Journal of Mathematical Psychology 90 (2019) 31–46

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

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

Confidence and varieties of bias

Andrew Heathcote [*], Eleanor Holloway, James Sauer

Department of Psychology, University of Tasmania, Australia

HIGHLIGHTS

- We test whether response bias has two different bases reflecting beliefs and utilities of choices.
- We predict accuracy/response time/confidence from Vickers (1979) balance-of-evidence theory.
- We test these predictions with ANOVA analyses and Bamber's (1979) state–trace analysis.
- The latter analyses use Prince et al. (2012) and Davis-Stober, et al.'s (2016) Bayes factors.
- Predictions were supported at the aggregate level, but support was more equivocal for individuals.
- We discuss why individual results are preferred over aggregate results in state–trace analysis. We draw lessons about obtaining clear results from individual state–trace analyses.

ARTICLE INFO

Article history:
Received 13 June 2018
Received in revised form 20 August 2018
Available online 29 October 2018

ABSTRACT

We test the proposition that response bias can have two different bases; reflecting either differing beliefs about the a priori likelihood of competing response alternatives, or their relative utilities. In evidence accumulation models, these two types of bias are thought to manifest as variations in the starting point for accumulation and threshold for responding, respectively. Although these two mechanisms are indistinguishable for linear accumulators in terms of accuracy and RT, Vickers' (1979) balance-of-evidence hypothesis predicts they have dissociable effects on confidence. We derived ten ordinal predictions from these models and confirmed them at the level of group averages using traditional ANOVA analyses of results from a new experiment that manipulated the probability of correct responses and the rewards associated with them. However, individual effects were more variable, particularly with respect to the reward manipulation. We then used Bamber's (1979) state–trace analysis to test the predicted dissociations using Bayes factors developed by Prince, Brown and Heathcote (2012) and Davis-Stober, Morey and Heathcote (2016). Once again, we found support at the aggregate level but more equivocal results for individuals. We discuss why individual results are to be preferred over aggregate results in state–trace analysis and draw the lesson that tailored designs are needed to obtain clear results from individual state–trace analyses.

In this paper we investigate the proposition that bias in decision making can have two very different bases. One basis, related to beliefs about inequalities in the prior probability of different outcomes, is grounded in a normative Bayesian approach to belief updating. The other basis, related to making choices that result in good consequence or minimize bad consequences, is grounded in normative approaches to utility maximization in theories of economic decision making. In both cases, the result is a tendency to favour some decisions over others, so the two bases can be hard to identify from simply observing what decisions are made. One potential way to discern the two is to elicit a judgement about decision confidence. If, for example, one option is chosen over another because it offers more reward, then a lower confidence judgement may signal an underlying belief that the rejected option is actually more likely, but a decision was made counter to that belief in order to maximize expected gains. Here we seek to use this potential dissociation between confidence judgements and decision making to assess the evidence for the two different varieties of bias associated with manipulations of the prior probability of two response options and of the rewards associated with each.

Evidence for a dissociation based on a simple interaction between the effects of two manipulations is notoriously unreliable (Dunn & Kirsner, 1988; Loftus, 1978), particularly for bounded dependent measures such as choice probability and confidence ratings, because of the possibility of a nonlinear mapping between latent states and manifest behaviours (e.g., floor and ceiling effects). Hence, we rely on state–trace analysis (Bamber, 1979) to provide

[*] Corresponding author.
E-mail address: andrew.heathcote@utas.edu.au (A. Heathcote).

https://doi.org/10.1016/j.jmp.2018.10.002
0022-2496/© 2018 Elsevier Inc. All rights reserved.

rigorous tests of whether the effects of our two manipulations can be explained by a single underlying latent variable, or whether more than one is required, supporting a cognitive basis for two distinct types of bias. State–trace analysis is based on assessing whether a plot of one dependent variable against another is strictly monotonic (i.e., always non-increasing or always non-decreasing), which indicates one latent variable is sufficient to explain the observed behaviour, or whether the plot is non-monotonic, which indicates more than one latent variable is required. In our context, we might expect that non-monotonic state–trace plots constructed from dependent variables affected by proportion and reward manipulations could provide evidence that there are two different bases for bias.

For inference, we rely on an encompassing-prior Bayes-factor method (Klugkist, Laudy, & Hoijtink, 2005), as applied to state–trace analysis by Davis-Stober, Morey, Gretton, and Heathcote (2016) and Prince, Brown, and Heathcote (2012), to analyse individual participants' state–trace results. We use Bayes factors because they are not only able to provide evidence favouring our hypothesis (non-monotonicity), but also evidence for the null-hypothesis (i.e., monotonicity), avoiding the criticism of frequentist null-hypothesis statistical testing that it is "violently biased against the null hypothesis" (Berger & Delampady, 1987, p. 330). We focus on individual-participant results because of the long-recognized problem – exemplified in the Condorcet voting paradox (de Condorcet & de Marquis, 1785, cited in Gehrlein, 2002; see also Gehrlein, 1983[alp]) – that averaging can distort ordinal relationships, and, by definition, monotonicity and non-monotonicity are about orderings. Prince et al. provided examples where the average of individually monotonic state–trace plots is non-monotonic, and vice versa, suggesting state–trace analysis of average data is potentially misleading.

In an Appendix to this paper, we review Vickers' (1979) balance-of-evidence hypothesis, which provides an explicit theoretical evidence-accumulation-model framework for understanding the joint effects of our two bias manipulations on accuracy, confidence and response time (RT). In these models, prior probability is thought to affect the starting-point of evidence accumulation, and reward to affect the threshold amount of evidence required to trigger a choice (Vickers & Lee, 1998). In linear accumulator models the two types of bias do not have an identifiably different effect on accuracy or RT. That is, the effect of a change of one type of bias on these measures can always be perfectly mimicked by an appropriate change in the other type of bias. However, their effects can be differentiated by confidence as determined by the "balance of evidence"; the difference between the amount of evidence in the winning vs. losing accumulator at the moment of choice. The only empirical evidence that we are aware of which bears on whether the predicted dissociations hold is from a number of unpublished experiments, whose results are summarized in a book chapter by Vickers (1985). Hence, we report the results of a new experiment, with accompanying analyses testing a set of ten ordinal predictions (derived from the theoretical framework developed in the Appendix, and summarized below) about confidence, accuracy and RT. We then apply state–trace analysis, first to the accuracy data, for which our theoretical framework predicts a monotonic effect, and to confidence data and to the joint effects of accuracy and confidence, where a non-monotonic effect is predicted. We end by discussing the psychological implications of our findings, and the strengths and weaknesses of our methodological approach to state–trace analysis.

1. Experiment

The experiment used a perceptual binary decision task, requiring participants to indicate the majority colour (blue or orange) in a checkerboard stimulus. Vickers and Lee (1998) argued that the perceived consequences of a response and a priori expectations about likely accuracy of a response produce dissociable effects on decision-making and confidence. However, we can find no empirical demonstrations of this dissociation. Thus, this experiment tests this claim. Bias was manipulated via rewards (influencing the utility of responses and hence, according to Vickers & Lee, 1998, thresholds) and via informing participants about the proportion of stimuli of each type (i.e., majority blue or orange and, according to Vickers and Lee, influencing beliefs prior to making a decision). Each correct response earned points, with differences in points for correct majority blue and orange used to manipulate rewards. Prior probability was manipulated through the proportion of trials in each block with majority blue and orange stimuli. Participants completed two sessions, and bias type was manipulated within subjects between sessions, with both types of bias favouring the same colour within an individual to avoid confusion about the bias direction.

Within each block of trials half of the stimuli were easy and half hard, where difficulty was manipulated through the strength of the majority (54% or 52%, for easy and hard trials, respectively). Using Prince et al.'s (2012) terminology, difficulty was included as a "trace" factor in the state–trace analysis. The state–trace analysis plots performance in two bias-type conditions against each other (i.e., bias type forms a "state" factor). The plot has separate lines for the bias-for condition (i.e., the bias favours the stimulus and hence the correct response) and bias-against condition (i.e., disfavouring the stimulus and hence the correct response), where each line joins the two levels of the trace factor (easy vs. hard). Bias direction constitutes the "dimension" factor whose joint action in the two bias-type conditions must be explained by changes in either a single underlying psychological ("latent") dimension, or by changes on more than one such dimension.

Non-monotonicity among all points in the state–trace plot (i.e., an inability to join all points with a single always non-decreasing or always non-increasing line) indicates that a single latent variable cannot explain the data. The trace factor is chosen based on an a priori expectation that it has a monotonic effect for both levels of the dimension factor which, in the case of difficulty, is a decrease in accuracy and confidence and an increase in RT. If this is the case, then any non-monotonicity in the plot can be unambiguously attributed to the effect of the dimension factor. Ideally, the trace factor should also have a comparable effect to the dimension factor in terms of magnitude in order to maximize the chances of detecting non-monotonicity, through causing the lines for each level of the dimensions factor to overlap on at least one axis. Without overlap, a state–trace plot must be monotonic even when the underlying system generating behaviour is controlled by more than one latent variable (see Prince et al., 2012).

As discussed, accumulator models account for response bias through effects on threshold placement, or the start-point for evidence accumulation (see Appendix for details). However, an alternative (or perhaps additional) possibility exists. The effects of bias on confidence may operate via a post-decision, metacognitive mechanism. Thus, participants might complete the evidence accumulation process in the usual manner but, upon reaching threshold, engage in a post-decision, time-consuming process where their confidence in the selected response is interrogated according to either a priori expectations about the likelihood of the available response options, or the perceived consequences of the available response options. We included unbiased blocks (with equal rewards and proportions) to test this possibility. Demonstrating equal or faster responding in bias blocks would make this account less likely, unless this additional metacognitive mechanism was accompanied by a much faster decision process, which should then be detectable as a marked decrease in accuracy in bias blocks.

A problem with evaluating the bias effect relative to an unbiased condition is that participants may have different levels of caution (i.e., the overall level of thresholds for both accumulators) between biased and unbiased conditions. This is plausible because they must know the nature of the bias condition (i.e., unbiased vs. biased towards a particular response) before the stimulus appears. This problem can be avoided by making the comparison between stimuli that are favoured and disfavoured by the bias. Because the stimulus is not known prior to the trial, thresholds cannot be differentially adjusted between these two conditions. Hence, apart from the check described in the last paragraph, our analyses focused on the results from the biased conditions so as to avoid being confounded by caution differences.

2. Method

2.1. Participants

32 participants underwent both bias manipulations in two, separate, 1-h sessions. Participants were aged 18 or over and were recruited from either the first-year psychology student participation pool at the University of Tasmania, Sandy Bay, or through posters placed around campus. Psychology students received 2 h course credit for participating, and other participants' expenses in attending were compensated with a $20 gift voucher. All participants provided informed consent. They were tested on separate computers in groups of no more than four.

2.2. Procedure

The task stimulus comprised a 32 × 32 grid of either majority blue or orange squares, 20 × 20 pixels in size displayed on a 24-inch computer screen. Participants were asked to indicate if there were more blue squares or more orange squares. The positions of blue and orange squares were randomized and updated 20 times per second. Decision difficulty was manipulated through the majority colour, either 52% (hard) or 54% (easy) of squares. Differential rewards were manipulated through points gained by a correct response, 300 points for orange and 100 for blue, or vice versa. Prior probability was manipulated by having either 75% majority blue stimuli in a block of trials or 75% majority orange stimuli. In the probability-manipulation blocks, each correct response received 200 points; in the reward blocks orange and blue stimuli occurred equally often, and in unbiased blocks both proportions and rewards were equal.

Sessions were at least one day apart and consisted of 8 blocks of 40 trials. Within each session, an equal number of biased and unbiased blocks were presented in a counterbalanced order, with proportion and reward manipulations in separate sessions, again counterbalanced over participants. Before each session, participants completed one unbiased and one biased practice block. For each trial, participants had four response options: high confidence that there are more blue squares, low confidence that there are more blue squares, low confidence that there are more orange squares, and high confidence that there are more orange squares. Each response was made by pressing the "s", "d", "j", and "k" keys, where "s" and "k" indicated high confidence and "d" and "j" indicated low confidence. Key pairs associated with each colour (either "s" and "d" or "j" and "k") were counterbalanced over participants, and during each trial coloured squares reminding participants of the colour-to-key mapping were present in the lower corners of the screen.

Before participants completed the task, instructions were view ed on the computer screen, which told them what the task involved, what keys to respond with to make their chosen response, and to rest their fingers on the keys during each block. Participants were instructed to press the spacebar when they had finished reading the instructions and were ready to complete the practice trials. After each trial, if participants responded with the correct colour, the speed in which they responded appeared on the screen for 500 ms, and they received the relevant number of points. The total points participants received remained at the top of the screen throughout the task. If participants responded incorrectly, the word 'incorrect' appeared on the screen. If participants responded either too slowly (over 2 s) or too quickly (less than 200 ms), "too slow", or "too fast" appeared on the screen; all of these messages appeared for 500 ms. Timing feedback was provided to discourage fast guessing and distracted responding. After the relevant message appeared there was a brief blank screen, followed by the next trial. At the end of a block of trials, the number of points earned, accuracy as a percentage, and mean RT were displayed. Participants were instructed to use this information to guide subsequent performance and asked to press the spacebar when they were ready to begin the next block. Participants were made aware of the reward and proportion manipulations before each block, for example: "During this block, you will receive 200 points each time you respond with the correct colour. During this block, 75% of the trials will have mostly blue displays. It is more likely that an individual trial will be mostly blue. Try to use this information to help you make the correct response".

3. Results

All descriptive analyses were conducted with the *lme4* R package (Bates, Maechler, Bolker, & Walker, 2014), using a Gaussian error model for the logarithm of RT and a binomial probit model for binary data.[1] ANOVA inferences were made via Wald χ^2 tests with type III sums of squares as implemented by the *car* package (Fox, Friendly, & Weisberg, 2013). Participant data was initially screened for fast guessing, although this is mainly associated with conditions unlike those in the present experiment, responding under strong speed emphasis and with more extreme bias manipulations (Noorbaloochi, Sharon, & McClelland, 2015). Most participants had no responses less than 0.2 s, with all but one having less than 0.3% and one having 3.4%. All such responses were removed (0.16% overall) from further analysis.

An initial analysis of RT and accuracy compared unbiased and biased conditions to see if there was evidence of an extra metacognitive judgement, and therefore longer RTs and possibly reduced accuracy, in the bias blocks. Contrary to the presence of an extra post-decisional metacognitive judgement, mean RT in the unbiased condition (0.737 s) was slower than in the reward (0.729 s), $\chi^2(1) = 10$, $p = .002$, and proportion (0.681 s), $\chi^2(1) = 685$, $p <$ .001, conditions, and accuracy was less in the unbiased condition (83.4%) than the reward (85.2%), $\chi^2(1) = 18.8$ $p < .001$, and proportion (86.2%), $\chi^2(1) = 44.7$, $p < .001$.

We next checked on the speed of error vs. correct responses in the unbiased condition. Mean RT for errors (0.766 s) was slower than for correct responses (0.732 s), $\chi^2(1) = 103, p < .001$, consistent with errors being mainly being driven by rate variability. This finding supports investigation of hypotheses based the latter assumption. All further analyses exclude the unbiased condition. Note that, overall, errors (0.756 s) were slower than corrects (0.699 s) in the biased conditions, $\chi^2(1) = 137$, $p < .001$, and there was no evidence this differed between reward (0.768 s vs. 0.723 s) and proportion (0.747 s vs. 0.67 s) bias, $\chi^2(1) = 0.43$,

[1] Note that our use of frequentist analysis here was motivated by the fact that Bayes factors for generalized linear model ANOVAs implemented in *lme4* are not yet readily available, and approximate approaches, such analyses of summary proportions for each individual transformed in an attempt to conform to a Gaussian error model, throw away information and are of questionable accuracy.

Table 1
Results of testing 10 predictions. The "Larger condition" column reports bias condition that is predicted to produce a larger value on the dependent variable, with the percentage of participants conforming to the prediction in brackets. The right adjacent column gives the exact binomial probability for one tailed test of that account. Accuracy and confidence differences ("Larger–Smaller) are reported as percentages for accuracy and confidence and in seconds for RT. The right adjacent columns report χ^2 values testing the difference (which all have one degree of freedom) with corresponding one-tailed p values.

	Reward bias					Proportion bias				
	Larger condition	p	Larger–smaller	χ^2	p	Larger condition	p	Larger–smaller	χ^2	p
Accuracy	For (65)	.08	3.5	26.4	<.001	For (88)	<.001	13.1	287	<.001
Correct Mean RT	Against (65)	.08	0.035	68.3	<.001	Against (96)	<.001	0.106	615	<.001
Error Mean RT	For (62)	.16	0.058	2.36	.06	For (92)	<.001	0.163	188	<.001
Correct Confidence	Against (46)	.72	6.2	21.8	<.001	For (69)	.038	15.7	237	<.001
Error Confidence	For (38)	.91	9.6	0.17	.34	Against (77)	.002	19.2	54.6	<.001

$p = .51$, supporting the applicability of accumulator models to the bias data. In accumulator models, the distance from start point to threshold, and hence RT, is smallest for the favoured response, which is the correct response for a favoured stimulus and the error response for the disfavoured stimulus. A significant three-way interaction found among response accuracy, bias type and bias direction, $\chi^2(1) = 240, p < .001$, is consistent with this prediction, with errors slower than corrects for stimuli favoured by the bias (reward: 0.788 s vs. 0.705 s, proportion: 0.821 s vs. 0.649 s) but almost as fast or faster for disfavoured stimuli (reward: 0.753 s vs. 0.742 s, proportion: 0.650 s vs. 0.744 s). These preliminary analyses support the applicability of the ten predictions we derived to the present data.

We now report tests of ten predictions about the order of accuracy, RT and confidence results in the bias-for vs. bias-against conditions. Predictions for bias for vs. bias against differences in accuracy are straightforward: accuracy will be greater in the bias-for case. For correct RT the opposite ordering holds, RT will be less in the bias-for case because the distance from start-point to threshold is less. For error RTs the ordering is the opposite (i.e., the same as for accuracy) because the bias is against the correct response but for the error response, and so the distance from start-point to threshold is less in the bias-against case. These predictions for confidence are derived in the Appendix based on Vickers' (1979) balance-of-evidence hypothesis for accumulator models (these predictions are followed by the number of the relevant equation in the Appendix):

(1) Accuracy will be greater in the bias-for condition under reward bias;
(2) Accuracy will also greater in the bias-for condition under proportion bias;
(3) Mean RT for correct responses will be greater in the bias-against condition under reward bias;
(4) Mean RT for correct responses will also be greater in the bias-against condition under proportion bias;
(5) Mean RT for error responses will be greater in the bias-for condition under reward bias;
(6) Mean RT for error responses is also greater in the bias-for condition under proportion bias;
(7) Confidence in correct responses with reward bias is greater in the bias-against condition (A.10);
(8) Confidence in correct responses with proportion bias will be greater in the bias-for condition (A.11);
(9) Confidence in error responses with reward bias will be greater in the bias-for condition (A.12);
(10) Confidence in error responses with proportion bias will be greater in the bias-for condition (A.13).

Before testing these 10 predictions, we screened participants for their use of confidence ratings. All participants gave a high confidence rating for one or more response but three participants in the reward condition never gave a low confidence rating, two different participants did the same in the proportion condition, and one participant did so in both. These participants were eliminated from all further analyses to avoid ceiling effects in confidence analyses, leaving a remaining sample of 26. Confidence was expressed in terms of the proportion of high confidence responses and analysed as a binary variable in the same way as accuracy.

Table 1 provides the directions of each of the 10 predicted effects, corresponding differences, and test statistics. On average every prediction was fulfilled, but the effects were substantially larger for the proportion than reward bias manipulation, by a factor of around three for accuracy and RT, and a factor of around two for confidence. Hence, while all effects were significant for the proportion manipulation, two failed to reach significance for the reward manipulation. These two were the error-related effects; as average accuracy was 85%, and so error responses were more than five times less common than correct responses, greater measurement error, and hence lesser power, would be expected for the error analyses. None of the interactions between these effects and difficulty approached significance, so there was no evidence that the effects were inconsistent for easy and hard choices.

There were substantial individual differences, particularly in confidence, and particularly under the reward manipulation. For the stronger proportion manipulation, the majority of participants displayed effects in the predicted direction at a level significant by a binomial test. This was not the case for the weaker reward manipulation, where the majority showed the predicted effect for accuracy and RT, but near equality or a minority held for the confidence measures, and no binomial test achieved significance. This pattern was the same at both levels of difficulty. Fig. 1 plots the overall effects for each participant sorted by magnitude. For the proportion manipulation the predicted effects are quite consistent, although less so for confidence, whereas for the reward manipulation they are relatively consistent only for RT.

Difficulty was included as a "trace" factor with the expectation that accuracy and confidence would decrease with difficulty and RT would increase, and that these effects would apply in both bias-for and bias-against conditions. Consistent with the latter expectation, no interactions between difficulty and bias direction approached significance for either the reward condition — accuracy: $\chi^2(1) = 0.49, p = .48$, correct confidence: $\chi^2(1) = 0.35, p = .55$, error confidence: $\chi^2(1) = 0.56, p = .44$, correct RT: $\chi^2(1) = 1.12, p = .29$, and error RT: $\chi^2(1) = 1.46, p = .23$ – or the proportion condition – accuracy: $\chi^2(1) = 0.06, p = .81$, correct confidence: $\chi^2(1) = 1.06, p = .3$, error confidence: $\chi^2(1) = 0.09, p = .76$, correct RT: $\chi^2(1) = 0.86, p = .35$, and error RT: $\chi^2(1) = .04, p = 84$. There were also strong main effects of the trace factor in the expected directions for all but error confidence, for both the reward condition — accuracy: $\chi^2(1) = 417, p < .001$, correct confidence: $\chi^2(1) = 20.8, p < .001$, error confidence: $\chi^2(1) = 0.04, p = .85$, correct RT: $\chi^2(1) = 161, p < .001$, and error RT: $\chi^2(1) = 7.3, p = .007$ – or the proportion condition – accuracy: $\chi^2(1) = 250, p < .001$, correct confidence: $\chi^2(1) = 25.1, p < .001$, error confidence: $\chi^2(1) = 0.97, p = .33$, correct RT: $\chi^2(1) = 192, p < .001$, and error RT: $\chi^2(1) = 7.1, p = .008$.

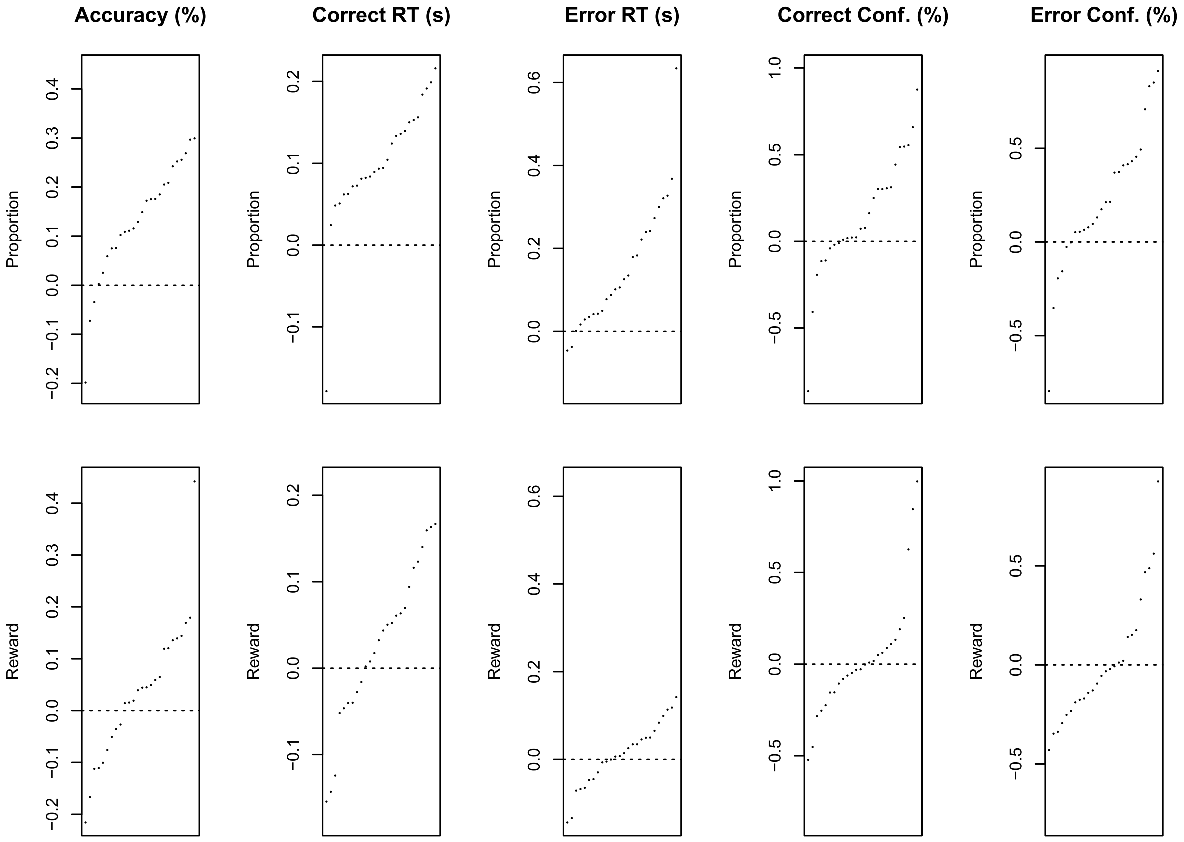
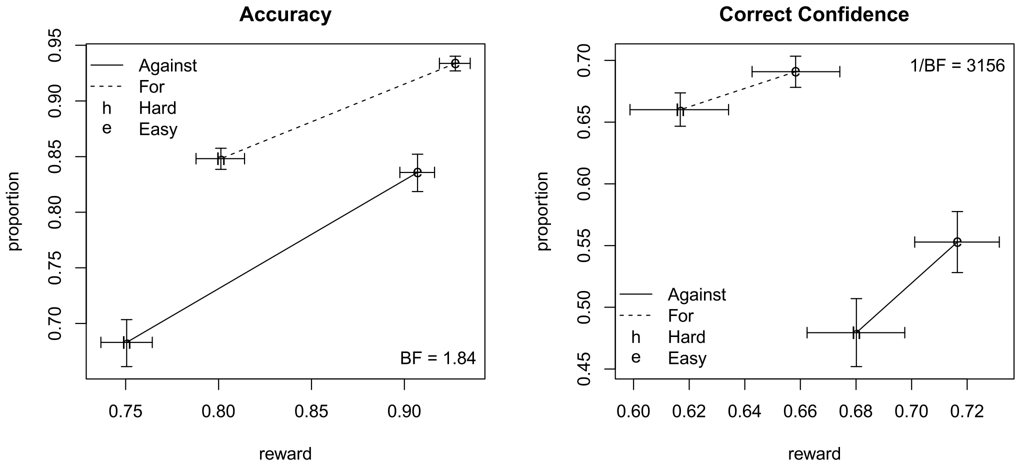


Fig. 1. Magnitudes of predicted effects (see Table 1 for definitions) for each of 26 participants, sorted separately in each panel by magnitude. Conf. = confidence.



(a) (b)

Fig. 2. Aggregated state–trace plot with 95% credible intervals based on 10,000 samples from a probit model with parameters estimated from data under a uniform prior for (a) accuracy and (b) confidence. The BF in the panels is the $\mathrm{BF}_{\mathrm{M.MN}}$ defined in Eq. (1).

4. State–trace analysis

We restricted our state–trace analysis to accuracy and confidence for correct responses, as four subjects had no error response in one or more cell of the 2 × 2 × 2 state–trace design, and other participants had very few errors, particularly in the higher accuracy (easy, bias-for) conditions, resulting in excessive measurement noise in confidence for error responses. Fig. 2 provides state–trace plots created by aggregating data over participants (i.e., based on total counts of either correct or high-confidence responses). Error bars are 95% credible intervals for the aggregate based on the binomial probit error model that we assume for all analyses (Davis-Stober et al., 2016). In Fig. 2(a), accuracy in the proportion and reward conditions constitute the state factor, bias direction (for vs. against) the dimension factor, and difficulty (hard vs. easy) the trace factor. For this case, a monotonic state–trace is predicted by the linear racing-accumulator framework. Fig. 2(b) has the same format, but it plots the probability of making a high-confidence response. For this case balance-of-evidence hypothesis predicts a non-monotonic state–trace.

We predict that the state–trace plot for accuracy will be monotonic because in linear accumulator models the two types of bias do not have an identifiably different effect on accuracy (as any change in start point can be perfectly mimicked by a change in threshold). Fig. 2(a) suggests that in aggregate the effect on accuracy is close to monotonic (following the increasing order over against-hard, for-hard, against-easy and for-easy conditions), with only a minor inconsistency between the hard-for and easy-against conditions (i.e., the former is clearly smaller than the latter in the reward condition, but very marginally greater in the proportion condition). To test for monotonicity vs. non-monotonicity without potential confounding due to averaging we calculated Bayes factors at the individual level. This computation required us to classify posterior outcomes (e.g., samples from the posterior) according to their conformity with various order restrictions. One restriction specific to the accuracy analysis is that the sample has chance or better accuracy (i.e., only samples with accuracy greater than 50% are retained for further analysis). The other restrictions, which are generic to the analyses of all state–trace results, enable calculation of posterior probabilities that, when divided by corresponding prior probabilities, constitute Bayes factors. Prior probabilities can be calculated analytically (see Prince et al., 2012, for details) based on the assumption that all orders are equally likely, but calculation of posterior probabilities requires numerical methods; we used Davis-Stober et al.'s (2016) Laplace approximation to the posterior distribution on a probit scale, which is essentially instantaneous to compute.[2]

For example, suppose Po(M) is the probability that posterior samples are monotonic (e.g., have the same order over the four trace × dimension factor conditions in both the reward and proportion conditions), Po(NM) = 1 −Po(M) is the probability of non-monotonic samples, and the corresponding prior probabilities are Pr(M) and Pr(NM), then the monotonic vs. non-monotonic Bayes factor is:

$$BF_{M.NM} = \frac{Po\,(M)\,/Pr\,(M)}{Po\,(NM)\,/Pr\,(NM)}. \qquad (1)$$

The numerator quantifies the degree to which monotonic samples are more common in the posterior than the prior, and similarly for the denominator with respect to non-monotonic samples. Division by the prior takes account of the fact that a non-monotonic result is much more likely by chance, as there are many more ways to be non-monotonic than monotonic. Multiplying the Bayes factor by the prior odds of each model produces posterior odds (i.e., the relative probability of each model after observing the data). Fig. 2(a) reports this Bayes factor for accuracy data obtained by aggregating responses over participants. It indicates that belief in a monotonic over non-monotonic model is changed by a factor of 1.84 by observing the aggregated data. If monotonic and non-monotonic models were considered equally likely a priori, the posterior odds in this case would be 1.84, and so the posterior probability for the monotonic model is 1.84/(1 + 1.84) ∼ .65, indicating the evidence is equivocal.

As the implications of aggregate results are uncertain for inferences about individual cognition, we repeated the analysis at the individual level. Fig. 3(a) shows on the abscissa the base 10 logarithm of the Bayes factors in (1) for each participant (indicated on the ordinate by a letter and sorted by the magnitude). Values above zero indicate the data increases belief in the monotonic model and those below zero in the non-monotonic model. Horizontal lines indicate a classification of belief change into equivocal (Bayes factors between 1/3 and 3, corresponding to $\log_{10}$ values of −0.477 and 0.477), positive (1/10–1/3 and 3–10), strong (1/100–1/10 and 10–100) and very strong (less than 1/100 or greater than 100). These discrete classifications simply aid description of the underlying Bayes factors, which provide a graded scale of belief change, and

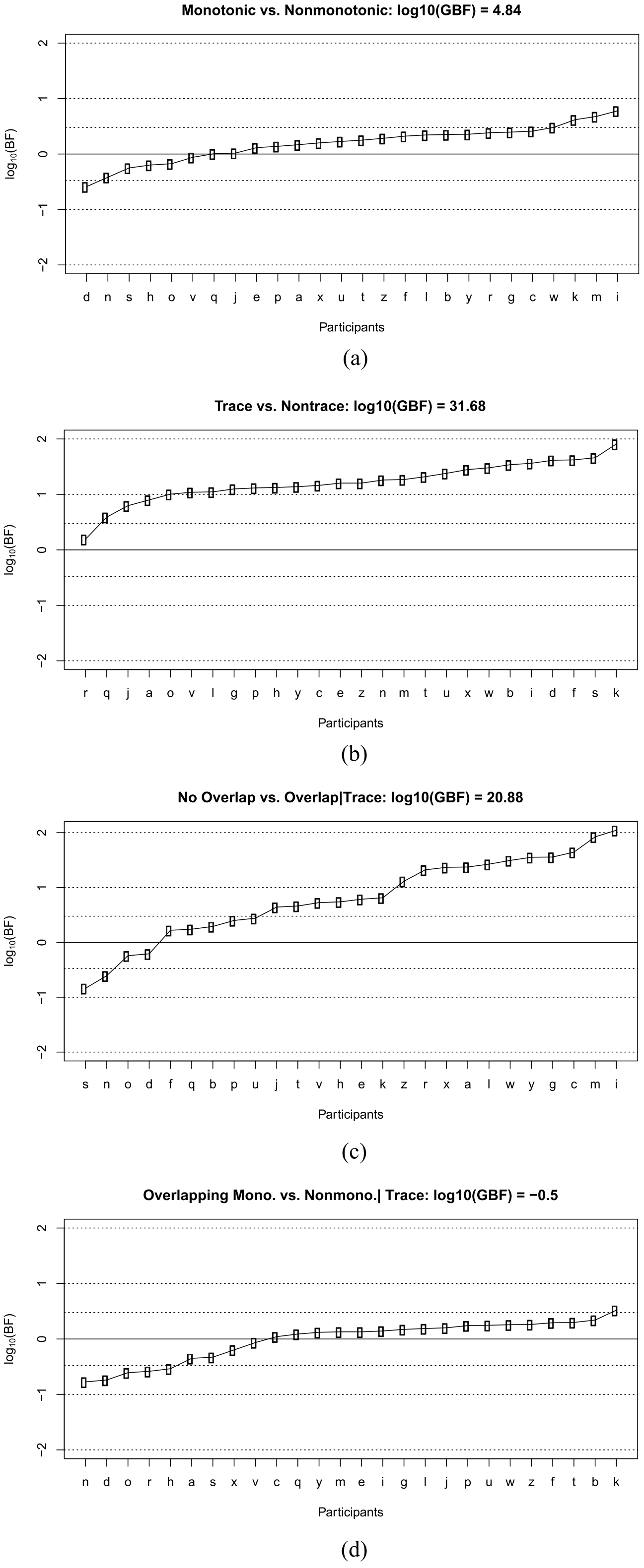


Fig. 3. Individual and group Bayes factors for the accuracy state–trace analysis.

[2] We checked the approximation with Monte-Carlo methods using 5 × 10^7 posterior samples per participant, and we found it to be accurate. To speed the check – as it took many CPU hours for each participant – we augmented Davis-Stober et al.'s (2016) software to allow the Monte-Carlo function to compute in parallel over participants. We also found and corrected a bug in this software for a case that was not used in the original paper, but which was used here, where no trace restriction is applied (i.e., neglecting monotonic cases where the order on one axis is the exact opposite of the order on the other axis).

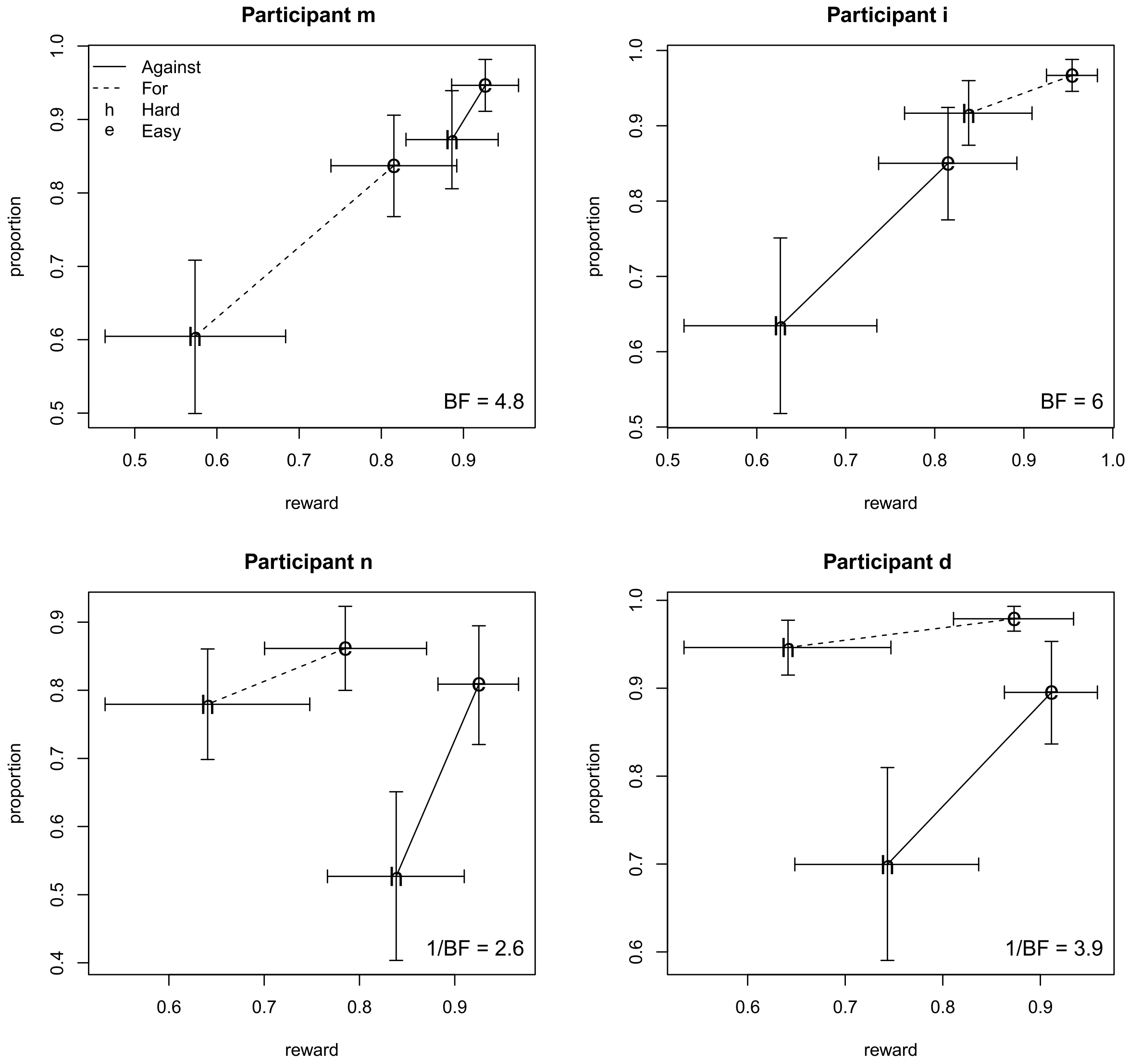


Fig. 4. Individual state–trace accuracy plots with 95% credible intervals for two participants with the strongest evidence for monotonicity in Fig. 3(a) (top row) and the two with the strongest evidence for non-monotonicity (bottom row). The BF in the panels is the $BF_{M,MN}$ defined in Eq. (1).

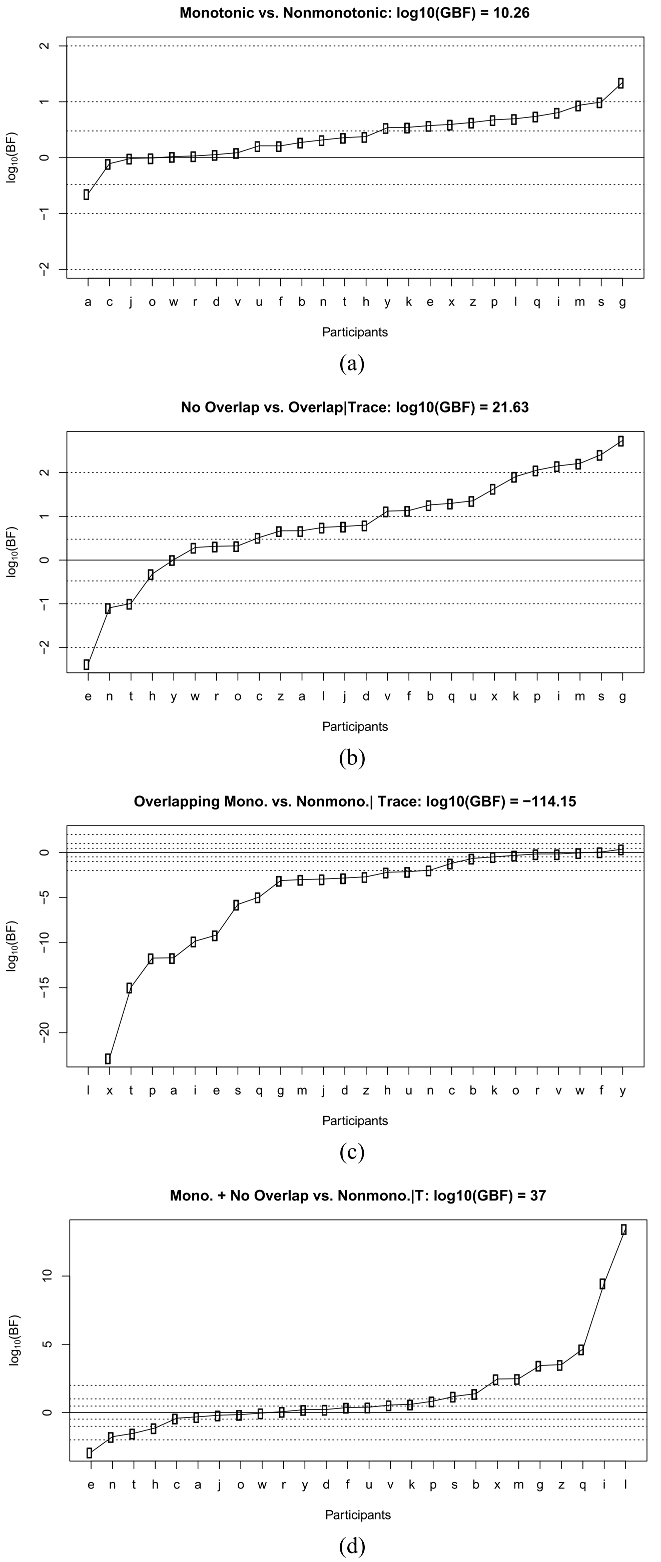
correspond (under an equal prior) to posterior probabilities over ranges approximately bounded by 0.75, 0.9 and 0.99.

Fig. 3(a) shows that all but five of the 26 individual level Bayes factors are equivocal, with four providing positive support for monotonicity and one for non-monotonicity. The title above Fig. 3(a) provides the "Group Bayes Factor" (GBF), which is the product of the individual Bayes factors (equivalently, the logarithm of the group Bayes factor is the sum of individual Bayes factor logarithms). Consistent with Bayes factors favouring monotonicity for majority of participants (20/26), the group Bayes factor very strongly favours monotonicity, by a factor of $\sim 10^{4.84}$. The group Bayes factor tests the hypothesis that all participants are monotonic vs. all participants are non-monotonic, which might be considered to be of limited utility if it is plausible that there may be heterogeneity among participants. In this case, their being only mild positive evidence for non-monotonicity for one individual participant suggests that homogeneity is plausible. Fig. 4 plots the two participants with the strongest evidence for monotonicity (top row) and the two with the strongest evidence for non-monotonicity (bottom row).

Prince et al. (2012) advocated using the trace order restriction – in the current context that accuracy decreases with difficulty – to exclude measurement noise and so sharpen the analysis. They provided a Bayes factor to test the trace model, which is shown for each participant in Fig. 3(b): These results favour the trace model for most participants, with only one equivocal case, three positive and the rest strong. However, Fig. 3(c) shows, again using a Bayes factor developed by Prince et al., that the traces do not, in the main, overlap (e.g., in the top row of Fig. 4 the lines joining points do not overlap on either axis). Prince et al. argued that in the absence of overlap, inferences about monotonicity vs. non-monotonicity are ambiguous. They provided a Bayes factor which excludes this ambiguous (non-overlapping) evidence and, as might be expected, its values displayed in Fig. 3(d) show almost entirely equivocal individual results, as well as an equivocal group Bayes factor.

4.1. Confidence analysis

Fig. 2(b) shows very strong support for non-monotonicity in an aggregate state–trace plot of confidence, consistent with the predictions of the balance-of-evidence hypothesis. However, as shown in Fig. 5(a), this was not the case at the individual level, with only one participant having positive support for non-monotonicity, whereas ten had positive support for monotonicity and one strong support, with the rest being equivocal. As a consequence, the GBF provided very strong support for monotonicity.



(a)

(b)

(c)

(d)

Fig. 5. Individual and group Bayes factors for the confidence state–trace analysis.

Fig. 6 displays the two participants with the strongest support for monotonicity (top row) and non-monotonicity (bottom row). It is evident that the putatively monotonic participants have very non-overlapping traces, and Fig. 5(c) shows this is commonly the case, with overlap supported for only four participants. As a result, inference conditional on the trace model – which as might be expected holds quite well for this data (GBF = 13.4, with support in all but 4 participants, and no strong results against the trace model) – depends strongly on whether non-overlapping samples are excluded (Fig. 5(c)) or included (Fig. 5(d)) as supporting monotonicity. Davis-Stober et al. (2016) and Prince et al. (2012) suggested they be excluded in order to provide a more stringent test of monotonicity. However, in their application any lack of overlap was minor, and monotonicity generally favoured. In the present data the lack of overlap means there is little support for the monotonic overlapping model, so it is dominated by the non-monotonic model. If on the other hand, non-overlapping samples are also taken to support monotonicity the result reverses, but as shown in Fig. 7, which plots the four participants for whom monotonicity received the strongest support in Fig. 5(d), it is clear that this support is spurious, and the reality is that lack of overlap makes any adjudication on monotonicity reliant on a large degree of extrapolation, and hence manifestly unreliable.

4.2. *Joint confidence–accuracy analysis*

In order to avoid the problems associated with non-overlap, we performed a state–trace analysis on the joint effects of accuracy and confidence. The balance-of-evidence hypothesis predicts a dissociation between the joint effects of bias type and bias direction on accuracy and confidence and, hence, a non-monotonic state–trace plot. Note that the most commonly applied model of bias – signal detection theory – predicts a monotonic state–trace plot because it has only one mechanism to set bias, a shift in its decision criteria along the evidence axis. For example, a uniform shift in decision criteria to lower values on the evidence axis (i.e., a bias favouring the option supported by higher values the decision axis) results in both higher accuracy and confidence for the favoured over the disfavoured option.

Fig. 8 depicts aggregate state–trace plots for easy and hard data separately, with confidence and accuracy constituting the state factor, bias type (reward vs. proportion) the dimension factor, and bias direction the trace factor. For both hard and easy conditions these plots appear non-monotonic, which would support the predictions of the balance-of-evidence hypothesis over those of signal-detection theory. Note that in these cases the designation to trace and dimension roles is arbitrary, as a monotonic effect that is consistent for both confidence and accuracy is not expected for either bias type or direction. Hence, analysis is restricted to the Bayes factor in Eq. (1) with no conditioning on a trace model.[3]

Fig. 8 shows that traces do overlap in the aggregate, and this was also true at the individual level (easy GBF = −140, hard GBF = −148), with only one participant with a Bayes factor favouring no overlap, and almost all other cases showing strong support for overlap. Fig. 8 also shows that at the aggregate level there is a strong dissociation between the joint effects of bias type and bias direction on accuracy vs. confidence, with Bayes factors that strongly favour non-monotonicity. Once again, we checked the aggregate pattern at the individual level in order to avoid potential averaging distortion and found indeterminate results (Fig. 9). Most participants had equivocal Bayes factors, and group Bayes factors differed in favouring non-monotonicity for the easy condition and monotonicity for the hard condition. Fig. 10 plots individual results for two participants (“i” and “x”) who consistently had among the highest evidence for monotonicity in the easy and hard conditions, and two participants (“t” and “d”) who had the highest evidence for non-monotonicity in both easy and hard. Good overlap is evident in all cases with only the participant with the highest evidence for non-monotonicity (“d”) evincing a pattern similar to the aggregate.

[3] Tests like those shown in Fig. 4(b) strongly rejected the Trace model, with $\log_{10}(\text{GBF}) = -87.1$ and -92 for easy and hard respectively. The same general pattern held when bias direction was the trace factor, with $\log_{10}(\text{GBF}) = -23.6$ and -26.5 respectively.

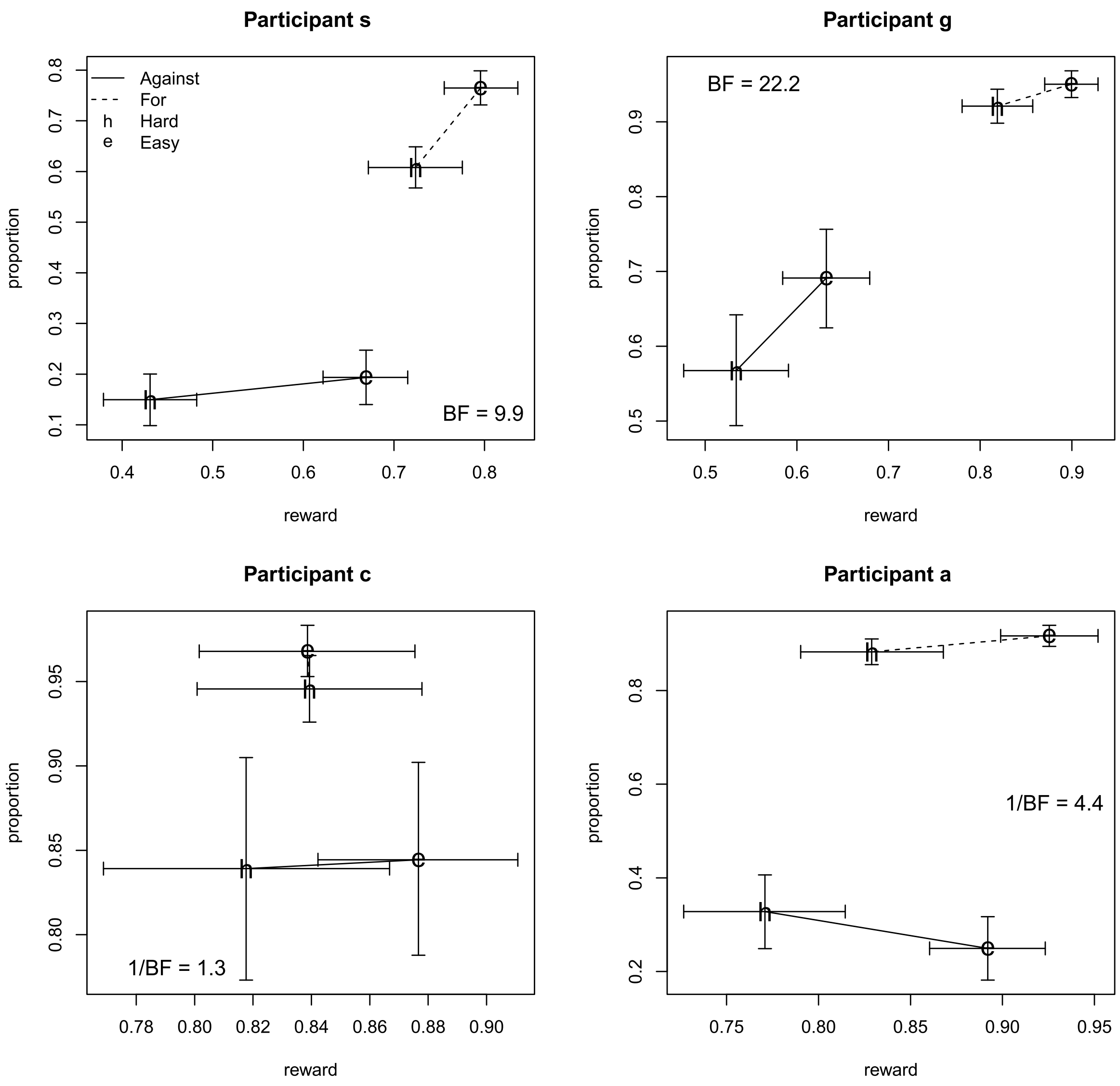


Fig. 6. Individual state–trace accuracy plots with 95% credible intervals for two participants with the strongest evidence for monotonicity in Fig. 5(a) (top row) and the two with the strongest evidence for non-monotonicity (bottom row). The BF in the panels is the $BF_{M,MN}$ defined in Eq. (1).

5. Discussion

We investigated whether there are two different bases for response bias as revealed by a divergence between decisions and by beliefs about those decisions as measured by confidence ratings. We derived from Vickers' (1979) balance-of-evidence hypothesis for linear evidence accumulation models ten ordinal predictions about effects on accuracy, confidence and RT of manipulations which and Vickers and Lee (1998) hypothesized selectively influence two types of bias: a manipulation of the *proportion* of trials favouring one or other binary response, and a manipulation of the *reward* favouring each response. We also included a no-bias condition as a baseline to check if participants enacted bias through appending to their decision process and extra time-consuming metacognitive judgement. We found equal or faster responding in bias compared to no-bias blocks along with greater accuracy, making any confounding from an extra metacognitive judgement very unlikely. A traditional ANOVA of average results found that all of the proportion-manipulation predictions were clearly confirmed. The results for the reward manipulations also generally favoured the predictions, but test results were less clear, due to a relatively weak effect of rewards that were subject to strong individual differences (for a similar finding see Mulder, Wagenmakers, Ratcliff, Boekel & Forstmann, 2012). Clearly it would be desirable to replicate these results with a more effective reward manipulation. However, given the substantial number of very specific predictions that were confirmed, these findings at least provide solid initial support for the proposition that there are two types of bias selectively influenced by reward and proportion manipulations that can be distinguished by a dissociation between effects on accuracy and RT and effects on confidence.

We then took a non-parametric approach to testing the predicted dissociation using state–trace analysis (Bamber, 1979). Sta te–trace analysis is based on a plot of results for one dependent variable against another. It is an attractive analysis method because it is relatively assumption free, requiring only that the mapping between latent variables determining behaviour and that observed or manifest behaviour are monotonic. Given this assumption, if a state–trace plot is monotonic it can be certainly inferred that only one latent variable is required to explain the observed data pattern. Conversely, if the state–trace plot is non-monotonic it can certainly be inferred that more than one latent variable is required. Note, however, that the latent variables must be at least partially independent and have identifiably different effects on the dependent variable used to quantify behaviour.

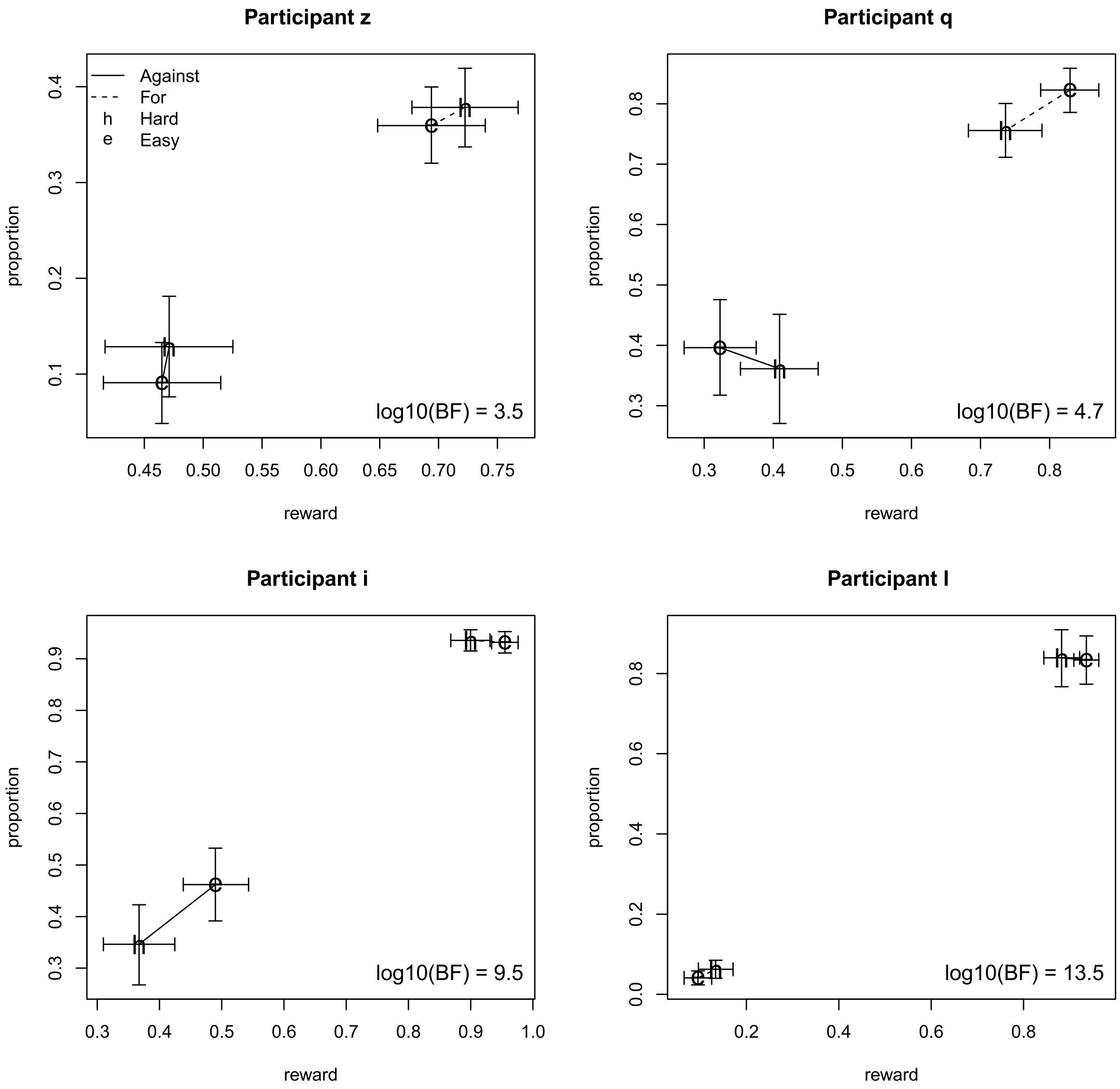
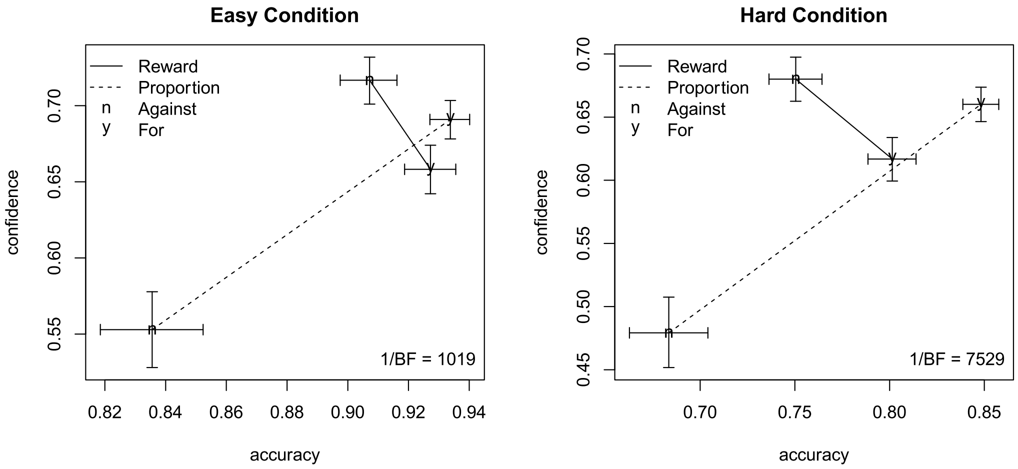


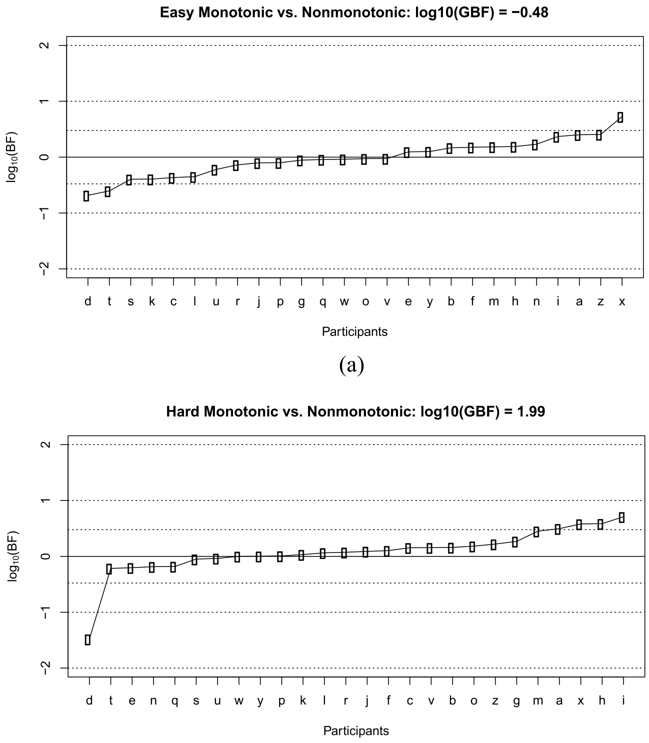
Fig. 7. Individual state–trace accuracy plots with 95% credible intervals for four participants with the strongest evidence for monotonicity in Fig. 5(d). The BF in the panels compare overlapping and non-overlapping monotonic samples to non-monotonic samples conditional on the trace model being true.



(a)

(b)

Fig. 8. Aggregated confidence vs. accuracy state–trace plot with 95% credible intervals based on 10,000 samples from a probit model with parameters estimated from data under a uniform prior for (a) easy and (b) hard conditions. The BF in the panels is the $BF_{M.MN}$ defined in Eq. (1).



(a)

(b)

Fig. 9. Individual and group Bayes factors for the accuracy vs. confidence state–trace analysis.

Formal inference about state–trace analysis requires an additional assumption about the nature of measurement error. We focused on binary data – accuracy as quantified by the probability of making a correct response and confidence as quantified by the probability of providing a high-confidence rating – and used Prince et al.'s (2012) Bayes factor approach to state–trace inference, which requires only the additional assumption that measurement error is binomial. Although we reported results aggregated over participants, we placed most weight in results for individuals because averaging can distort the ordinal relationships that state–trace analysis relies on. That is, from a monotonic state–trace plot of data aggregated over participants it cannot be certainly inferred that any participant has a monotonic state–trace plot (i.e., that their observed behaviour was generated by variation in a single latent variable). Similarly, it cannot be certainly inferred from a non-monotonic state–trace plot of aggregated data that any participant has a non-monotonic state–trace plot (i.e., that their observed behaviour was generated by variation in more than one latent variable).

Overall, our state–trace analysis was consistent with the prediction from linear evidence accumulation models that the two types of bias cannot be distinguished in accuracy data. Although a few participants did display some evidence of non-monotonicity (see Fig. 5, bottom row), individual Bayes factors provided barely positive support for non-monotonicity in only 1/26 cases, which might plausibility be attributed to measurement noise. If taken at face value, this finding of monotonicity underlines the fact that state–trace analysis cannot detect the action of two distinct latent variables (e.g., the starting point of evidence accumulation and evidence threshold required to trigger a response) if their effect on a manifest measure (e.g., accuracy) is mediated through a third common latent variable (e.g., the distance between starting-point and threshold).

However, we believe a conclusion in favour of a monotonic effect on accuracy requires qualification based on Prince et al.'s (2012) argument that, in the type of design we adopted, strong evidence for monotonicity requires "overlap". This type of design tests whether manipulation of a "dimension" factor – in our case the direction of the bias manipulation – has a monotonic vs. non-monotonic effect on the two dependent variables that define the axes of the state–trace plot – in our case accuracy in the reward vs. the proportion conditions – which together constitute a "state" factor. This design also includes a third "trace" factor that can be assumed to have a monotonic effect, here a manipulation of decision difficulty, where more difficult decisions are assumed to reduce accuracy in both the reward and proportion conditions. The trace factor is included to bring the dependent-variable measurements for the state and dimension levels into an overlapping range. This is evident graphically in a state–trace plot by an overlap on at least one axis of lines joining conditions with the same dimension-factor levels ("traces"). We did not find much support for overlap at the individual level, and when we excluded non-overlapping evidence for monotonicity, results favouring monotonicity became equivocal. Examination of individual state–trace plots revealed that the lack of overlap was not great, and so this issue could likely be addressed by a marginally stronger manipulation of difficulty.

We next applied state–trace analysis to the same design, but with confidence in the proportion vs. reward conditions constituting the state factor. Here the balance-of-evidence hypothesis predicts a non-monotonic plot, and indeed we found strong evidence for non-monotonicity at the aggregate level. However, at the individual level many findings were equivocal. A more extreme version of the problem with overlap also emerged: when we excluded non-overlapping evidence for monotonicity results for the majority of participants strongly favoured non-monotonicity, but they strongly favoured monotonicity when it included non-overlapping evidence. We were able to address the problem of overlap in further analyses that constituted the two levels of the state factor from accuracy vs. confidence. At the aggregate level, this produced strong support for non-monotonicity in separate state–trace plots for the easy and hard conditions. However, once again at the individual participant level results were equivocal. For some participants it is possible that equivocal results were due to a failure of the selective influence, so that the bias manipulations affected both starting points and thresholds, at least some trials. Overall, we draw the conclusion that the present experiment does not provide clear state–trace evidence either supporting or disconfirming the balance-of-evidence hypothesis or, more generally, that there are two bases for bias.

Although these state–trace analyses did not provide a basis to draw psychological conclusions, we believe they illustrate a series of important methodological lessons. First, we view the ability of a Bayes-factor based analysis to show that evidence is equivocal as a strength. In contrast, a non-significant frequentist result could indicate either that the null holds or that the evidence is equivocal. Although consideration of statistical power can shed some light on this dilemma in a quantitative sense, frequentism lacks the direct quantitative interpretability of the Bayes factor approach. Second, the present results underline for us the importance of taking into account overlap in a state–trace plot. For example, we believe it is clearly unreasonable to attempt to make any conclusions about monotonicity vs. non-monotonicity from highly non-overlapping plots, such as those shown in Fig. 8. Third, the ease with which we can construct Bayes factors, conditional on the trace model holding, that either exclude non-overlapping samples (as previously proposed by Prince et al., 2012) or include them (a new Bayes factor constructed here) illustrates the flexibility of the encompassing-prior approach to test a variety of ordinal hypotheses that are germane to state–trace analysis. Given that the Laplace approximation introduced by Davis-Stober et al. (2016) essentially solves

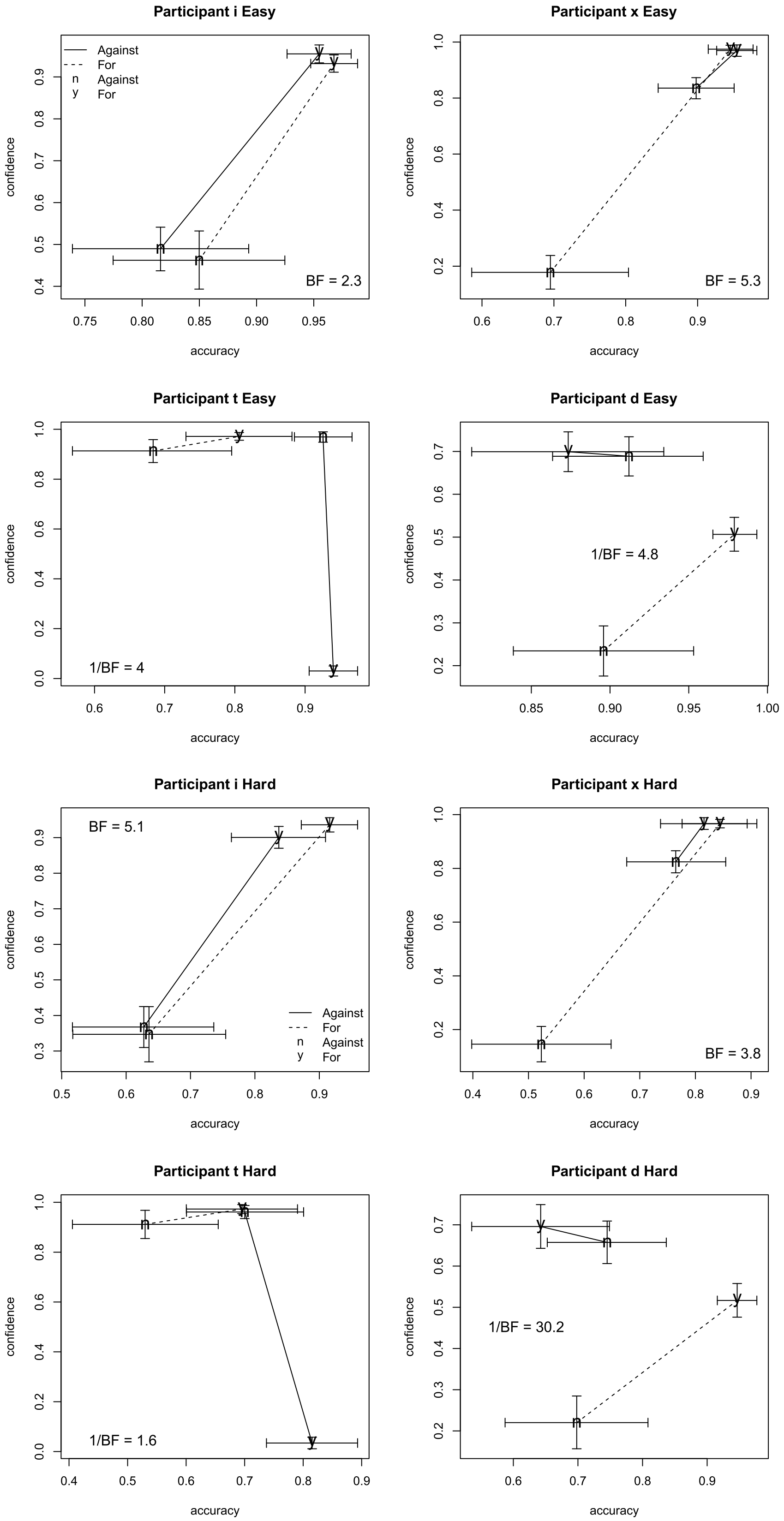


Fig. 10. Individual state–trace confidence vs. accuracy plots with 95% credible intervals for two participants with consistently high evidence for monotonicity in the easy (top row) and hard (third row) conditions, the two participants with the strongest evidence for non-monotonicity in both the easy (second row) and hard (bottom row) conditions. The BF in the panels is the $BF_{M.MN}$ defined in Eq. (1).

any computational bottleneck associated with this approach, and this approximation was shown to be accurate not only in their data, but also in the present case where each participant performed far fewer trials, there seems to us to be no practical impediment to the adoption of this approach. Still, we advise it is prudent to check the approximation with simulation methods, and for that purpose we augmented their software to take advantage of multiple cores and have made it, and scripts using it to perform all of the analyses presented here, freely available (osf.io/fzn9a).

A fourth lesson for us is that several stark contrasts between highly certain inferences based on the aggregate and much less certain inferences at the individual level make us cautious about relying on the former approach. This is not to say that we believe that the aggregate inferences are wrong. Indeed, they correspond very well to the predictions made by theories with strong converging empirical support, linear evidence accumulation models (e.g., Brown & Heathcote, 2008), and the balance-of-evidence hypothesis (e.g., Vickers, 2001). Certainly, there was also no evidence that *all* participants could have the opposite pattern to the aggregate, even though it may be possible to construct such examples (e.g. Prince et al., 2012, Fig. 5). However, the present results clearly illustrate cases where strong non-monotonicity at the aggregate level (e.g., Fig. 8) can occur when at least some underlying participant effects appear quite monotonic (e.g., participants "i" and "x" in Fig. 10). Less striking, and perhaps more easily attributed to measurement noise, apparently monotonic aggregate results (Fig. 3(a)) occurred when at least some participants contributing to the average appeared quite non-monotonic (participants "n" and "d" in Fig. 5).

Although the dangers of averaging are well known, not only in terms of ordinal hypotheses but also in terms of non-linear functional hypotheses (e.g., Brown & Heathcote, 2003; Estes, 1956; Heathcote, Brown, & Mewhort, 2000), this has to be balanced against the potentially beneficial effects of averaging with respect to ameliorating measurement noise (Cohen, Sanborn, & Shiffrin, 2008). However, we contend that if inferences are based on averages, particularly with respect to state–trace analysis which depends so crucially on joint orders, in each particular case either credible evidence must be presented that averaging is not misleading or claims about the strength of such inferences appropriately circumscribed. Because any such circumscription detracts from the most attractive feature of state–trace analysis – its ability to make strong conclusions based on minimal assumptions – we finish by discussing ways to avoid such limitations. Before doing so, however, we acknowledge that even individual analysis can be subject to averaging distortion to the degree that within an individual there is inhomogeneity in dimensionality over trials or over items where they are inhomogeneous (e.g., when using word stimuli). Our items were quite homogeneous, but it is possible that participants used inconsistent strategies across the course of the experiment (e.g., following our selective influence assumptions on only a subset of trials), in which case the results of our individual analysis would potentially be subject to averaging distortion. Unfortunately, this type of averaging effect is harder to address than averaging over participants, and of course the existence of this problem does not mitigate the potential issues associated with averaging over participants or in any way reduce the importance of attempting to addressing them.

In the case of non-linear functional hypotheses, hierarchical modelling (Shiffrin, Lee, Kim, & Wagenmakers, 2008) is one approach that can garner the benefits of averaging with respect to measurement noise while avoiding averaging distortion (e.g., Evans, Brown, Mewhort, & Heathcote, in press). However, we are not aware of any proof that this is the case for state–trace analysis. Indeed, it is clearly the case that even though monotonicity may hold for population parameters, at least some individuals sampled from that population may be non-monotonic, and similarly when non-monotonicity holds for population parameters some samples of individuals may be monotonic. In cases where interest focuses on the latent structure within individuals, as it clearly does here, a hierarchical approach seems to us of questionable utility. This led Davis-Stober et al. (2016) to develop a method that, instead of directly testing the monotonicity of the average, tests whether the average falls within the convex hull of all monotonic orders that also respect an ordering on the trace factor. If every participant is monotonic then the average must fall within the convex hull. For inference, they used the encompassing prior methodology based on a binomial error model to develop an "Aggregated Bayes factor" that can be estimated from the ratio of the proportion of samples from the posterior that fall within the convex hull in the posterior and prior. This approach takes advantage of the reduction in measurement noise due to averaging. We did not use this approach in the present case because it has not yet been extended to testing non-monotonicity and cases in which the trace order cannot be assumed. However, we believe such extensions hold some promise and should be pursued.

An alternative approach to avoiding equivocal individual results due to measurement noise is to collect better individual data. Recently, Smith and Little (in press) pointed out that many of the most reliable results in Psychology come from the psychophysical tradition, which focuses on identifying functional relationships that hold within an individual by precise and carefully calibrated measurement of a small set of observers with high precision (see also Kolossa & Kopp, 2018, for an example from cognitive neuroscience where individual data quality is paramount). This sort of approach focuses not only on collecting a large number of experimental trials per participant, but also on adjusting experimental conditions in order to focus measurement on areas of design space that provide high information gain (e.g., where the target function changes rapidly and on regions that yield information about particular aspects such as asymptotic performance). It seems likely that any state-analysis that seeks to characterize the dimensionality of latent structures within individual participants can benefit from a similar approach. In particular, the levels of dimension and trace factors can be calibrated on an individual basis to maximize overlap, potentially using a non-factorial design (see Prince et al., 2012 or further discussion) then a large number of experimental trials run for each participant. Although this approach might not be suited to all participant populations and research topics, it is well suited to tests of broad theoretical positions that propose latent structures common to all members of a population.

In closing, we have shown that the balance-of-evidence hypothesis makes clear predictions (i.e., predictions that are not confounded by model mimicry) about two type of bias and that these predictions are testable based on the selective influence of a proportion manipulation on one type of bias and a selective influence of a reward manipulation on the other type of bias. State–trace analysis can potentially provide an unambiguous test for the presence of two types of bias when applied to individual performance as long as the selective-influence assumptions hold consistently across trials. Our results here provide a promising start, but more needs to be done. In particular, a design with more trials, and with a better calibrated difficulty manipulation, will be required, as well as a way of increasing the effect of the reward manipulation. In future work we plan to purse these design improvements, including investigating the potential for a variable reward schedule, which is known to produce behaviours that are more rapidly acquired and frequently repeated (Ferster & Skinner, 1957), to produce a larger bias effect.

Acknowledgements

Andrew Heathcote acknowledges ARC Discovery Project DP160101891.

Appendix. Modelling bias

A.1. Racing accumulators

Accumulator models propose that choice options are represented by processes that sum a sequence of samples of evidence. The first process to accumulate a threshold amount of evidence triggers a response corresponding to the option it represents. Accumulator models can explain both the choice that is made and the time it takes to make that choice (i.e., the time taken to accrue the threshold amount of evidence in the winning accumulator). Fig. A.1 illustrates a linear deterministic accumulator model (Brown & Heathcote, 2008; Heathcote & Love, 2012) for a binary choice between left and right options. The thick dashed horizontal lines indicate decision thresholds (with the same value, b, for both accumulators). The thin slanted dashed lines indicate evidence accumulation totals, which are assumed to increase with time at a constant rate $v > 0$. In Fig. A.1(i), the starting point is the same for both accumulators, and the left rate, v_L, is greater than the right rate, v_R, so the left accumulator wins the race (i.e., the left option is chosen) at time t_1.

Bias towards a response increases the probability with which that response is made, so bias towards a correct response will increase accuracy, and a bias towards the wrong response will decrease accuracy. In accumulator models, bias towards a particular response can be induced either by lowering the threshold of the corresponding accumulator or by increasing the point at which it starts accumulating evidence relative to other accumulators. However, when accumulation is linear, as illustrated in Fig. A.1, shifts in the starting point of evidence accumulation and shifts of threshold amount of evidence required to make a response do not have identifiably different effects on RT and choice probability (i.e., one can always be perfectly mimicked by the other). The same lack of identifiability also occurs in stochastic accumulator models, where accumulation varies randomly from moment to moment. In all of these models, decision time is a decreasing function of the distance between the starting evidence total and the amount of evidence required to trigger a response. As both types of bias towards a response reduce this distance, they both decrease RT (i.e., the sum of decision time and the time for non-decision processes, such as stimulus encoding and response production) for the favoured response and increase the probability that it will win the race.

A.2. The balance of evidence

Vickers (1979) proposed an extension that allows accumulator models to account for confidence judgements. Confidence is proportional to the "balance of evidence", in Fig. A.1(i) the difference in the evidence total in the winning accumulator, $b = \mathrm{v}_L t_1$, and total in the losing accumulator at the same time, $\mathrm{v}_R t_1$: BoE $= b - \mathrm{v}_R t_1 = t_1(\mathrm{v}_L - \mathrm{v}_R)$. Confidence will be greater (i.e., the balance of evidence, or BoE, is larger) when the margin of the loss is greater (i.e., the loser has accrued little evidence). In contrast, confidence will be low if the winner beats the loser by only a narrow margin, and so the balance of evidence is small.

Vickers and Lee (1998) pointed out that start-point and threshold bias act in a different way on confidence judgements made using the balance of evidence. In particular, they note that in "...shifting the starting position towards threshold ... responses will also be made with increased confidence (since they benefit from a 'starting bonus', equivalent to the size of the shift in starting position) ... In contrast, where the threshold ... is shifted towards the starting position, the observer expects each stimulus to be equally likely, and there is no addition to the confidence evaluation ... Indeed, the reduction of the ... threshold will produce a reduction in the confidence with which ... responses are made" (p. 187). We demonstrate their point using a single trial of linear deterministic accumulator (see Vickers, 1979 for discussion of the stochastic case): Fig. A.1(ii) illustrates the effect of threshold bias and Fig. A.1(iii) start-point bias, where the bias always favours the left response. The left response always wins the race because its rate of accumulation, v_L, is greater than the rate for the right accumulator, $v_L > v_R$. Note that if the left response always wins in the unbiased case, then it must also be true in the biased case as the bias favours the left response.

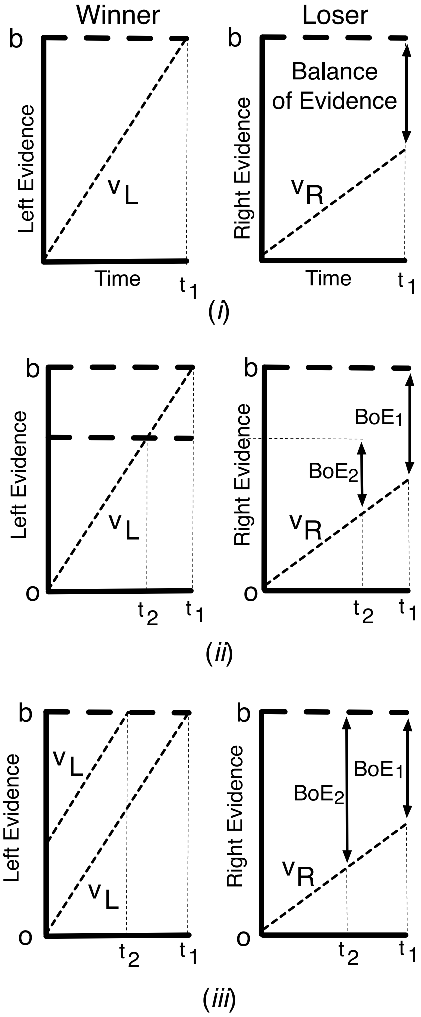


Fig. A.1. The balance-of-evidence hypothesis and bias effects in a binary linear deterministic accumulator model. Fig. A.1(i) shows the balance of evidence in a case where the participant is unbiased. Fig. A.1(ii) shows the effect of threshold bias (i.e., reduced threshold for L compared to R) on the balance of evidence. Fig. A.1(iii) shows the effect of start-point bias (i.e., higher start-point in the accumulator for L compared to R) on the balance of evidence.

Consider first the case of a left bias mediated by a reduced upper threshold for the left accumulator, which is illustrated in Fig. A.1(ii) along with the unbiased case for easy comparison. In the unbiased case, both accumulators start at zero and have the same threshold, b. A left decision happens at time $t_1 = b/v_L$. At t_1 the right accumulator is still below threshold (as its rate $v_R < v_L$), with evidence $v_R t_1 = b v_R / v_L$, and so the balance of evidence is:

$$BoE_U = b - b v_R / v_L = b(1 - v_R / v_L). \tag{A.1}$$

In the biased case, suppose the reduced left threshold is at pb (indicated by the lower thick dashed horizontal line in the left accumulator in Fig. A.1(ii)), where $p < 1$. A decision is made earlier at $t_2 = pb/v_L$, and so the evidence in the losing accumulator is reduced to $v_R t_1 = pbv_R/v_L$. Thus, the balance of evidence for the threshold bias case is:

$$BoE_{\mathrm{TB}} = pb(1 - v_{\mathrm{R}}/v_{\mathrm{L}}). \tag{A.2}$$

This is reduced relative to the unbiased case by a factor p (i.e., $BoE_{\mathrm{U}} > BoE_{\mathrm{TB}}$), with the unbiased case greater by:

$$BoE_{\mathrm{U}} - BoE_{\mathrm{TB}} = b(1 - p)(1 - v_{\mathrm{R}}/v_{\mathrm{L}}) > 0. \tag{A.3}$$

The reduction occurs because in the unbiased case the left accumulator has more time to accrue evidence than in the biased case, and because it does so at a faster rate than the right accumulator it gains a bigger advantage and hence a larger balance of evidence.

Next, consider the case of a left bias mediated by a higher start point for the left accumulator, illustrated in Fig. A.1(iii). The unbiased case – where left and right accumulators have the same threshold – is as before. In the biased case, the left accumulator starts higher (as indicated by the upper slanting line in the left accumulator in Fig. A.1(iii)), at a position $p \times b$, which is conveniently expressed as a proportion $p < 1$ of the threshold. A decision is now made earlier at $t_2 = pb/v_L$. Hence, the evidence in the losing accumulator is reduced to $v_R t_1 = pbv_{\mathrm{R}}/v_{\mathrm{L}}$, and so balance of evidence with start-point bias is:

$$BoE_{\mathrm{SB}} = b(1 - pv_{\mathrm{R}}/v_{\mathrm{L}}). \tag{A.4}$$

This is increased relative to the unbiased case (i.e., $\mathrm{BoES}_{\mathrm{SB}} > \mathrm{BoE}_{\mathrm{U}}$), with the biased case greater by:

$$BoE_{\mathrm{SB}} - BoE_{\mathrm{U}} = b(1 - p)v_{\mathrm{R}}/v_{\mathrm{L}} > 0. \tag{A.5}$$

The increase occurs because the evidence in the winner is always b (i.e., it does not change with start point) but an increase in start point must decrease the evidence in the loser (and hence increase the BoE) because the winner finishes earlier, leaving less time for the loser to accrue evidence before the winner triggers the response.

The cases illustrated in Fig. A.1 do not take account of the full complexity of real evidence accumulation models, which typically account for errors through random variation from trial to trial in accumulation rates and start points (i.e., random biases) and/or through stochastic variability. Proofs in these cases require more than the simple analysis given here, and likely require consideration of explicit distributional assumptions, which vary among both deterministic (e.g., Terry et al., 2015) and stochastic (e.g., Logan, Van Zandt, Verbruggen, & Wagenmakers, 2014; Ratcliff & Smith, 2004) accumulator models. Predictions may also vary depending on specific parameter values, and so require simulation and/or parametric model fitting to investigate. However, a simple analysis of linear deterministic accumulator models can still make some progress with respect to predictions about correct and error responses where there is only trial-to-trial rate variability, at least when certain simplifying approximations are made, and we develop and test these predictions here. Note that in the unbiased case, the varying-rate model is restricted to predicting that errors are slower than correct responses. Slow errors usually occur with careful responding when speed is not emphasized over accuracy, so we adopt that setting in our experiment and check whether it was effective by testing the relative speed of unbiased correct and error responses.

A.3. Rate variability

In the unbiased case with rate variability, both correct and error responses will have higher rates for the winner than the loser, and so (A.1) continues to hold, with $\mathrm{BoE}_{\mathrm{U}}$ tending to be larger for correct than error responses because the ratio $\mathrm{v}_{\mathrm{R}}/\mathrm{v}_{\mathrm{L}}$ will tend to be closer to one for error responses (i.e., the difference in rates between winning and losing accumulators will be smaller on average for error than correct responses). In the biased case, however, complications can arise because bias can cause the response to be different from the unbiased case. A change in response can occur because it is the accumulator with the lower rate (i.e., the accumulator corresponding to the correct response) that is favoured by the bias. If the effect of the bias is sufficient to overcome the rate advantage of the accumulator corresponding to the wrong response, then a correct response will be made in the biased case. Hence, correct responses will be a mixture of cases with the winner having both higher and lower rates than the loser, and error responses will be made up of only cases where the bias was insufficient to overcome the rate advantage for the wrong accumulator. As a result, exact predictions will be both distribution and parameter dependent. Again, to make progress with a simple analysis, we assume an approximation in which response switching between biased and unbiased cases does not occur.

For correct responses under this approximation, all of the deductions about the biased cases remain the same, and so (A.2)–(A.5) still hold. For error responses, the bias will now be against the winner (i.e., against the wrong response). Keeping with the convention in Fig. A.1 that the left accumulator wins, for both types of bias the winner will achieve threshold at $t_1 = b/v_{\mathrm{L}}$. In the threshold bias case, the loser starts at zero and so at t_1 will have achieved a level $v_{\mathrm{R}} t_1 = bv_{\mathrm{R}}/v_{\mathrm{L}}$ (note that the lower threshold for the loser is irrelevant as we have stipulated the case where it has not crossed that threshold at t_1). Hence, the balance of evidence in this "threshold biased against" case is:

$$BoE_{\mathrm{TBA}} = b(1 - i_{\mathrm{R}}/v_{\mathrm{L}}). \tag{A.6}$$

This is exactly the same as for the unbiased case and so:

$$BoE_{\mathrm{U}} - BoE_{\mathrm{TBA}} = 0. \tag{A.7}$$

For the "start-point biased against" case, the loser starts at pb and so at t_1 will have a level $pb + \mathrm{v}_{\mathrm{R}} t^1 = pb + b\mathrm{v}_{\mathrm{R}}/\mathrm{v}_{\mathrm{L}}$. Hence, balance of evidence is:

$$BoE_{\mathrm{SBA}} = b(1 - p - v_{\mathrm{R}}/v_{\mathrm{L}}). \tag{A.8}$$

In this case the sign of the difference is the opposite to the case for a correct response:

$$BoE_{\mathrm{SBA}} - BoE_{\mathrm{U}} = -bp < 0. \tag{A.9}$$

In summary, for errors the difference in evidence between the biased and unbiased cases is either zero or the opposite to correct responses.

A.4. "Bias-for" vs. "bias-against" conditions

A problem with evaluating the bias effect relative to an unbiased condition is that participants may have different levels of caution (i.e., the overall level of thresholds for both accumulators, e.g., b in the notation developed for Fig. A.1) between biased and unbiased conditions. This is plausible because they must know the nature of the bias condition (i.e., unbiased vs. biased towards a particular response) before the stimulus appears. This problem can be avoided by making the comparison between stimuli that are favoured and disfavoured by the bias. Because the stimulus is not known prior to the trial, thresholds cannot be differentially adjusted between these two conditions. Hence, we now develop predictions for comparisons between stimuli that are favoured and disfavoured by the bias, which we then test in the experiment.

We first examine predictions for correct responses. For the threshold case, we require the difference between (A.2) and (A.6), which shows that confidence in a decision consistent with the bias will be *less* than confidence in decisions against the bias (i.e., $BoE_{\mathrm{TBFc}} < BoE_{\mathrm{TBAc}}$):

$$BoE_{\mathrm{TBAc}} - BoE_{\mathrm{TBFc}} = (1-p)b(1-v_{\mathrm{R}}/v_{\mathrm{L}}) > 0. \tag{A.10}$$

For the start-point case, we require the difference between (A.4) and (A.8), which shows that confidence in a decision consistent with the bias will be *greater* than confidence in decisions against the bias (i.e., $BoE_{\mathrm{SBFc}} > BoE_{\mathrm{SBAc}}$):

$$BoE_{\mathrm{SBFc}} - BoE_{\mathrm{SBAc}} = b[p + (1-p)v_{\mathrm{R}}/v_{\mathrm{L}}] > 0. \tag{A.11}$$

For error responses, we need only swap the roles of bias-for and bias-against equations and so the directions of the confidence differences reverse. Hence, for the threshold case, confidence in decisions consistent with the bias will be *greater* than confidence in decisions against the bias (i.e., $BoE_{\mathrm{BFTe}} > BoE_{\mathrm{TBAe}}$):

$$BoE_{\mathrm{TBAe}} - BoE_{\mathrm{TBFe}} = -(1-p)b(1-v_{\mathrm{R}}/v_{\mathrm{L}}) < 0. \tag{A.12}$$

For the start-point case confidence, confidence in decisions consistent with the bias will be *less* than confidence in decisions against the bias (i.e., $BoE_{\mathrm{SBFe}} < BoE_{\mathrm{SBAe}}$):

$$BoE_{\mathrm{SBFe}} - BoE_{\mathrm{SBAe}} = -b[p + (1-p)v_{\mathrm{R}}/v_{\mathrm{L}}] < 0. \tag{A.13}$$

Predictions for bias for vs. bias against differences in accuracy are straightforward: accuracy will be greater in the bias-for case. For correct RT the opposite ordering holds, RT will be less in the bias-for case because the distance from start-point to threshold is less. For error RTs the ordering is the opposite (i.e., the same as for accuracy) because the bias is against the correct response but for the error response, and so the distance from start-point to threshold is less in the bias-against case.

References

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

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. *R package version*, *1*(7), 1–23.

Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, *2*, 317–352.

Brown, S., & Heathcote, A. (2003). Averaging learning curves across and within participants. *Behavior Research Methods, Instruments, & Computers*, *35*(1), 11–21.

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*, 153–178.

Cohen, A. L., Sanborn, A. N., & Shiffrin, R. M. (2008). Model evaluation using grouped or individual data. *Psychonomic Bulletin & Review*, *15*(4), 692–712. http://dx.doi.org/10.3758/PBR.15.4.692.

Davis-Stober, C., Morey, R. D., Gretton, M., & Heathcote, A. (2016). Bayes factors for state-trace analysis. *Journal of Mathematical Psychology*, *72*, 116–129.

de Condorcet, M., & de Marquis, C. (1785). *Essai sur l'application de l'analyse à la probabilité devdécisions rendues à la pluralité de voix, imprimerie royal*. Paris: Imprimerie Royale.

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

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

Evans, N. J., Brown, S. D., Mewhort, D. J. K., & Heathcote, A. (in press). Refining the law of practice. Psychological Review.

Ferster, C. B., & Skinner, B. F. (1957). *Schedules of reinforcement*. East Norwalk, CT, US: Appleton-Century-Crofts.

Fox, J., Friendly, M., & Weisberg, S. (2013). Hypothesis tests for multivariate linear models using the car package. *The R Journal*, *5*(1), 39–52.

Gehrlein, W. V. (1983). Condorcet's paradox. *Theory and Decision*, *15*(2), 161–197. http://dx.doi.org/10.1007/bf00143070.

Gehrlein, W. V. (2002). Condorcet's paradox and the likelihood of its occurrence: Different perspectives on balanced preferences. *Theory and Decision*, *52*(2), 171–199. http://dx.doi.org/10.1023/a:1015551010381.

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review*, *7*(2), 185–207.

Heathcote, A., & Love, J. (2012). Linear deterministic accumulator models of simple choice. *Frontiers in Cognitive Science*, *3*, 292. http://dx.doi.org/10.3389/fpsyg.2012.00292.

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

Kolossa, A., & Kopp, B. (2018). Data quality over data quantity in computational cognitive neuroscience. *NeuroImage*, *172*, 775–785. http://dx.doi.org/10.1016/j.neuroimage.2018.01.005.

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

Logan, G. D., Van Zandt, T., Verbruggen, F., & Wagenmakers, E.-J. (2014). On the ability to inhibit thought and action: General and special theories of an act of control. *Psychological Review*, *121*(1), 66–95.

Mulder, M. J., Wagenmakers, E.-J., Ratcliff, R., Boekel, W., & Forstmann, B. U. (2012). Bias in the brain: a diffusion model analysis of prior probability and potential payoff. *Journal of Neuroscience*, *32*(7), 2335–2343. http://dx.doi.org/10.1523/JNEUROSCI.4156-11.2012.

Noorbaloochi, S., Sharon, D., & McClelland, J. L. (2015). Reward information biases a fast guess process in perceptual decision making under deadline pressure: Evidence from behavior, evoked potentials, and quantitative model comparison. *Journal of Neuroscience*, *35*(31), 10989–11011. http://dx.doi.org/10.1523/JNEUROSCI.0017-15.2015.

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

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review*, *111*(2), 333–367. http://dx.doi.org/10.1037/0033-295X.111.2.333.

Shiffrin, R., Lee, M., Kim, W., & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, *32*(8), 1248–1284. http://dx.doi.org/10.1080/03640210802414826.

Smith, P. L., & Little, D. R. (in press). Small is beautiful: In defense of the small-N design. Psychonomic Bulletin & Review.

Terry, A., Marley, A. A. J., Barnwal, A., Wagenmakers, E.-J., Heathcote, A., & Brown, S. D. (2015). Generalising the drift rate distribution for linear ballistic accumulators. *Journal of Mathematical Psychology*, *68–69*, 49–58.

Vickers, D. (1979). *Decision processes in visual perception*. Academic Press.

Vickers, D. (1985). Antagonistic influences of performance change in detection and discrimination tasks. In G. Cognition (Ed.), *Proceedings of the XXIII international congress of psychology*: *vol. 3. Information processing, and motivation* (pp. 79–115). New York: North Holand.

Vickers, D. (2001). Where does the balance of evidence lie with respect to confidence? In E. Sommerfeld, R. Kompass, & T. Lachmann (Eds.), (pp. 148–153). Presented at the *Proceedings of the 17th annual meeting of the International Society for Psychophysics*.

Vickers, D., & Lee, M. D. (1998). Dynamic models of simple judgments: I. Properties of a self-regulating accumulator module. *Nonlinear Dynamics, Psychology, and Life Sciences*, *2*(3), 169–194.