Registered Report

QJEP

Quarterly Journal of Experimental Psychology
2021, Vol. 74(1) 122–134

DOI: 10.1177/1747021820963033
qjep.sagepub.com
SAGE

Does speed of recognition predict two-alternative forced-choice performance? Replicating and extending Starns, Dubé, and Frelinger (2018)

Anne Voormann[1], Annelie Rothe-Wulf[1], Jeffrey J Starns[2] and Karl Christoph Klauer[1]

Abstract

Does the speed of single-item recognition errors predict performance in subsequent two-alternative forced-choice (2AFC) trials that include an item with a previous error response? Starns, Dubé, and Frelinger found effects of this kind in two experiments and accounted for them in terms of continuous memory-strength signal guiding recognition decisions. However, the effects of error speed might just as well only reflect an artefact due to an error-correction strategy that uses response latency as a heuristic cue to guide 2AFC responses, elicited through confounding factors in their experimental design such as error-correction instructions and feedback. Using two conditions, a replication condition, replicating the procedure from Starns et al., and an extension condition (each $n = 130$), controlling for the named shortcomings, we replicated the error speed effect. In both conditions, speed of errors in a single-item recognition task was predictive of subsequent 2AFC performance, including the respective error item. To be more precise, fast errors were associated with decreased 2AFC performance. As there was no interaction with the factor condition, the results support the idea that speed of single-item recognition responses reflects the amount of memory information underlying the respective response rather than being used for a simple error-correction strategy to improve 2AFC performance.

Keywords

Recognition memory; discrete-state models; diffusion model

Received: 31 January 2019; revised: 13 July 2020; accepted: 16 July 2020

An ongoing discussion in recognition memory concerns the question of whether decision makers have access to a continuous memory-strength signal when making recognition decisions or whether such decisions are based on discrete states (for a review see Pazzaglia et al., 2013 but see also Batchelder & Alexander, 2013; Dubé et al., 2013). Theories assuming continuous memory-strength signals postulate that individuals have access to a graded summary of the accumulated or sampled evidence from memory (e.g., Banks, 1970; Ratcliff, 1978; Van Zandt, 2000). Thus, when conducting a recognition decision, as in deciding whether a person has been previously encountered, the decision is based on the amount of familiarity elicited by this person relative to a critical value. In contrast, theories based on discrete states acknowledge that there might be a continuous sampling of evidence from memory, but decisions are ultimately mediated via discrete "detect" or "uncertainty" states (e.g., Luce, 1963; Province & Rouder, 2012; Snodgrass & Corwin, 1988). Thus, in the above-mentioned case, the person is either detected as being known or, if not detected, one guesses based on contextual information. With the present study, we want to contribute to this discussion through replication and extension of a previously conducted test proposed to discriminate those two mechanisms.

[1]Department of Psychology, University of Freiburg, Freiburg, Germany
[2]Department of Psychological and Brain Sciences, University of Massachusetts Amherst, Amherst, MA, USA

Corresponding author:
Anne Voormann, Department Psychology, University of Freiburg, D-79085 Freiburg, Germany.
Email: anne.voormann@psychologie.uni-freiburg.de

Generally, recognition describes the ability to identify previously experienced situations as such. Thus, in a typical recognition memory test (single-item recognition task) participants learn a list of words in a study phase and judge in a subsequent test phase whether the presented words have previously been studied ("old" items) or not ("new" items). Responses are coded in one of four response categories, termed *hits*, *misses, correct rejections* and *false alarms*: a studied word can be judged to be an old item (hit) or a new item (miss); a new item can be judged to be a new item (correct rejection) or an old item (false alarm).

Another frequently used paradigm is the two-alternative forced-choice (2AFC) task that differs from the single-item recognition task with respect to the test phase. In the 2AFC task, two words are presented at test, one target and one lure. Participants' task is to indicate which word is the target (Green & Swets, 1966).

To discriminate whether such decisions are based on continuous or discrete information is quite difficult. Researchers developed various methods to distinguish between these two possibilities using both model comparison methods and experimental designs evoking qualitatively different behavioural expectations. For example receiver operating characteristic curves (ROCs; Bröder & Schütz, 2009), first and second choice responses (Kellen & Klauer, 2011), tests for conditional independence (Kellen et al., 2015; Province & Rouder, 2012), model selection based on minimum description length (Kellen et al., 2013; Klauer & Kellen, 2015), or ranking tasks (Kellen & Klauer, 2014) were used to discriminate between continuous and discrete processing of memory evidence. The results of these studies are mixed, some favouring continuous models and others discrete-state models.

Recently, Starns et al. (2018) investigated this question by combining a single-item recognition task with a subsequent 2AFC task. First, participants worked on a single-item recognition task making "old" versus "new" judgements. Then, a 2AFC task followed, in which participants were asked to identify the target in a pair of words that combined an item they had misclassified in the single-item recognition task and an item they had correctly responded to. The authors expected that the speed of errors in the single-item recognition task would predict the performance in 2AFC trials that include the error items. Specifically, Starns et al. (2018) hypothesised that fast errors in the single-item recognition task are associated with lower performance within the 2AFC task than slow errors. This hypothesis rests on a theory of memory retrieval in which recognition decisions are based on a diffusion process of sampling evidence from memory (Ratcliff, 1978). In this framework, fast errors—compared with slow errors—reflect higher degrees of misleading evidence elicited by a memory probe, which should affect decisions in the 2AFC task involving these same memory probes. In contrast, based on discrete accounts such as the two-high threshold model (2HTM; Snodgrass & Corwin, 1988), Starns et al. (2018) predicted no differences in performance of a 2AFC task between fast and slow errors.

In the following, we will elaborate on the different theories and their predictions regarding fast and slow errors. Then, we will discuss the results of Starns et al. (2018) and possible confounds within their paradigm motivating the present study.

Continuous models

As mentioned above, continuous models assume access to a continuous familiarity signal as a basis for recognition decisions. For response selection, the accumulated familiarity needs to either reach a threshold or exceed a critical value. For example in signal detection theory, one of the first continuous recognition memory models, the perceived familiarity is compared with a response criterion (Banks, 1970). Each time the perceived familiarity exceeds the criterion the response "old" is elicited, whereas if the familiarity falls below the criterion participants make a "new" response.

Diffusion model

One prominent example of continuous accounts, the diffusion model (Ratcliff, 1978), can be understood as a dynamic extension of signal detection theory. In this account, information is accumulated in a diffusion process. The drift rate of the diffusion process is the rate at which information accrues from memory and represents the strength of the memory signal (Ratcliff & Starns, 2009) comparable to the level of perceived familiarity within signal detection theory. Evidence accumulation terminates once the accumulated amount of evidence reaches one of two boundaries, an upper one associated with the "old" response, or a lower one associated with the "new" response. Thus, within this framework, the higher the drift rate of a specific stimulus, the higher the strength of the memory signal elicited by this stimulus. Consequently, the higher the drift rate the faster a threshold is reached and the associated response delivered.

Within the diffusion model, erroneous responses occur each time the accumulated information reaches the incorrect threshold. This can occur because of two mechanisms. On one hand, the average drift rate can point to the correct threshold, but because a noisy diffusion process characterises the aggregation of evidence, variability exists within the sampling process and, hence, the incorrect threshold can be reached by accident (avoidable errors; Ratcliff, 2014). In the case of recognition memory, this could be a target word which was initially categorised as "new" but with further consideration, one would identify the error and correctly classify it as being

"old." On the other hand, variability of the drift rates exists across trials. Thus, a stimulus can have a drift rate pointing in the direction of the incorrect threshold. In the case of recognition memory this could be a target which is not memorised, for example, due to inattention during study (Ratcliff, 2014). Consequently, evidence accumulates towards the threshold eliciting a "new" response and error reaction times (RTs) would be an indicator of the strength of that evidence.

Therefore, under certain conditions, fulfilled in Starns et al.'s (2018) experiments, the speed of errors is diagnostic of the drift rate likely underlying the erroneous decision.[1] Specifically, fast errors are associated with higher drift rates pointing towards the false threshold than slow errors, and thus they reflect systematically misleading information from memory.

Based on this rationale for errors, Starns et al. (2018) hypothesised that fast errors in a single-item recognition task should on average result in lower performance within subsequent 2AFC trials consisting of an error item and a word correctly responded to compared with slow errors. This is because items with fast errors are likely to elicit stronger misleading memory information than items with slow errors.

Discrete-state models

Discrete-state models or threshold models of recognition memory have in common that responses are based on a combination of a number of discrete mental states, typically detection and uncertainty states (Riefer & Batchelder, 1988). In addition, as soon as a certain state is entered, only the information about the state remains while all information about how this state was reached is lost. There exist different models such as the one-high threshold model (Blackwell, 1963), the 2HTM (Snodgrass & Corwin, 1988), or the one-low threshold model (Luce, 1963). These models differ in the number and type of thresholds as well as associated mental states.

2HTM

The most popular threshold model is the 2HTM with three different states, a detection state for old words, a detection state for new words, and an uncertainty state (Snodgrass & Corwin, 1988). With a certain probability d_o (d_n), a target (lure) enters the respective detection state leading to the corresponding response "old" ("new"). Each time detection fails, guesses determine the response out of a state of uncertainty. Thus, within this framework, errors only occur if a stimulus could not be detected correctly and the response is guessed incorrectly. In the case of recognition memory, this could be a target which is not detected as such and then guessed to be "new" or a lure which is not detected as lure and then guessed to be "old."

Generally, it is possible to draw inferences from RTs to response states within the 2HTM. For example, detecting a stimulus might occur faster than not detecting a stimulus and guessing the response (Klauer & Kellen, 2018). However, conditional on having entered a mental state, responses and RTs should not depend upon further mnemonic evidence. Under this "information loss" assumption, the speed of errors arising from the uncertainty state should not contain information on the underlying memory strength.

Thus, as errors can only occur out of the same underlying state of uncertainty, this framework allows no inference about the amount of memory evidence based on response speed. Therefore, fast and slow erroneous reactions do not differ in their informational value. Based on these considerations, Starns et al. (2018) predicted no difference in the performance on fast and slow error trials in a subsequent 2AFC task for the 2HTM.

Two-low threshold *model*

Starns et al. (2018) also proposed a discrete-state model that allows for misleading retrieval, the two-low threshold model (2LTM). This model has the same form as the 2HTM discussed above, except that it allows for the possibility that a lure item could misleadingly produce the "detect old" state and a target could misleadingly produce the "detect new" state, necessitating additional parameters to represent the probability that these outcomes will occur. Under the assumption that participants respond more quickly from the detect states than from the more ambiguous uncertainty state, the 2LTM predicts, contrarily to the 2HTM, a relationship between single-item error speed and subsequent 2AFC accuracy that matches the qualitative pattern predicted by the diffusion model. That is, errors in this model are a mixture of guesses from a state of uncertainty and responses based on misleading retrieval, and faster errors should tend to come from the latter category. Misleading retrieval impairs 2AFC performance, creating lower accuracy for trials with a fast-error word than trials with a slow-error word. Thus, the 2LTM, like the diffusion model, predicts a relationship between error speed and 2AFC accuracy.

Findings by Starns et al. and possible shortcomings

As predicted by the diffusion model account, but not by the 2HTM, Starns et al. (2018) found that the performance in 2AFC trials including fast errors from the single-item recognition task was worse compared with the performance in trials including slow errors.

However, in Starns et al.'s (2018) experiments, the 2AFC task always paired one item erroneously responded to with one item correctly responded to in the single-item

recognition task, and participants were instructed to correct their previous error. Participants' task was thus to identify the error item to correct their previous error by a correct 2AFC decision. In a state of uncertainty, participants searching for the previously misclassified item might then well use response latency as a heuristic cue to identify the likely error. To be more precise, in the absence of any real or valid information from memory, participants might utilise response latency to aid their decision, surmising that an item for which they previously required a long time to come to a decision is likely to be the item with the error. Response latency might thereby bias guessing processes rather than reflect underlying memory dynamics.

To rule out strategic effects of this kind, Starns et al. (2018) examined as part of their Experiment 2B the strategy that participants used during the 2AFC task by means of a retrospective survey. In a direct question asking whether participants considered response times from the single-item recognition test to inform their answers on the subsequent 2AFC task, half of the participants indicated that they used this strategy. This surprising outcome strengthens the assumption that a metacognitive strategy of error identification based on response latency causes the results. Note, however, that Starns et al. (2018) interpreted the outcomes of the strategy survey as evidence against the heuristic use of response latency to identify the likely error item in a state of uncertainty. This conclusion was based on two observations: (1) only two participants mentioned a latency-based strategy in an open-ended question that preceded the direct question and (2) participants who reported that they *did not* consider the latency of their previous responses showed an effect of error RT on subsequent 2AFC performance that was statistically indistinguishable from the effect shown by participants who *did* report this strategy.

The usage of a metacognitive guessing strategy based on response latency could have been encouraged because Starns et al. (2018) implemented an error feedback within their design. They constructed their 2AFC trials in such a way that each trial consisted of an error item and a word correctly responded to. In addition, they informed participants about this composition of 2AFC trials and instructed them to correct previously made errors.

Note that the diffusion model's predictions in no way depend on the procedure of describing the 2AFC task as a chance to correct previous errors, nor do they depend on the practice of creating all 2AFC trials by pairing an item with a previous error and an item with a previous correct response. According to the diffusion model account, fast errors should be associated with impaired performance, compared with slow errors, in trials in which the error item is paired with an item correctly responded to. This should occur even if participants are given more traditional 2AFC instructions and even if some of the 2AFC trials contain words that both received a previous error or both received a previous correct response, meaning that the 2AFC trial does not provide feedback by revealing that one of the words must have received a previous error response.

Although the diffusion model's predictions do not change when the error-feedback component of the 2AFC task is eliminated, previous studies showed that feedback can have an impact on results within memory tasks. For instance Malejka and Bröder (2016) showed that findings by Starns et al. (2008) of source memory for unrecognised items may have been an artefact of providing feedback, allowing participants to infer that they previously responded incorrectly to the item in question. Removing the feedback eliminated the evidence for source memory for unrecognised items, leading Malejka and Bröder (2016) to conclude that the effect was an artefact of the design providing feedback. Motivated by this preceding evidence, we propose an extension of Starns et al.'s (2018) design that removes the error feedback.

Aims and hypotheses

In the light of the shortcomings mentioned, our goals were twofold: first, we intended to replicate Starns et al.'s (2018) results, closely following their design. In particular, the replication condition uses the same construction of 2AFC trials and the same error-correction instruction as Starns et al. (2018). Second, we implemented an extension condition, in which we removed the error feedback by using a more traditional construction of the 2AFC trials and omitting the error-correction instruction.

Because this study is an instance of a so-called adversarial collaboration between Anne Voormann, Annelie Rothe-Wulf, and Karl Christoph Klauer on the one hand and Jeffrey J. Starns on the other hand, we pre-registered the experimental procedure and hypotheses prior to data collection. For the replication condition, both groups predicted that there would be a difference in the 2AFC performance as a function of the speed of errors in the single-item recognition task such that fast errors lead to worse 2AFC performance than slow errors.

To assess the possibility that this effect, if it replicated, reflects a metacognitive strategy of error-detection on the basis of response latency, we introduced the extension condition. Jeffrey J. Starns predicted a difference in the 2AFC performance of fast and slow errors for this condition just as for the replication condition, based on the diffusion model account. Thus, significant effects of speed of errors are expected in both the extension and replication conditions. Because participants will not be in an error-detection mode in the extension condition, Anne Voormann, Annelie Rothe-Wulf, and Karl Christoph Klauer predicted to the contrary that the effect would be eliminated if it relies on a metacognitive strategy of error detection. Thus, an interaction between speed of errors and experimental

condition was predicted along with a significant effect of speed of errors in the replication condition, but not in the extension condition.

Other effect patterns than the ones just considered are possible. For example, a pattern with no interaction due to no effect in both conditions as well as no interaction but a significant effect of speed of errors in the replication condition and no such effect in the extension condition would be ambiguous. We therefore based our power planning on the interaction effect. As detailed below, this also resulted in satisfactory power for the individual condition-wise tests for effects of error speed on 2AFC accuracy so that the absence of individually significant effects in the replication condition or in both conditions could be clearly interpreted as a failed replication.

Methods

We pre-registered this study at the Quarterly Journal of Experimental Psychology as pre-registered report; the approved pre-registration protocol as well as all materials, analysis scripts and data files are publicly available on OSF (https://osf.io/ejucx/).

In general, our experimental design combined Starns et al.'s (2018) procedure of Experiments 1 and 2 following Experiment 2 as closely as possible while implementing slightly larger single-item recognition test blocks similar to Experiment 1. This permitted a more balanced construction of 2AFC trials in the extension condition. In addition, a pilot-study suggested slightly higher accuracies and slower responses for German participants in a single-item recognition task that thus provided fewer critical 2AFC trials with previous errors for our participants compared with Starns et al.'s (2018) participants (see Table 1). We expected that introducing slightly larger study-test cycles compared with Starns et al.'s (2018) Experiment 2 would reduce performance slightly (Cary & Reder, 2003) and produce hit and false alarm rates that are comparable to Starns et al.'s (2018) Experiment 2. Apart from that, the design and the procedure followed Starns et al. (2018) with the small modifications explicated below.

Participants

In total, 268 participants took part in this study, from which three had to be excluded because of computer problems, two aborted the experiment prior to final data collection, and three participants did not satisfy the inclusion criteria defined as a difference between hit rates and false alarm rates of at least 0.1 within the single-item recognition test. This resulted in the pre-registered number of 260 valid datasets ($n = 130$ per condition). Age ranged from 17 to 46 years with a mean of $M = 23.94$ years ($SD = 4.89$). Participants were recruited from the participant pool of the department Social Psychology and Methodology, University of Freiburg. All participants spoke German as a first language and received either partial course credit or a monetary reward for their participation.

Table 1. Mean hit and false alarm rate and median latency (RT) in Starns et al.'s (2018) Experiment 2 and for 10 pilot participants.

	Hit		False alarm	
	Rate	RT	Rate	RT
Starns et al. (2018)	0.64	919 ms	0.26	1,150 ms
Pilot participants	0.74	1,030 ms	0.16	1,349 ms

RT: reaction time.

To determine the number of needed valid datasets, we conducted an a-priori power analysis, using the observed effect of speed of errors on 2AFC performance reported by Starns et al. (2018) in their analysis-of-variance (ANOVA). The smallest effect size across Starns et al.'s (2018) experiments for this critical effect was $d_z = 0.31$ (range: 0.31–0.34). Assuming a possible interaction in the direction that there is an effect of size $d_z = 0.31$ in the replication condition but a null effect in the extension condition, the effect for the group comparison will be $d = 0.31$ (assuming equal variances in the two groups). To find an effect of this size given $\alpha = .05$ and $\beta = .80$ in a one-tailed independent two-sample *t*-test, 130 valid datasets per condition are necessary. Across the replication and extension conditions, we thus planned to collect valid datasets from 260 participants. The power for detecting an effect of speed of errors with effect size $d_z = 0.31$ will thereby be $\beta = .97$ in a one-tailed *t*-test with $\alpha = .05$ in each condition with $n = 130$.

Design

Our study implemented two conditions: a replication condition and an extension condition. In both conditions, we presented three study-test cycles, consisting of one practice cycle and two experimental cycles. Each test phase included several single-item recognition blocks and subsequent 2AFC blocks. The two conditions differed in the construction and number of the 2AFC trials and in their instructions.

Materials and list composition

Words were randomly drawn from a wordpool consisting of 639 neutral German nouns taken from a study by Lahl et al. (2009). The words were four to eight letters long with ratings medium in valence (ranging from 3.5 to 6.5 on an 11-point scale) and low in arousal (ranging from 0.5 to 4.5 on an 11-point scale). All words were approximately equally frequent according to the log frequency ratings obtained for each word via WordGen (ranging from 0.3 to 2.9; Duyck et al., 2004).

Departing from Starns et al. (2018), the study list consisted of 28 words in the practice cycle and 80 words in the experimental cycles. Words were presented in groups of four during the study phase, the first and last group of each study list serving as filler words.

The single-item recognition test lists consisted of two blocks of 20 words each in the practice cycle and of six blocks of 18 words each in both experimental cycles. Targets and lures were counterbalanced within blocks. Furthermore, the first single-item recognition test block of each experimental cycle started with four additional warm-up trials. Warm-up trials consisted of two targets, taken from the first four filler words in the study list, and two lures and were discarded for analysis.

In each 2AFC trial, a target and a lure were presented. In the replication condition, a 2AFC trial always paired either a miss with a correct rejection or a hit with a false alarm as classified on the basis of responses in the immediately preceding single-item recognition task. For each block, as many critical trials as possible were constructed, as in Starns et al. (2018). In the extension condition, each 2AFC block comprised 10 trials in the practice cycle and 9 trials in the experimental cycles, including all possible pairings based on the performance of the single-item recognition task in addition to the critical trials (hit—false alarm; hit—correct rejection; miss—false alarm; miss—correct rejection). For the creation of the 2AFC pairs, critical pairs (hit—false alarm; miss—correct rejection) were, however, favoured, so that there were as many critical pairs as would have been constructed in the replication condition.

The 2AFC instructions differed between conditions. The replication condition implemented the original instructions from Starns et al. (2018) translated into German. Here, participants were informed that every 2AFC trial consisted of one word correctly responded to in the single-item recognition task and one word with an erroneous previous response. Furthermore, participants were instructed to identify and correct previously made errors. In the extension condition, participants were told that one of the words in each 2AFC trial is a studied one, the other one a not-studied one. Here, participants were simply asked to try their best to select the previously studied word.

Procedure

The experiment was programmed in C++, and one session lasted about 30 min. Participants were randomly assigned to conditions. The procedure followed the one described in Starns et al. (2018).

Prior to the experiment, participants provided informed consent. After the instructions, the study-test cycles started. In the study phase, words were presented sequentially at screen centre for 1,900 ms. Like in Starns et al. (2018), each group of four study items was followed by a recall task cued by position for one of the four words (e.g., "recall the second word"). This method was used to ensure that participants encode all words presented during study. If the participant's typed responses were incorrect, an error message appeared for 1,000 ms. After a blank screen for 500 ms, the next four study items proceeded. Words selected for the recall task were excluded from both the single-item recognition task and the 2AFC task, except for in the practice cycle. After all study words were presented, the recognition task began with alternating blocks of single-item and 2AFC trials.

In each single-item recognition trial, one word was presented at screen centre until a response was given. Participants used the "Y" and "-" key of a German QWERTZ keyboard to indicate old and new words. The response labels "ALT" and "NEU" (German for "OLD" and "NEW") were visible below the stimulus aligned horizontally with the response keys to be used for the old/new responses.

After each block of single-item recognition trials, participants worked through the 2AFC task. As in Starns et al.'s (2018) Experiment 2, in each trial, the two words were presented successively for 1,000 ms each, starting with the left word. Then, both words were presented together until a response was given. This procedure should ensure that participants considered both words and did not respond based on the memory signal of only one word (Starns et al., 2017). The "Y" key of a German QWERTZ keyboard was used to indicate the left word as the target and the "-" key to indicate the right word. Again, labels signalling the response mapping were visible during the presentation of both words. The left/right positions of targets and lures were counterbalanced across trials of each block of 2AFC trials.

Data elimination and analyses

Data elimination and analysis methods, including quality checks, were parallel to the analyses of Starns et al. (2018). Additional to the individual *t*-tests for effects of speed of errors on the 2AFC performance within each condition, an ANOVA included the factors error speed (fast vs. slow) and trial type (2AFC trial consisting of words previously responded to "studied"—S–S vs. "not studied"—N–N), and the factor experimental condition (replication vs. extension) to investigate whether there is an interaction between the speed of errors and the experimental conditions.

In a Bayesian hierarchical logistic regression patterned after the one described in Starns et al. (2018), we also included a possible effect of condition on slope to examine whether the condition has an influence on the error RT slope. The Bayesian hierarchical logistic regression was conducted in addition to the ANOVA to allow error RTs within single-item recognition trials to be included as a

Table 2. Results from a mixed ANOVA on proportion correct in two-alternative forced-choice trials including condition as between-subject factor (replication vs. extension) and trial type (studied–studied (S–S) vs. not-studied–not-studied (N–N)) and error speed (fast vs. slow) as within-subject factors.

	MSE	*F*(1,256)	η_g^2	*p*
Condition	0.03	1.26	0.002	.26
Trial type	0.02	106.83***	0.09	<.001
Condition × trial type	0.02	4.22*	0.004	.04
Error speed	0.03	22.21***	0.02	<.001
Condition × error speed	0.03	0.36	0.0003	.55
Trial type × error speed	0.02	3.8	0.003	.052
Condition × trial type x error speed	0.02	0.04	<.0001	.84

ANOVA: analysis of variance; MSE: mean squared error; *F*: *F*-value; η_g^2: generalised η^2; *p*: *p*-value.
*$p < .05$.
***$p < .001$.

continuous predictor rather than discretised as in the ANOVA as well as to take into account the different trial numbers per trial type and participant.

Results

Hit (H) and false alarm (FA) rates did not differ significantly between both conditions (Hs: $t(258) = 0.92$, $p = .356$; $d = 0.11$, 95% confidence interval (CI) on standardised effect size [−0.13, 0.36]; FAs: $t(258) = 1.65$, $p = .101$; $d = 0.20$, 95% CI [−0.04, 0.45]). Mean rates were $M_H = .66$ and $M_{FA} = .19$ in the replication condition, and $M_H = .64$ and $M_{FA} = .17$ in the extension condition. Furthermore, the number of critical 2AFC trials was comparable between conditions, $M = 56.5$, $SD = 15.3$ in the replication condition and $M = 56.1$, $SD = 15.6$ in the extension condition, $t(258) = 0.21$, $p = .838$; $d = 0.03$, 95% CI [−0.22, 0.27]. Critical 2AFC trials consisted of one erroneous and one correct single-item recognition response so that both responses in the single-item recognition task were either "studied" or "not-studied."

Forced-choice performance

To investigate the effect of condition on the error speed effect, we conducted a mixed ANOVA with condition (replication vs. extension) as a between-subject factor and error speed in single-item recognition trials (fast vs. slow) and trial type (S–S vs. N–N) as within-subject factors. Mean percent correct in critical 2AFC trials served as the dependent variable. Parallel to Starns et al. (2018), we discarded trials with single-item recognition decisions faster than 400 ms or slower than 8 s from analysis. To discretise error speed, we used a median split including only the error RTs of single-item recognition trials that were included in critical trials. We separately evaluated the median for each participant and word type (target vs. lure). All single-item recognition responses being smaller than the respective median RT were categorised as fast errors, while single-item recognition responses slower than or equal to the respective median RT were categorised as slow errors. Two participants from the extension condition had to be discarded from analyses because both committed only one false alarm and thus median RTs could not be computed. Table 2 lists the complete results of the ANOVA. In the following, we will focus on the effects relevant for our hypothesis and on significant effects.

As can be seen from Table 2, the ANOVA revealed no effect of condition, $\eta_g^2 = .002$, and 95% CI = [0.00, 0.03], indicating that there was no significant 2AFC performance differences between conditions, $P(\text{correct} | \text{replication}) = .66$, $SD_{\text{Rep}} = .09$, $P(\text{correct} | \text{extension}) = .65$, and $SD_{\text{Ext}} = .09$. Replicating the results of Starns et al. (2018), there was a significant effect of error speed on 2AFC performance, $\eta_g^2 = .021$ and 95% CI = [0.0002, 0.07]. Performance in fast error trials, $P(\text{correct} | \text{fast error}) = .64$, $SD = .17$, was depressed compared with performance in slow error trials, $P(\text{correct} | \text{slow error}) = .69$, $SD = .16$. Importantly, there was no interaction between condition and error speed, $\eta_g^2 = .0003$, 95% CI = [0.00, 0.02], revealing that the conditions did not differ significantly in the size of the effect of error speed on 2AFC performance (see Figure 1). Conducting two planned *t*-tests to evaluate the effect of error speed on 2AFC performance within each condition using an adjusted alpha of $\alpha = .025$ to correct for multiple tests revealed a significant effect of error speed in both the replication, $t(129) = 4.44$, $p < .001$, $d_z = 0.36$, and 95% CI = [0.15, 0.57], as well as the extension condition, $t(127) = 4.53$, $p < .001$, $d_z = 0.52$, and 95% CI = [0.27, 0.76]. In both conditions, performance in fast error trials, $P(\text{correct} | \text{fast error, replication}) = .63$, $SD_{\text{Rep}} = .11$, $P(\text{correct} | \text{fast error, extension}) = .62$, $SD_{\text{Ext}} = .12$, is worse than in slow error trials, $P(\text{correct} | \text{slow error, replication}) = .68$, $SD_{\text{Rep}} = .10$, $P(\text{correct} | \text{slow error, extension}) = .68$, $SD_{\text{Ext}} = .11$.

In addition to these effects central to our research question, there was a significant effect of trial type, $\eta_g^2 = .090$ and 95% CI = [0.03, 0.17], driven by more incorrect responses in 2AFC trials combining a miss and correct

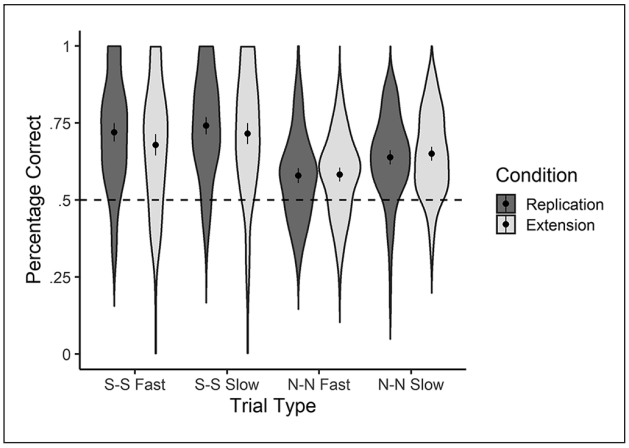

Figure 1. Mean (dots), 95% confidence intervals (whiskers), and distributions of percentage correct in two-alternative forced-choice performance in critical trials per trial type and condition (replication vs. extension). S–S denotes trials consisting of a previous false alarm and hit (both responses "studied"), N–N trials denotes trials consisting of a previous miss and correct rejection (both responses "not studied"). The category fast combines trials with reaction times (RTs) faster than median RT and category slow combines trials with RTs equal to or slower than median RT. The dashed line represents chance performance.

rejection, P(correct | N–N) = .61, SD = .14, than in 2AFC trials combining a false alarm and a hit, P(correct| S–S) = .71, SD = .19. This might reflect the fact that misses result from unsuccessful encoding or retrieval, while false alarms only involve misleading retrieval. As noted by Starns et al. (2018), the diffusion model predicts higher accuracy on S–S than N–N trials with typical parameter values in recognition tasks. Furthermore, trial type interacted significantly with condition, $\eta_g^2 = .004$ and 95% CI = [0.00, 0.03]. Again, using an adjusted alpha of $\alpha = .025$ for two post hoc t-tests, the effect of trial type was significant within both the replication, $t(258) = 8.32$, $p < .001$, $d_z = 1.03$, and 95% CI = [0.77, 1.29], as well as the extension condition, $t(254) = 5.06$, $p < .001$, $d_z = 0.63$, and 95% CI = [0.38, 0.89]. In both conditions, there were more incorrect responses in N–N trials, P(correct | N–N trials, replication) = .61, SD_{Rep} = .10, P(correct | N–N trials, extension) = .62, SD_{Ext} = .09, than in S–S trials, P(correct | S–S trials, replication) = .73, SD_{Rep} = .13, P(correct | S–S trials, extension) = .70, SD_{Ext} = .15. Because participants tended to make only a few false alarms, all results involving S–S trials have to be interpreted with caution.

Bayesian hierarchical logistic regression

We now turn to analyses that use single-item error RT as a continuous predictor in a logistic regression instead of just classifying trials as "fast" or "slow." We used a hierarchical Bayesian approach that matches the one reported by Starns et al. (2018).

Model details. The model predicted 2AFC accuracy on each trial with the following equation

$$logit(p_{jk}) = i_j + es_j ert_{jk} + cs_j crt_{jk} + eci_j ert_{jk} crt_{jk}$$

where j indexes a particular participant and k indexes a particular trial. p denotes the probability of a correct response, i the intercept, es the slope for error RT, cs the slope for correct RT, and eci the interaction in error and correct RT slopes. ert_{jk} and crt_{jk} are z-scores of the log-transformed error and correct single-item RT for the forced-choice item with a previous error and correct response. Following Starns et al. (2018), we took log RTs to attenuate the positive skew in the distributions and converted them to z-scores to aid interpretation of the logistic slopes.[2] We z-transformed RTs for each participant, trial type, and previous response type (correct and incorrect) separately. To ensure stable estimates for participants with low trial counts for a particular type of error (e.g., someone who almost never made false alarms), the standard deviations for the z-scores were based on all error trials (for *ert*) or correct trials (for *crt*) from the respective participant pooled across targets and lures. Due to this procedure, all participants could be included in the Bayesian hierarchical logistic regression.

As in Starns et al. (2018), we estimated all of the logistic regression parameters separately for S–S and N–N trials. This corresponds to the following logistic regression equation for a given trial k within a given participant j

$$\begin{aligned} logit(p_{jk}) = {} & (i_j + r_{jk} di_j) + (es_j + r_{jk} des_j) ert_{jk} \\ & + (cs_j + r_{jk} dcs_j) crt_{jk} \\ & + (eci_j + r_{jk} deci_j) ert_{jk} crt_{jk} \end{aligned}$$

where all parameters starting with a d represent the difference in parameter values between N–N and S–S trials and r_{jk} codes the trial type for participant j on trial k with S–S coded as 0 and N–N coded as 1. (All other symbols have the same meaning as in the previous equation.)

Our model assumed that the logistic regression parameters followed Gaussian distributions across participants, and we estimated separate distributions for the replication and extension condition to evaluate the effect of condition. Each of the 8 participant-level parameters (4 logistic parameters × 2 trial types) had a separate mean (μ) and standard deviation (σ) defining this across-participant distribution for the two conditions (replication and extension). We index the across-participant parameters with subscripts corresponding to the type of regression parameter and superscripts for the condition; for example, μ_i^{Rep} is the mean of the across-participant distribution of intercept parameters in the replication condition collapsed over both trial types. We omit the condition superscript for estimates that combine information from both conditions.

Following Starns et al. (2018), we used uninformative priors on the parameters for the across-participant distributions: Gaussian distributions with $M = 0$ and $SD = 100$ for the μ parameters and uniform distributions from 0.05 to 10 for the σ parameters.[3] We used JAGS to sample from posterior distributions. For each model, we ran 4 MCMC chains that each supplied 2,500 samples from the posterior distribution after the results were thinned by taking every 100th sample. The goal of thinning was to ensure that effective sample size (ESS) for all critical parameters was similar to the number of samples stored in the output files due to a reduction of autocorrelations between two successive retained samples. ESS estimates for average logistic parameter values were at or above 8,427. The Gelman–Rubin statistic (Gelman & Rubin, 1992) was $\hat{R} \leqslant 1.01$ for all of the parameters we report (Brooks & Gelman 1998). We computed 90% posterior intervals (PIs) by taking the .05 and .95 quantiles of the posterior samples, and we discuss any effect with a PI that excludes zero (meaning that there was at least a 95% chance of a non-zero effect in the specified direction).

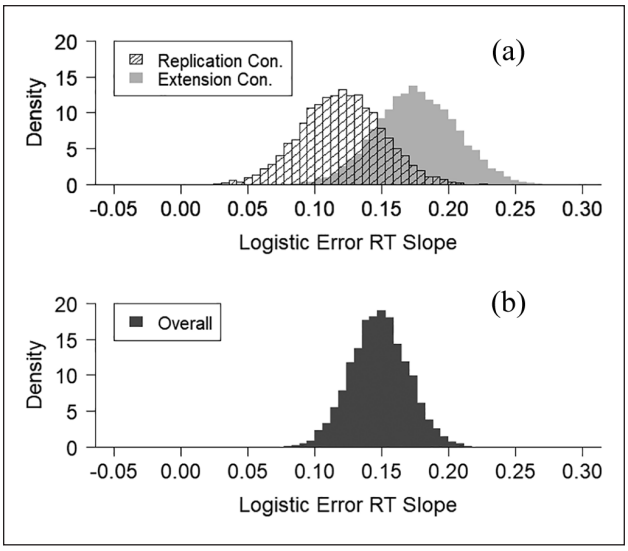

Figure 2. Posterior distributions for the average error RT logistic regression slope (μ_{es}) across participants separately for (a) replication and extension condition as well as (b) collapsed across groups.

Logistic results. The most critical results involve the slope on error RT, *es*. Figure 2a shows posterior distributions for the average error slope across participants in the replication (μ_{es}^{Rep}) and extension (μ_{es}^{Ext}) conditions. Both conditions provided strong evidence for a positive relationship between error RT and 2AFC accuracy, indicating that slower single-item error responses had a higher probability of a correct response in the subsequent 2AFC test. Contrary to our expectations, the sample provided some evidence that the error slope was lower in the replication condition, $Md_{Rep} = 0.12$, 90% PI = [0.07, 0.17], than the extension condition, $Md_{Ext} = 0.18$, and 90% PI = [0.13, 0.23]. However, the posterior interval on the slope difference included zero, $Md = -0.06$, and 90% PI [−0.13, 0.02], so we conclude that there was essentially no difference between the conditions. Most critically for our purposes, the results show clear evidence against the idea that the extension condition would eliminate or substantially reduce the error speed effect on subsequent 2AFC performance. Figure 2b shows the posterior for error RT slope collapsed across conditions (and trial type). The median, $Md = 0.15$, as well as the range of posterior slope values, 90% PI = [0.11, 0.18], indicates that the current error speed effect was very similar to the one reported by Starns et al. (2018) for Experiment 2 ($Md = 0.15$, 90% PI = [0.09, 0.21]).

In the following, we will discuss a few additional parameter results that were highlighted by Starns et al. (2018). For a complete depiction of the results, we refer interested readers to the OSF analysis files (https://osf.io/ejucx/). The results provided suggestive evidence that the error RT slope was lower for S–S trials than for N–N trials (posterior medians of $Md_{S\text{-}S} = 0.13$ and $Md_{N\text{-}N} = 0.17$, respectively), but the posterior interval on the difference included zero, 90% PI = [−0.11, 0.03]. Thus, there seems to be no meaningful trial-type effect on error RT slope. As in Starns et al. (2018), and as predicted by all known decision models, we observed a negative relationship between single-item recognition RT for correct responses and 2AFC accuracy. The parameter for the across-participant average slope on correct RT (μ_{cs}) indicated a clear negative relationship, posterior $Md = -0.26$ and 90% PI = [−0.30, −0.23]. This effect indicates that faster correct responses on single-item recognition trials predicted higher 2AFC accuracy when the item reappeared in a 2AFC trial. As with the error slope, the correct slope was closer to zero for S–S than N–N trials (posterior medians of $Md_{S\text{-}S} = -0.23$ and $Md_{N\text{-}N} = -0.30$, respectively). The 90% posterior interval on the difference excluded zero, but included some values that could be considered functionally null [−0.15, −0.003]. Therefore, we cannot make a confident inference that there is a meaningful trial-type effect on correct RT slope.

Posterior distributions for the across-participant average intercept coefficient (μ_i) were higher for S–S trials, $Md_{S\text{-}S} = 0.93$, and 90% PI = [0.87, 1.00], than N–N trials, $Md_{N\text{-}N} = 0.45$, and 90% PI = [0.41, 0.49], with clear evidence that this was a meaningful difference, 90% PI on difference [0.40, 0.56]. These estimates correspond to the overall proportion correct on the 2AFC task of .72 and .61 for S–S and N–N trials, respectively. Starns et al. (2018) demonstrated that the diffusion model predicts higher performance for S–S trials with parameter values that are typical for recognition memory tasks. We will discuss the reasons behind it in the section "Discussion".

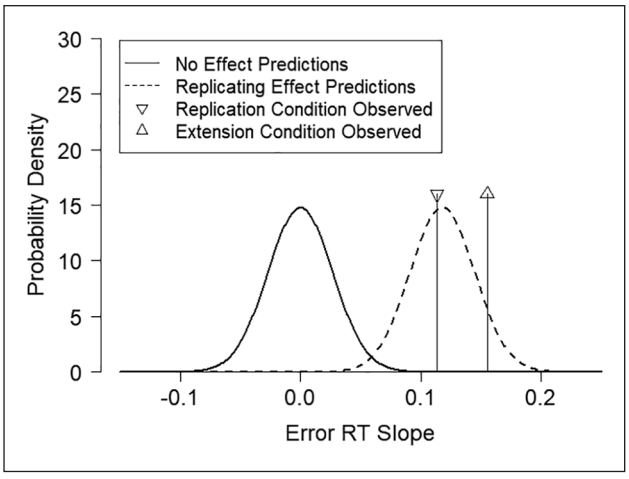

Figure 3. Distributions of predicted error RT slopes for a "no effect" hypothesis (solid line) as well as a "replicating effect" hypothesis (dashed line) assuming an effect size based on the Starns et al. (2018) results. Vertical lines mark the observed maximum likelihood slope coefficients for the replication and extension condition.

Exploratory inspection of predictions for logistic point estimates

Before collecting data, we created an additional OSF preregistration that also documented predictions for the logistic slope on error RT (https://osf.io/ejucx/). The predictions came from model simulations informed by the logistic regression results from Starns et al. (2018). Simulation code and detailed methods are available on OSF. The simulation used a signal detection model to determine the number of 2AFC trials for each simulated participant, which we assumed would produce a realistic distribution of trial counts across participants. To check this assumption, we reran the prediction simulations using the actual trial counts for each participant in the current experiment. This produced nearly identical results compared with the original simulations.

Figure 3 shows the distribution of predicted error RT slopes for the "no effect" and "replicating effect" hypotheses, with vertical lines marking the observed slope coefficients. The coefficients were produced by pooling trials across participants and getting the maximum likelihood estimate for the slope parameter, which is the same method used for each simulated experiment in the prediction simulations. The results from both conditions were much more consistent with predictions based on a "replicating effect" than a "null effect" hypothesis. The degree of support for the two hypotheses can be quantified by taking the likelihood ratio from the prediction distributions at the observed slope value. The slope estimate in the replication condition was about 5,000 times more likely for the "replicating effect" hypothesis than the "no effect" hypothesis, and the estimate in the extension condition was over a million times more likely under the "replicating effect" hypothesis than the "no effect" hypothesis. Thus, the results provide very strong evidence against a null effect in both conditions. These results mirror the conclusions from the Bayesian modelling, but they also show that we anticipated the magnitude of the error slope effect.

Discussion

The aim of the present study was to evaluate whether the effect of error speed in a single-item recognition response on the accuracy of a subsequent 2AFC task is based on misleading memory information or evoked through an error-correction strategy. Introducing an extension condition in which confounding aspects of the experimental design were eliminated, we replicated the effect of error speed in a single-item recognition response on accuracy in a subsequent 2AFC task applying two different statistical methods: one using error RT as a dichotomous factor and another using error RT as a continuous predictor. Both tests indicated that faster errors in a single-item recognition task go along with a decreased subsequent 2AFC performance in trials composed of a previous error and correct response. Based on these results, we can conclude that the effect of error speed is not eliminated by de-emphasising error correction and adding forced-choice trials with two words that were both correctly classified on the earlier single-item trials. This point is perhaps best supported by Figure 3, which shows that the observed logistic slope in the extension condition was much more consistent with our predictions for a replicating effect of error speed than for a null effect of error speed. This result also indicates that the finding is not unique to the particular methods of the Starns et al. (2018) study, and thus provides evidence against the idea that the effect is produced by an RT-based strategy for correcting error responses. However, it is also possible that the methods in the extension condition did not disrupt the error-correction strategy as we intended.

Starns et al. (2018) originally investigated the occurrence of the error speed effect as a critical test to evaluate whether recognition errors can be based on misleading evidence in addition to ambiguous evidence (i.e., failed retrieval). In the following we will elaborate on the different mechanisms allowing an error speed effect.

Misleading evidence as source of the error speed effect

In the framework of the diffusion model (Ratcliff, 1978), the occurrence of the error speed effect can be explained by a higher amount of misleading information elicited by fast errors compared with slow errors (Starns et al., 2018). The idea is that the value of familiarity is accumulated over time for each word. The more evidence exists either to the correct or to the incorrect response, the faster a response is elicited. Thus, both the speed of errors and the speed of correct responses in a single-item recognition task should

be predictive of the probability of a correct response in a subsequent 2AFC trial that includes items from the single-item recognition task (Starns et al., 2018). This can be concluded as the same accumulating mechanisms underlie correct and incorrect responses. But, contrarily to error RTs, faster correct responses in a single-item recognition task should result in a higher accuracy on subsequent 2AFC task. Although we did not include speed of correct responses as a factor in our ANOVA, correct RTs were included in the logistic regression. Evaluating the slope of correct RTs, the effect appears in the theoretically meaningful direction. This strengthens the assumption that the error speed effect is induced through the accumulation of misleading information from memory for recognition decisions.

Additional to the error speed effect, Starns et al. (2018) predicted based on the diffusion model an interaction between error speed and trial type as long as the proportion of avoidable errors is lower for targets than for lures.[4] A higher proportion of avoidable errors should lead to a higher proportion of 2AFC trials that can generally be corrected and thus to a higher overall performance as well as a smaller error speed effect. Therefore, next to an overall smaller performance, which is evident in our study through the main effect of trial type as well as the lower intercept, they predicted the error speed effect to be higher in 2AFC trials involving a previously misclassified target (N–N trials) than in trials involving a previously misclassified lure (S–S trials). Starns et al. found no evidence for a difference of the error speed effect across trial types and noted this as a failed prediction of the diffusion model. In our study, the interaction slightly failed to reach significance in the ANOVA (see Table 2) and the posterior intervals for the interaction within the logistic regression included zero. Nevertheless, the effect went in the direction predicted by the diffusion model. The failure to detect the effect might be caused by the small size of the effect. However, other mechanisms also impact a possible interaction as an increase in memory for targets due to multiple presentation, once in the study phase and once in the single-item recognition test, and a misleading memory trace for lures in the 2AFC task created by the previous appearance in the single-item recognition task.

Mistaken detection as source of the error speed effect

In the framework of discrete-state models, the most prominent example, the 2HTM, cannot account for the error speed effect as mentioned already in the introduction and would require modification as explained in the next section. This is because in the 2HTM all recognition errors within a single-item recognition task result from the same underlying uncertainty state (Province & Rouder, 2012; Snodgrass & Corwin, 1988). RTs in the framework of RT-MPTs (RT multinomial processing trees) are explained as the sum of the process times leading to the respective response as well as encoding and response execution times (Klauer & Kellen, 2018). Thus, within this framework, RTs can only reflect response certainty as long as different underlying mental processes can result in the same response which is given for correct (they can result out of correct detections or correct guesses) but not for incorrect responses. Therefore, the 2HTM can account for the relationship between correct RTs in single-item recognition and subsequent 2AFC performance but not for the relationship with error RTs.

Starns et al. (2018) discussed a version of discrete-state models which can in principle account for the error speed effect, the 2LTM. This model incorporates mistaken detection as well as incorrect guessing as possible mechanisms underlying recognition errors. And because two different processing paths can underlie errors in this model, the 2LTM can account for the error speed effect in the framework of RT-MPTs: Assuming that detection is on average faster than not detecting and guessing, as it was demonstrated in Klauer and Kellen (2018), regardless whether it is a mistaken or a correct detection, fast errors should result mostly out of a mistaken detection while slow errors should result mostly out of incorrect guesses.

In addition, within the 2LTM the proportion of correct 2AFC responses in fast error trials allows some conclusions about the amount of mistaken detections. If fast errors consist only of mistaken detection, performance should be at or below chance performance, because incorrectly detected items can only be corrected in subsequent 2AFC responses if paired with a correctly detected item. However, some mistaken detections might result within the subsequent 2AFC task in a different discrete state, guessing or even correct detection, increasing the probability of a corrected 2AFC response. In addition, we used a median split to categorise fast and slow errors. Thus, if the proportion of mistaken detection is low, fast errors will also include incorrect answers based on guessing, lifting performance on 2AFC trials above chance levels, while still being lower than for slow errors, if and when mistaken detection at least occasionally appears in the 2AFC trials. As is evident from Figure 1, 2AFC performance is above chance performance for most participants for both fast misses (N–N trials) as well as fast false alarms (S–S trials) revealing that the proportion of mistaken detection seems to be low.

Discrete versus continuous mediation of recognition decisions

Regarding the ongoing discussion of whether recognition decisions are mediated through discrete states or are based on a continuous memory-strength signal, no clear conclusions can be drawn based on our results as both the diffusion model, through misleading evidence, as well as the 2LTM, through mistaken detection, can account for the present results. Nevertheless, the 2HTM in its current form has difficulties explaining the error speed effect. The error speed effect indicates that some targets (lures) can elicit strong misleading evidence from memory suggesting

misleadingly that they are lures (targets). This leads to fast errors in the single-item recognition test and to low accuracy in the subsequent 2AFC test. The 2HTM does not include a mechanism for misleading memory evidence, so that the 2HTM seems falsified.

However, even under the 2HTM, evidence from memory would be expected to be misleading for targets not attended to during study. Targets which were not attended to during study and as a consequence were not encoded in memory should be expected to function as lures despite their nominal status as targets. As a result, such targets would elicit misleading evidence in favour of their being lures, whatever the model that mediates recognition decisions. Similarly, considering lures, target-lure similarity could function as a comparable mechanism for lures. Lures being sufficiently similar to studied targets could elicit a response from memory similar to that evoked by targets (Brainerd & Reyna, 1990) and hence elicit misleading evidence in favour of their being targets, whatever the model that mediates recognition decisions. For such reasons, it would be interesting for future research to investigate whether these two manipulations, attention during study and/or target-lure similarity, moderate the error speed effect for targets and lures, respectively.

Conclusion

We found a clear relationship between the speed of error responses in single-item recognition and accuracy in forced-choice recognition in both the replication and extension condition. Thus, the results are consistent with the predictions of the diffusion model and the 2LTM, which assume that fast errors tend to be associated with stronger misleading evidence than slow errors, but are not consistent with the 2HTM without additional assumptions. The results also show that de-emphasising error correction and including forced-choice trials does not eliminate the error-speed effect, suggesting that the effect is not dependent on an error-correction strategy engendered by the Starns et al. (2018) paradigm.

Furthermore, regardless of whether the error speed is correlated with the strength of misleading memory information (continuous models) or the likelihood of mistaken detection (2LTM), the type of information used seems to be item-specific and stable across the two tasks to some extent. Thus, some targets seem to be processed similar to lures, perhaps because they were not attended to during study, and some lures seem to be processed similar to targets, perhaps because they are associated to other studied targets as per Brainerd and Reyna's (1990) gist memory. If this account is correct, manipulating target-lure similarity as well as participant's attention during study should be found to moderate the present effect of error speed.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: A.V. received support from the Deutsche Forschungsgemeinschaft (DFG; GRK 2277), Research Training Group "Statistical Modelling in Psychology" (SMiP). The preparation of this article was supported by Grant No. 1454868 from the National Science Foundation awarded to J.J.S.

ORCID iD

Anne Voormann https://orcid.org/0000-0002-5024-5116

Data accessability statement

The approved pre-registration protocol as well as all materials, experimental files, analysis scripts and raw data files are publicly available on OSF (https://osf.io/ejucx/).

Notes

1. This reasoning only holds for certain parameter constellations. Starns et al. (2018) assessed these parameter constellations within a simulation and found that there exists a correlation between the speed of errors and the two-alternative forced-choice (2AFC) performance only if the proportion of avoidable errors is low.
2. Unfortunately, Starns et al. (2018) neglected to mention the transformations that they applied to the reaction time (RT) data. Their data are publicly available at https://osf.io/w72pp/ and can be used to verify that the logistic results reported in the article are based on participant-level *z*-scores for log RTs. The OSF page for the current project also includes a pre-registered prediction simulation that used the same RT transformation (https://osf.io/ejucx/).
3. Starns et al. (2018) used uniform priors ranging from 0 to 10 for σ parameters. But because JAGS has sampling problems when allowing σ to range from 0 upwards, we adjusted the lower boundary while keeping the prior nearly as uninformative as in the original manuscript. In addition, we increased the number of final samples slightly, compared with Starns et al. (2018) to achieve more accurate parameter estimates.
4. As discussed by Starns et al. (2018), fits of the diffusion model to recognition data consistently show that targets have a higher proportion of drift rates in the "wrong" direction (towards the threshold for an incorrect response) than lures, both because the mean of the drift rate distribution tends to be closer to zero for targets than for lures and because the drift rate distribution tends to be more variable across trials for targets than for lures. This means that a higher proportion of errors for targets (targets called "new") are unavoidable with additional retrieval time compared with errors for lures (lures called "old"). When drifts are in the wrong direction, taking extra time just tends to produce stronger support for the wrong answer.

References

Banks, W. P. (1970). Signal detection theory and human memory. *Psychological Bulletin*, *74*(2), 81–99. https://doi.org/10.1037/h0029531

Batchelder, W. H., & Alexander, G. E. (2013). Discrete-state models: Comment on Pazzaglia, Dube, and Rotello (2013).

Psychological Bulletin, *139*(6), 1204–1212. https://doi.org/10.1037/a0033894

Blackwell, H. R. (1963). Neural theories of simple visual discriminations. *Journal of the Optical Society of America*, *53*(1), 129–160. https://doi.org/10.1364/JOSA.53.000129

Brainerd, C. J., & Reyna, V. F. (1990). Gist is the grist: Fuzzy-trace theory and the new intuitionism. *Developmental Review*, *10*(1), 3–47. https://doi.org/10.1016/0273-2297(90)90003-M

Bröder, A., & Schütz, J. (2009). Recognition ROCs are curvilinear: Or are they? On premature arguments against the two-high-threshold model of recognition. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *35*(3), 587–606. https://doi.org/10.1037/a0015279

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*(4), 434–455. https://doi.org/10.2307/1390675

Cary, M., & Reder, L. M. (2003). A dual-process account of the list-length and strength-based mirror effects in recognition. *Journal of Memory and Language*, *49*(2), 231–248. https://doi.org/10.1016/S0749-596X(03)00061-5

Dubé, C., Rotello, C. M., & Pazzaglia, A. M. (2013). The statistical accuracy and theoretical status of discrete-state MPT models: Reply to Batchelder and Alexander (2013). *Psychological Bulletin*, *139*(6), 1213–1220. https://doi.org/10.1037/a0034453

Duyck, W., Desmet, T., Verbeke, L. P. C., & Brysbaert, M. (2004). WordGen: A tool for word selection and nonword generation in Dutch, English, German, and French. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 488–499. https://doi.org/10.3758/BF03195595

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

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. John Wiley & Sons.

Kellen, D., & Klauer, K. C. (2011). Evaluating models of recognition memory using first- and second-choice responses. *Journal of Mathematical Psychology*, *55*(3), 251–266. https://doi.org/10.1016/j.jmp.2010.11.004

Kellen, D., & Klauer, K. C. (2014). Discrete-state and continuous models of recognition memory: Testing core properties under minimal assumptions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(6), 1795–1804. https://doi.org/10.1037/xlm0000016

Kellen, D., Klauer, K. C., & Bröder, A. (2013). Recognition memory models and binary-response ROCs: A comparison by minimum description length. *Psychonomic Bulletin & Review*, *20*(4), 693–719. https://doi.org/10.3758/s13423-013-0407-2

Kellen, D., Singmann, H., Vogt, J., & Klauer, K. C. (2015). Further evidence for discrete-state mediation in recognition memory. *Experimental Psychology*, *62*(1), 40–53. https://doi.org/10.1027/1618-3169/a000272

Klauer, K. C., & Kellen, D. (2015). The flexibility of models of recognition memory: The case of confidence ratings. *Journal of Mathematical Psychology*, *67*, 8–25. https://doi.org/10.1016/j.jmp.2015.05.002

Klauer, K. C., & Kellen, D. (2018). RT-MPTs: Process models for response-time distributions based on multinomial processing trees with applications to recognition memory. *Journal of Mathematical Psychology*, *82*, 111–130. https://doi.org/10.1016/j.jmp.2017.12.003

Lahl, O., Göritz, A. S., Pietrowsky, R., & Rosenberg, J. (2009). Using the World-Wide Web to obtain large-scale word norms: 190,212 ratings on a set of 2,654 German nouns. *Behavior Research Methods*, *41*(1), 13–19. https://doi.org/10.3758/BRM.41.1.13

Luce, R. D. (1963). A threshold theory for simple detection experiments. *Psychological Review*, *70*(1), 61–79. https://doi.org/10.1037/h0039723

Malejka, S., & Bröder, A. (2016). No source memory for unrecognized items when implicit feedback is avoided. *Memory & Cognition*, *44*(1), 63–72. https://doi.org/10.3758/s13421-015-0549-8

Pazzaglia, A. M., Dube, C., & Rotello, C. M. (2013). A critical comparison of discrete-state and continuous models of recognition memory: Implications for recognition and beyond. *Psychological Bulletin*, *139*(6), 1173–1203. https://doi.org/10.1037/a0033044

Province, J. M., & Rouder, J. N. (2012). Evidence for discrete-state processing in recognition memory. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(36), 14357–14362. https://doi.org/10.1073/pnas.1103880109

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108. https://doi.org/10.1037/0033-295X.85.2.59

Ratcliff, R. (2014). Measuring psychometric functions with the diffusion model. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(2), 870–888. https://doi.org/10.1037/a0034954

Ratcliff, R., & Starns, J. J. (2009). Modeling confidence and response time in recognition memory. *Psychological Review*, *116*(1), 59–83. https://doi.org/10.1037/a0014086

Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review*, *95*(3), 318–339. https://doi.org/10.1037/0033-295X.95.3.318

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, *117*(1), 34–50. https://doi.org/10.1037/0096-3445.117.1.34

Starns, J. J., Chen, T., & Staub, A. (2017). Eye movements in forced-choice recognition: Absolute judgments can preclude relative judgments. *Journal of Memory and Language*, *93*, 55–66. https://doi.org/10.1016/j.jml.2016.09.001

Starns, J. J., Dubé, C., & Frelinger, M. E. (2018). The speed of memory errors shows the influence of misleading information: Testing the diffusion model and discrete-state models. *Cognitive Psychology*, *102*, 21–40. https://doi.org/10.1016/j.cogpsych.2018.01.001

Starns, J. J., Hicks, J. L., Brown, N. L., & Martin, B. A. (2008). Source memory for unrecognized items: Predictions from multivariate signal detection theory. *Memory & Cognition*, *36*(1), 1–8. https://doi.org/10.3758/MC.36.1.1

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(3), 582–600. https://doi.org/10.1037/0278-7393.26.3.582