual-process theorists to find changes between contingent task conditions that (a) increase the ability for reasoning to discriminate correct from incorrect responses and decrease its bias to endorse when metacognitive heuristics are improved and (b) decrease the power for reasoning to discriminate correct from incorrect responses and increase the bias to endorse when metacognitive heuristics are impaired. By comparison, single-process theorists need only refute claims that such vectors have been found.[20]

If it is possible to find contingent task conditions that produce sufficiently many and large forbidden instances of the forbidden sign vector (+ – – +) to compensate for the risk of measurement error, then we can use Stephens et al.'s (2018) SDA to directly refute SPT and confirm DPT. Otherwise, if it is impossible to do so, then we can use their SDA to instead directly confirm SPT and refute DPT. As a result, the burden of proof is on DPT. This shouldn't be surprising, though. Both single- and dual-process theorists agree that reasoning constitutes a type of processing, but dual-process theorists add that this type of processing consists of two further sub-types of processing (T1 and T2). Since dual-process theorists propose that we include two sub-

[20] While it would be ideal for me to make a concrete suggestion about what contingent task conditions might produce this forbidden sign vector, I myself am inclined to SPT, so I'm skeptical that (+ – – +) can be found in any debiasing study of the kind that I've described.

types of processing to our *cognitive ontology* (Poldrack et al., 2011) in addition to the overarching type of reasoning, the principle of parsimony puts the onus on them to justify the additional ontological commitments of DPT, not on single-process theorists to justify the rejection of those ontological commitments.

§6. Conclusion

Should we add T1 and T2 reasoning to our cognitive ontology in addition to the overarching type of reasoning? I've argued that answering this question requires a carefully coordinated strategy among theorists, modellers, and experimenters. First, theorists must use theoretical considerations to develop a coherent formulation of DPT. In this paper, I argued that there are two constraints on a coherent formulation of DPT: the defining properties of T1 and T2 reasoning must be (a) duals of each other and (b) discrete. Since popular formulations of DPT fail to satisfy these conditions, I proposed a new formulation of DPT: T1 and T2 reasoning are partly defined as reasoning that precedes and follows metacognitive control. This formulation of DPT makes a coherent and plausible claim about cognitive ontology.

Second, experimenters, modellers, and theorists must coordinate to design relevant experiments. In the past, problematic formulations of DPT were used to try to isolate and dissociate default and intervening reasoning. The failure of these experiments to find evidence for DPT (Stephens et al., 2018, 2019) could be due to the truth of SPT or to the failure of these formulations to isolate and dissociate T1 and T2 reasoning. Our formulation of DPT enables us to more effectively isolate default and intervening reasoning. We can use suitable biasing conditions to selectively impair metacognitive heuristics (e.g., processing fluency, conflict), which decreases the probability of metacognitive control and isolates default reasoning. Likewise, we can use suitable debiasing conditions to selectively improve metacognitive heuristics, which increases the probability of metacognitive control and isolates intervening reasoning. This ensures that when SDA decides whether a type-difference in processing is necessary to explain the response rates, that this type-difference must be between default and intervening reasoning—not between different default responses to different tasks (or the same tasks under different instructions).

Third, modellers can then use formal analysis to determine whether models with one or two types of processes are best for explaining response rates in these experiments. In the past, dual-process theorists have used linear analysis to misinterpret response rate data as evidence for DPT. In this paper, I've reviewed arguments from the cognitive modelling literature that monotonic analysis is more appropriate for explaining response rates. I've focused on Stephens et al.'s (2018) innovative approach to monotonic analysis, which combines (a) signed difference analysis (Dunn & James, 2003; Dunn & Anderson, 2017), (b) signal detection theory (Macmillan & Creelman, 2005), (c) conjoint monotonic regression (Kalish et al., 2016), and (d) bootstrapping approximations of conjoint monotonic regression. If we modify their approach to explain response rates in conditions that more effectively isolate default and intervening reasoning, then this monotonic analysis will directly adjudicate the debate between SPT and DPT.

Fourth, finally, we proceed from modelling back to theory. The single or dual types of reasoning may only be partly defined by their *extrinsic relations* to metacognitive control—as preceding, following, or being modulated by it. Our new formulation of DPT is consistent with the possibility that they are also partly defined by some of their *intrinsic properties*. In fact, SDA would provide one way to do this: it would suggest the cognitive algorithms that are used to generate responses within a given experimental context. We could then generalize across experimental contexts to determine the general kinds of cognitive algorithms that reasoning functions to use. Finally, we could revise our definitions of the types of reasoning so that they are also partly defined by their functions to implement these general kinds of cognitive algorithms.[21] From there, of course, we can start a new cycle—using our new, more elaborate theories to guide new, more elaborate experiments, which we can model to test our new more elaborate theories, and so on.

Acknowledgments: I thank Valerie Thompson for her incisive comments: they have led this paper in a completely different direction than I thought it could go, and I'm very pleased with where it has ended up. I thank Wim De Neys for a formative discussion about this debate in his reply to my commentary on his 2021 paper in *Perspectives on Psychological Science*: it was instrumental in the development of the experimental strategy that I've proposed here. Finally, I thank Reviewer 3, Sara Aronowitz, and Sahana Srivathsa for their helpful feedback on the various drafts of this paper.

[21] In doing this, we'd be approaching a level of understanding about reasoning that is predicted to be possible by the computational theory of mind (e.g., Marr, 1982; Pylyshyn, 1984; Piccinini & Craver, 2011).

References

Ackerman, R., & Thompson, V. A. (2017). Meta-reasoning: Monitoring and control of thinking and reasoning. *Trends in Cognitive Sciences, 21(8)*, 607–617. https://doi.org/10.1016/j.tics.2017.05.004

Andersson, L., Eriksson, J., Stillesjö, S., Juslin, P., Nyberg, L., & Wirebring, L. K. (2020). Neurocognitive processes underlying heuristic and normative probability judgments. *Cognition*, *196*, 104153. https://doi.org/10.1016/j.cognition.2019.104153

Bago, B., & De Neys, W. (2017). Fast logic?: Examining the time course assumption of dual process theory. *Cognition*, *158*, 90–109. https://doi.org/10.1016/j.cognition.2016.10.014

Bago, B., & De Neys, W. (2019). The Smart System 1: Evidence for the intuitive nature of correct responding on the bat-and-ball problem. *Thinking & Reasoning*, *25*(3), 257–299. https://doi.org/10.1080/13546783.2018.1507949

Bago, B., & De Neys, W. (2020). Advancing the specification of dual process models of higher cognition: A critical test of the hybrid model view. *Thinking & Reasoning*, *26*(1), 1–30. https://doi.org/10.1080/13546783.2018.1552194

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

Bonner, C., & Newell, B. R. (2010). In conflict with ourselves? An investigation of heuristic and analytic processes in decision making. *Memory & Cognition*, *38*(2), 186–196. https://doi.org/10.3758/MC.38.2.186

Botvinick, M. M., & Cohen, J. D. (2014). The computational and neural basis of cognitive control: Charted territory and new frontiers. *Cognitive Science*, *38*(6), 1249–1285. https://doi.org/10.1111/cogs.12126

Bourgeois-Gironde, S., & Vanderhenst, J.-B. (2009). How to open the door to System 2: Debiasing the bat and ball problem. In S. Watanabe, A. P. Bloisdell, L. Huber, & A. Young (Eds.), *Rational animals, irrational humans* (pp. 235–252). Keio University Press.

De Neys, W. (2012). Bias and conflict: A case for logical intuitions. *Perspectives on Psychological Science*, *7*(1), 28–38. https://doi.org/10.1177/1745691611429354

De Neys, W. (2014). Conflict detection, dual processes, and logical intuitions: Some clarifications. *Thinking & Reasoning*, *20*(2), 169–187. https://doi.org/10.1080/13546783.2013.854725

De Neys, W. (2017). *Dual Process Theory 2.0*. Routledge. https://doi.org/10.4324/9781315204550

De Neys, W. (2021). On dual- and one-process models of thinking. *Perspectives on Psychological Science*, 1745691620964172. https://doi.org/10.1177/1745691620964172

De Neys, W., & Pennycook, G. (2019). Logic, fast and slow: Advances in dual-process theorizing. *Current Directions in Psychological Science*, *28*(5), 503–509. https://doi.org/10.1177/0963721419855658

De Neys, W., Schaeken, W., & d'Ydewalle, G. (2005). Working memory and everyday conditional reasoning: Retrieval and inhibition of stored counterexamples. *Thinking & Reasoning, 11*, 349–381. http://dx.doi.org/10.1080/13546780442000222

De Neys, W., Vartanian, O., & Goel, V. (2008). Smarter than we think: When our brains detect that we are biased. *Psychological Science*, *19*(5), 483–489. https://doi.org/10.1111/j.1467-9280.2008.02113.x

Dewey, A. R. (2021). Reframing single- and dual-process theories as cognitive models: Commentary on De Neys (2021). *Perspectives on Psychological Science, 16*(6): 1428–31. https://doi.org/10.1177/1745691621997115

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

Dunn, J. C., & Anderson, L. (2018). Signed difference analysis: Testing for structure under monotonicity. *Journal of Mathematical Psychology*, *85*, 36–54. https://doi.org/10.1016/j.jmp.2018.07.002

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

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

Elqayam, S., & Evans, J. S. B. T. (2011). Subtracting "ought" from "is": Descriptivism versus normativism in the study of human thinking. *Behavioral and Brain Sciences*, *34*(5), 233–248. https://doi.org/10.1017/S0140525X1100001X

Evans, J. St. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, *59*(1), 255–278. https://doi.org/10.1146/annurev.psych.59.103006.093629

Evans, J. S. B. T. (2010). *Thinking twice: Two minds in one brain.* Oxford University Press.

Evans, J. S. B. T. (2017). Dual-process theory: Perspectives and problems. In W. De Neys (Ed.), *Dual Process Theory 2.0* (pp. 137–55). Routledge. https://doi.org/10.4324/9781315204550

Evans, J. S. B. T., & Curtis-Holmes, J. (2005). Rapid responding increases belief bias: Evidence for the dual-process theory of reasoning. *Thinking & Reasoning*, *11*(4), 382–389. https://doi.org/10.1080/13546780542000005

Evans, J. St. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, *8*(3), 223–241. https://doi.org/10.1177/1745691612460685

Gangemi, A., Bourgeois-Gironde, S., & Mancini, F. (2015). Feelings of error in reasoning—In search of a phenomenon. *Thinking & Reasoning*, *21*(4), 383–396. https://doi.org/10.1080/13546783.2014.980755

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, *1*(1), 107–143. https://doi.org/10.1111/j.1756-8765.2008.01006.x

Handley, S. J., & Trippas, D. (2015). Dual processes and the interplay between knowledge and structure: A new parallel processing model. In B. H. Ross (Ed.), *Psychology of Learning and Motivation* (Vol. 62, pp. 33–58). Academic Press. https://doi.org/10.1016/bs.plm.2014.09.002

Heit, E., & Rotello, C. M. (2010). Relations between inductive reasoning and deductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 805–812. http://dx.doi.org/10.1037/a0018784

Hirt, E. R., & Markman, K. D. (1995). Multiple explanation: A consider-an-alternative strategy for debiasing judgments. *Journal of Personality and Social Psychology, 69*(6), 1069–1086. https://doi.org/10.1037/0022-3514.69.6.1069

Hogarth, R. M., & Soyer, E. (2011). Sequentially simulated outcomes: Kind experience versus nontransparent description. *Journal of Experimental Psychology: General, 140*(3), 434–463. https://doi.org/10.1037/a0023265

Hoover, J. D., & Healy, A. F. (2017). Algebraic reasoning and bat-and-ball problem variants: Solving isomorphic algebra first facilitates problem solving later. *Psychonomic Bulletin & Review*, *24*(6), 1922–1928. https://doi.org/10.3758/s13423-017-1241-8

Hoover, J. D., & Healy, A. F. (2021). The bat-and-ball problem: A word-problem debiasing approach. *Thinking & Reasoning*. https://doi.org/10.1080/13546783.2021.1878473

Kahneman, D., & Frederick, S. (2005). A Model of Heuristic Judgment. In Holyoak, K. & Morrison, R.G. (eds.) *The Cambridge handbook of thinking and reasoning* (pp. 267–293). Cambridge University Press.

Kahneman, D. (2011). *Thinking, fast and slow.* New York: Farrar, Straus & Giroux.

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

Keren, G. (2013). A tale of two systems: A scientific advance or a theoretical stone soup? Commentary on Evans & Stanovich (2013). *Perspectives on Psychological Science*, *8*(3), 257–262. https://doi.org/10.1177/1745691613483474

Keren, G., & Schul, Y. (2009). Two is not always better than one: A critical evaluation of two-system theories. *Perspectives on Psychological Science*, *4*(6), 533–550. https://doi.org/10.1111/j.1745-6924.2009.01164.x

Klauer, K. C., Beller, S., & Hütter, M. (2010). Conditional reasoning in context: A dual-source model of probabilistic inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 298–323. http://dx.doi.org/10.1037/a0018705

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

Kruglanski, A. W., & Gigerenzer, G. (2011). Intuitive and deliberate judgments are based on common principles. *Psychological Review*, *118*(1), 97–109. https://doi.org/10.1037/a0020762

Kruglanski, A. W. (2013). Only one? The default interventionist perspective as a unimodel—Commentary on Evans & Stanovich (2013). *Perspectives on Psychological Science*, *8*(3), 242–247. https://doi.org/10.1177/1745691613483477

Larrick, R. P., Morgan, J. N., & Nisbett, R. E. (1990). Teaching the use of cost-benefit reasoning in everyday life. *Psychological Science, 1*(6), 362–370. https://doi.org/10.1111/j.1467-9280.1990.tb00243.x

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

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*(1), 19–40. https://doi.org/10.1037/1082-989X.7.1.19

Marr, D. (1982). *Vision: A computational investigation into the human reformat and processing of visual information.* San Francisco: W. H. Freeman and Company.

Mastrogiorgio, A., & Petracca, E. (2014). Numerals as triggers of system 1 and system 2 in the 'bat and ball' problem. *Mind & Society*, *13*(1), 135–148. https://doi.org/10.1007/s11299-014-0138-8

Mata, A. (2020). An easy fix for reasoning errors: Attention capturers improve reasoning performance. *Quarterly Journal of Experimental Psychology*, *73*(10), 1695–1702. https://doi.org/10.1177/1747021820931499

Melnikoff, D. E., & Bargh, J. A. (2018). The mythical number two. *Trends in Cognitive Sciences*, *22*(4), 280–293. https://doi.org/10.1016/j.tics.2018.02.001

Mevel, K., Borst, G., Poirel, N., Simon, G., Orliac, F., Etard, O., Houdé, O., & De Neys, W. (2019). Developmental frontal brain activation differences in overcoming heuristic bias. *Cortex*, *117*, 111–121. https://doi.org/10.1016/j.cortex.2019.03.004

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

Newstead, S. E. (2000). Are there two different types of thinking? *Behavioral and Brain Sciences*, *23*(5), 690–691. https://doi.org/10.1017/S0140525X0049343X

Oldrati, V., Patricelli, J., Colombo, B., & Antonietti, A. (2016). The role of dorsolateral prefrontal cortex in inhibition mechanism: A study on cognitive reflection test and similar tasks through neuromodulation. *Neuropsychologia*, *91*, 499–508. https://doi.org/10.1016/j.neuropsychologia.2016.09.010

Osman, M. (2004). An evaluation of dual-process theories of reasoning. *Psychonomic Bulletin & Review*, *11*(6), 988–1010. https://doi.org/10.3758/BF03196730

Osman, M. (2013). A case study: Dual-process theories of higher cognition—Commentary on Evans & Stanovich (2013). *Perspectives on Psychological Science*, *8*(3), 248–252. https://doi.org/10.1177/1745691613483475

Poldrack, R. A., Kittur, A., Kalar, D., Miller, E., Seppa, C., Gil, Y., Parker, D. S., Sabb, F. W., & Bilder, R. M. (2011). The cognitive atlas: Toward a knowledge foundation for cognitive neuroscience. *Frontiers in Neuroinformatics*, *5*. https://doi.org/10.3389/fninf.2011.00017

Pennycook, G. (2017). A perspective on the theoretical foundation of dual process models. In W. De Neys (Ed.), *Dual Process Theory 2.0* (pp. 5–27). Routledge. https://doi.org/10.4324/9781315204550

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015). What makes us think? A three-stage dual-process model of analytic engagement. *Cognitive Psychology*, *80*, 34–72. https://doi.org/10.1016/j.cogpsych.2015.05.001

Pennycook, G., Neys, W. D., Evans, J. St. B. T., Stanovich, K. E., & Thompson, V. A. (2018). The mythical dual-process typology. *Trends in Cognitive Sciences*, *22*(8), 667–668. https://doi.org/10.1016/j.tics.2018.04.008

Piccinini, G., & Craver, C. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese*, *183*(3), 283–311. https://doi.org/10.1007/s11229-011-9898-4

Pylyshyn, Z. (1984). *Computation and cognition: Toward a foundation for cognitive science.* Cambridge, MA: MIT Press.

Rips, L. J. (2001). Two kinds of reasoning. *Psychological Science*, *12*(2), 129–134. https://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

Samuels, R. (2009). The magical number two, plus or minus: Dual process theory as a theory of cognitive kinds. In K. Frankish & J. St B. T. Evans (eds.) *Two minds: Dual processes and beyond* (pp. 129–146). OUP.

Shynkaruk, J. M., & Thompson, V. A. (2006). Confidence and accuracy in deductive reasoning. *Memory & Cognition*, *34*(3), 619–632. https://doi.org/10.3758/BF03193584

Simon, G., Lubin, A., Houdé, O., & Neys, W. D. (2015). Anterior cingulate cortex and intuitive bias detection during number conservation. *Cognitive Neuroscience*, *6*(4), 158–168. https://doi.org/10.1080/17588928.2015.1036847

Singmann, H., & Klauer, K. C. (2011). Deductive and inductive conditional inferences: Two modes of reasoning. *Thinking & Reasoning*, *17*(3), 247–281. https://doi.org/10.1080/13546783.2011.572718

Sio, U. N., & Ormerod, T. C. (2009). Does incubation enhance problem solving? A meta-analytic review. *Psychological Bulletin, 135*(1), 94–120. https://doi.org/10.1037/a0014212

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*(1), 3–22. https://doi.org/10.1037/0033-2909.119.1.3

Smith, E. R., & DeCoster, J. (1999). Associative and rule-based processing: A connectionist interpretation of two-process models. In S. Chaiken (ed.) *Dual-process theories in social psychology* (pp. 323–336). The Guilford Press.

Stanovich, K. E. (2004). *The robot's rebellion: Finding meaning in the age of Darwin.* Chicago, IL: University of Chicago Press.

Stanovich, K. E. (2011). *Rationality and the reflective mind.* New York: Oxford University Press.

Stephens, R. G., Dunn, J. C., & Hayes, B. K. (2018). Are there two processes in reasoning? The dimensionality of inductive and deductive inferences. *Psychological Review*, *125*(2), 218–244. https://doi.org/10.1037/rev0000088

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

Teuber, H. L. (1955). Physiological psychology. *Annual Review of Psychology*, *6*, 267–296.

Thompson, V. A. (2009). Dual-process theories: A metacognitive perspective. In J. S. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 171–195). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199230167.003.0008

Thompson, V. A., & Morsanyi, K. (2012). Analytic thinking: Do you feel like it? *Mind & Society*, *11*(1), 93–105. https://doi.org/10.1007/s11299-012-0100-6

Thompson, V. A., & Johnson, S.C. (2014) Conflict, metacognition, and analytic thinking. *Thinking & Reasoning*, 20, 215–244. https://doi.org/10.1080/13546783.2013.869763

Thompson, V. A., Prowse Turner, J. A., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology*, *63*(3), 107–140. https://doi.org/10.1016/j.cogpsych.2011.06.001

Thompson, V. A., Turner, J. A. P., Pennycook, G., Ball, L. J., Brack, H., Ophir, Y., & Ackerman, R. (2013). The role of answer fluency and perceptual fluency as metacognitive cues for initiating analytic thinking. *Cognition*, *128*(2), 237–251. https://doi.org/10.1016/j.cognition.2012.09.012

Vartanian, O., Beatty, E. L., Smith, I., Blackler, K., Lam, Q., Forbes, S., & De Neys, W. (2018). The reflective mind: Examining individual differences in susceptibility to base rate neglect with fMRI. *Journal of Cognitive Neuroscience*, *30*(7), 1011–1022. https://doi.org/10.1162/jocn_a_01264

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