Bias in confidence: A critical test for discrete-state models of change detection

Samuel Winiger

University of Zurich

Henrik Singmann

University of Zurich

University of Warwick

David Kellen

Syracuse University

Author Note

Samuel Winiger and Henrik Singmann, Department of Psychology. David Kellen, Department of Psychology. Authors received support from the Swiss National Science Foundation Grant 100014_165591. We thank Nathalie Rieser for her help with data collection. We also thank Robert M. Nosofsky and an anonymous reviewer for their many valuable comments. All materials, data, and modeling scripts are available on the Open Science Framework: https://osf.io/es2rw

Correspondence should be sent to David Kellen, Department of Psychology, 430 Huntington Hall, NY 13244. (Email: davekellen@gmail.com).

Abstract

Ongoing discussions on the nature of storage in visual working memory have mostly focused on two theoretical accounts: On one hand we have a discrete-state account postulating that information in working memory is supported with high fidelity for a limited number of discrete items by a given number of "slots", with no information being retained beyond these. In contrast with this all-or-nothing view, we have a continuous account arguing that information can be degraded in a continuous manner, reflecting the amount of resources dedicated to each item. It turns out that the core tenets of this discrete-state account constrain the way individuals can express confidence in their judgments, excluding the possibility of *biased confidence judgments*. Importantly, these biased judgments are expected when assuming a continuous degradation of information. We report two studies showing that biased confidence judgments can be reliably observed, a behavioral signature that rejects a large number of discrete-state models. Finally, complementary modeling analyses support the notion of a mixture account, according to which memory-based confidence judgments (in contrast with guesses) are based on a comparison between graded, fallible representations, and response criteria.

Keywords: Visual Working Memory, Change Detection, Critical Test, Discrete-State Models, Confidence

Research in working memory is concerned with our ability to hold and maintain representations of information over a short amount of time. This ability is closely associated with key human faculties such as reasoning (Süß, Oberauer, Wittmann, Wilhelm, Schulze, 2002) and text comprehension (Daneman & Merikle, 1996), and has predictive value in important domains such as academic achievement (e.g., Bayliss, Jarrold, Gunn & Baddeley, 2003). In recent years, considerable efforts have been made in the study of working memory in the visual domain, with particular focus on its *capacity* and *storage mode*. At this point, it is well established that visual working memory (VWM) has limited capacity, in the sense that there is an upper limit in the amount of information that one can maintain in working memory at a given time (e.g., Cowan, 2001). There is, however, an ongoing discussion concerning the way information can be stored. This discussion has focused mostly on two theoretical accounts: On one hand we have *discrete-state* or *slot models*, which assume that items are either stored in memory with high precision (each is stored in a *slot*) or not at all (e.g., Luck & Vogel, 1997; Rouder et al., 2008; Zhang & Luck, 2008). On the other hand, we have *continuous resource models* postulating that information can be degraded in a more graceful manner, with the quality of each representation in VWM being determined by the amount of resources dedicated to it (e.g., Bays & Husain, 2008, van den Berg, Shin, Chou, George, & Ma, 2012; Wilken & Ma, 2004).

Like in many other research domains, the comparison of models of VWM is often predicated on a quantification of their ability to fit the *entire* data coming from some experimental design. These fits are made possible through a number of auxiliary assumptions, some parametric (e.g., latent distributions are Gaussian), others more substantive (e.g., processes are selectively influenced by certain experimental manipulations). Despite its successful track record, this approach to model comparison raises a number of concerns (Birnbaum, 2011; Kellen, 2019): For instance, a violation of any of the auxiliary assumptions made is likely to compromise the conclusions of a model-comparison exercise. This possibility is nothing more than the famous Duhem-Quine thesis (Duhem, 1954; Quine, 1963): Consider a theory $\mathcal{T}$, that along with

auxiliary assumptions $\mathcal{A}$, makes the observable prediction $\mathcal{O}$. The failure to observe $\mathcal{O}$ (i.e., $\bar{\mathcal{O}}$ is observed instead) does not imply a rejection of $\mathcal{T}$, given that $\mathcal{A}$ might be at fault. For example, in the context of response-time modeling, Jones and Dzhafarov (2014) showed that the ability of diffusion and ballistic accumulator models to successfully describe people's responses is entirely dependent on a number of auxiliary distributional assumptions. The critical role of $\mathcal{A}$ is also reflected in the (necessary) tinkering that takes place during model development (for a discussion and examples, see Shiffrin & Nobel, 1997).

Also challenging is the fact that goodness-of-fit measures, even when corrected for model flexibility, do not necessarily privilege the portions of the data that are most informative from a theoretical standpoint. This issue has been discussed at length by prominent theoreticians such as Rozeboom, who argued for the importance of determining the empirical support for each of the different formal propositions that constitute a theory, rather than looking at omnibus support measures (e.g., Rozeboom, 1970, 2008).[1] To illustrate the point being made here, let us consider a couple of notable observations coming from physics and biology: 1) clocks on satellites orbiting the Earth run differently from clocks on Earth (e.g., Burns, 2017), 2) there are humans with two blood types (i.e., blood-group chimeras; Dunsford et al., 1953). Both observations have important theoretical implications. But because they are very specific or rare, they are likely to be downplayed in any model-comparison exercise that considers the fit to the "entire" data on time measurement or human genetics, along with a premium on model parsimony. There is a real possibility that cruder models that cannot accommodate these specific observations might end up striking a better overall compromise between parsimony and fit. For instance, imagine a scenario in which one would make the case for 'first-generation' globalist models of memory, despite their inability to account for null list-strength effects (Ratcliff, Clark, & Shiffrin, 1990), on the grounds that they can account for a number of other effects in the literature at the

[1] "... astute evidence appraisal focuses on select features of the hypothesis at issue with only secondary confidence adjustments, if any, in its remainder. Holistic acceptance/rejection is for amateurs." (Rozeboom, 2008, p. 1123).

time (ca. the 1990s) *and* are simpler than its competitors.[2]

In light of these concerns, it is important for researchers to also consider some of the alternatives available in their toolboxes. The goal of the present work is to do so by comparing models of VWM using a *critical-test approach* (e.g., Birnbaum, 2008, 2011; Kellen & Klauer, 2014, 2015; Stephens, Dunn, & Hayes, 2018). The idea behind it is very simple: Identify a specific prediction that contrasts different *families of* models, and restrict all testing efforts to it. The end result is a strong inference that speaks directly to the theoretical commitments of each model (Platt, 1964). In some cases, the sets of *permissible outcomes* associated with different families of models, let us say $\mathcal{O}_{\mathcal{M}_1}$ and $\mathcal{O}_{\mathcal{M}_2}$, are *mutually exclusive* (e.g., Birnbaum, 2008). What this means is that the set of permissible outcomes for one family of models corresponds to the set of *forbidden outcomes* of a competing family and vice-versa. In other cases, the permitted outcomes of one family are a *subset* of the permitted outcomes of another; e.g., $\mathcal{O}_{\mathcal{M}_1} \subset \mathcal{O}_{\mathcal{M}_2}$. These differences have important implications in the way researchers engage in critical testing: In the first case, the only concern is to ensure that our experimental design will not yield observations at the boundary of the different mutually-exclusive predictions (e.g., Birnbaum, 2008). In the second case, researchers need to place their efforts towards reliably observing outcomes that are outside the subset $\mathcal{O}_{\mathcal{M}_1}$ (for an excellent example, see Stephens et al., 2018). In either case, it is possible to dismiss a broad families of models and shift our focus towards more promising options.

The remainder of this paper is organized as follows: First, we will discuss discrete-state and continuous models of VWM in the context of one of the main experimental paradigms used to compare them, the *change-detection task* (e.g., Luck & Vogel, 1997; Rouder et al., 2008; Wilken & Ma, 2004). We will then discuss the auxiliary assumptions that are typically made in this context, and propose a new critical test that does not require them. At the locus of this test is the specific way these models handle confidence judgments and the possibility of *biased confidence*, which we will define later on. We will then report two experiments studies (one of them

[2]For purely rhetorical purposes, let us assume without further discussion that more recent models such as REM (Shiffrin & Steyvers, 1997) are more complex or flexible.

pre-registered) that show the presence of biased confidence judgments, which are forbidden by discrete-state models. For the sake of readability, we will first discuss somewhat simplified versions of the discrete-state and continuous model, and later on show that the results of the critical test also hold across a number of more complex model variants. Finally, we will report complementary model fits showing that, among the remaining candidate accounts, there is support for a mixture account in which memory slots provide graded, fallible representations.

Continuous and Discrete-State Models of the Change-Detection Task

In each trial of the change-detection task, illustrated in Figure 1, participants study a distributed array of items (e.g., squares) that vary on a single dimension such as color. After a brief presentation and some delay, one item location is probed with a test item and participants have to judge whether the color of that item has changed relative to the previous presentation, responding either "`same`" or "`change`". Thus, in any change-detection task there are at least two trial types, Change trials (with correct response "`change`") and Same trials (with correct response "`same`"). In addition, researchers typically manipulate the number of squares presented in a given trial (e.g., Luck & Vogel, 1997). It is also common to manipulate the proportion of Same and Change trials across different test blocks (e.g., Donkin, Nosofsky, Gold, & Shiffrin, 2013; Donkin, Tran, & Nosofsky, 2014; Rouder et al., 2008). In a few cases, participants are

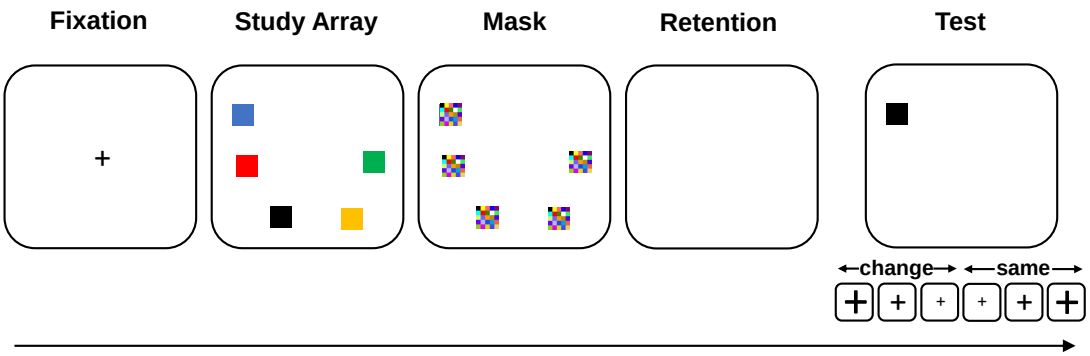

Figure 1. Illustration of the procedure of the change-detection task in visual working memory. Depicted is a change trial. Note that an eight-point confidence rating scale was used in the experiments reported here (the German verbal labels were `unterschiedlich` (`change`), and `gleich` (`same`). Snapshots of one of the test trials can be found in the Supplemental Materials.

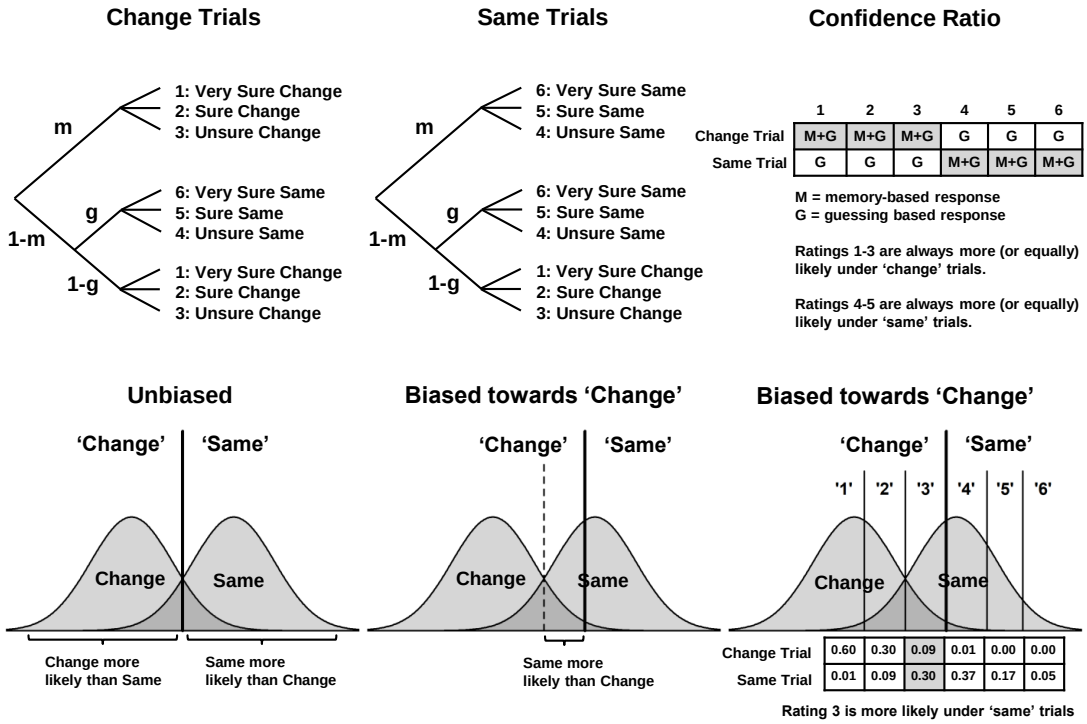

Figure 2. Top row: Discrete-state model for Change and Same trials, and the mixture components associated with each confidence level. In both trees, parameter m denotes the probability of an item being stored in working memory (in a slot), whereas parameter g denotes the probability of guessing "`same`". The confidence levels are described by one of two components, with '$M + G$' indicating that a given response confidence level in a given type of trial can be reached via both memory and guessing processes, and G indicating that it can only be reached via guessing. For clarity, the state-response mapping parameters ξ and γ (associated with memory and guessing states, respectively) are omitted. Bottom row: Illustration of the continuous resource model under unbiased and biased response criteria (criteria correspond to the vertical solid lines). The likelihood ratio associated with a given response corresponds to the average relative height of the two distributions within the region associated with that response. It is shown how the model can predict biased confidence judgments (e.g.,"3: Unsure Change" is more likely under Same trials), something that the discrete-state model cannot do. Note that the Gaussian distributions and the equally-spaced criteria are merely illustrative; they are not required for the discussed predictions to hold. For clarity, models for a six-point confidence rating scale are shown in the figure, whereas an eight-point confidence rating scale was used in the experiment reported here.

also requested to indicate how confident they are in their responses (e.g., Wilken & Ma, 2004; Ricker, Thiele, Swagman, & Rouder, 2017; see also Rademaker, Tredway, & Tong, 2012; van den Berg, Yoo, & Ma, 2017).

According to the *discrete-state model* illustrated in the top row of Figure 2, the tested item is stored in memory with probability m. Because the item is stored with high precision, a correct response is always expected.[3] With probability $1 - m$, no

[3]Please note that this assumption of high precision is only plausible when applying the model to experimental designs where the stimuli are highly discriminable (e.g. distinct color changes in Change trials, such as blue $\longrightarrow$ red). As discussed later on, our experiments were designed in order to make this assumption plausible. Later on, we will discuss the implications of relaxing this high-precision assumption (despite the experimental design) in different ways.

information about the item is stored and a guess has to be made, with response "`same`" being made with probability g, and response "`change`" being made with probability $1 - g$. The discrete-state model can be extended to the case of confidence ratings by introducing confidence-mapping parameters for memory-based ($\boldsymbol{\xi}$) and guessing-based responses ($\boldsymbol{\gamma}$). These mapping parameters are traditionally assumed to be conditionally independent, such that their values do not depend on the value of m — all that matters is the discrete state one is in, not the probability of entering such state (for discussions, see Kellen & Klauer, 2015; Klauer & Kellen, 2010).

In the case of the *continuous resource model*, illustrated in the bottom row of Figure 2, the information available for a tested item can be represented as a sample from a latent-strength distribution, one for cases in which the item has changed, and another one for when it is the same. Both distributions are established on a latent 'memory-strength' scale. Individuals judge the tested item by comparing its value with a response criterion τ, responding "`same`" when the value is larger than it, otherwise responding "`change`". The continuous models can also be easily extended to accommodate confidence ratings. Specifically, one can introduce additional criteria $\boldsymbol{\tau}$, such that the intervals defined by them are mapped onto different levels of the confidence-rating scale (see Figure 2).

Response-Bias Manipulations and ROCs

The motivation behind the manipulating of the proportion of Same and Change trials is the belief that it selectively influences participants' response biases, which are captured in each model by either the guessing probability g or the response criterion τ. Under this selective-influence assumption, we can derive clear predictions from both models about the way that the probability of "`same`" responses in Same and Change trials covaries across values of g or τ. These predictions are commonly referred to as the models' *Receiver Operating Characteristic* (ROC) functions (Green & Swets, 1966; Kellen & Klauer, 2018). We illustrate these ROCs in Figure 3. Binary-response ROCs have been used in a number of VWM studies (e.g., Donkin et al., 2013; 2014; Rouder et

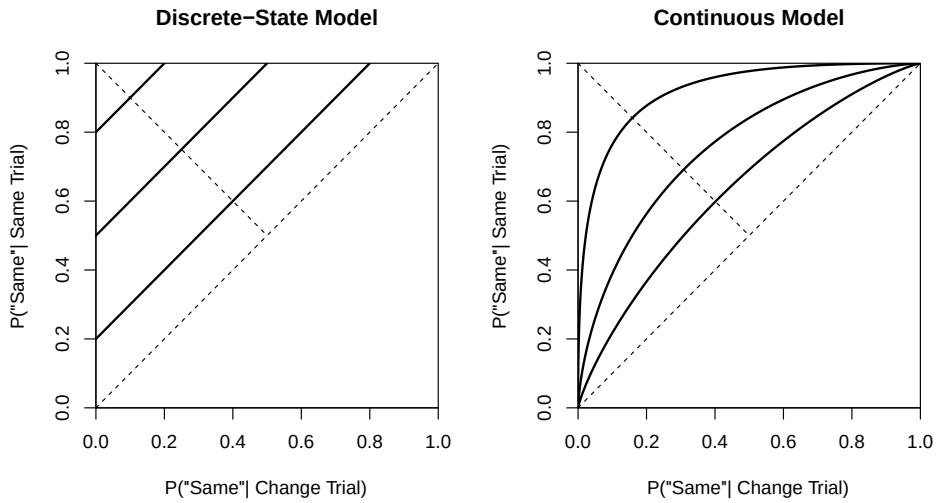

Figure 3. Illustration of different Receiver Operating Characteristic (ROC) functions predicted by the discrete-state and continuous models for binary response data (i.e., no confidence rating responses).

al., 2008). Of course, it is entirely possible that response-bias manipulations also affect memory processes; i.e., there is no selective influence (e.g., Balakrishnan, 1999; Van Zandt, 2000; see also Ashby, 1983; Diederich & Busemeyer, 2006; Ratcliff, 1981). If that is the case, then it is no longer possible to distinguish between both models based on the shape of the binary-response ROCs.

Confidence Ratings and ROCs

ROCs can also be constructed using confidence-ratings. Instead of being based on binary response proportion across different response biases, the ROC is constructed from the cumulative distributions across the confidence scale, from maximum-confidence "`change`" to maximum-confidence "`same`". Wilken and Ma (2004) compared continuous and discrete-state models under the assumption that memory-based responses in the latter are deterministically mapped onto *maximum-confidence* judgments, which enforces the prediction of linear confidence-rating ROCs. The reliance on this assumption turns out to be fatal, given that it has been thoroughly dismissed by a number of theorists (Bröder & Schütz, 2009; Erdfelder & Buchner, 1998; Falmagne, 1985; Krantz, 1969; Luce, 1963; Malmberg, 2002; Kellen & Klauer, 2014, 2015; Klauer & Kellen, 2010; Province & Rouder, 2012). Allowing memory-based responses to include the range of the confidence scale that is consistent with the binary judgment enables the

model to accommodate the curved confidence-rating ROCs that are typically observed.[4]

More recently, Ricker et al. (2017) proposed a new comparison relying on the notion of conditional independence, using a modified change-detection task. Instead of judging whether the color at test was the same presented before in that specific location (see Figure 1), participants were asked to choose between two colors. Also, they manipulated choice difficulty, such that the two colors presented were more or less distinct. Ricker et al. found that the confidence ratings obtained across choice difficulty levels did not conform with the assumption of conditional independence, therefore speaking against a discrete-state account. The problem with Ricker et al.'s conclusion is that it is questionable whether one could even begin to assume that conditional-independence holds in their experimental paradigm. As discussed by Kellen and Klauer (2015, p. 547), it is only reasonable to assume that conditional independence holds when dealing with experimental paradigms in which there are no 'external features' informing the participant of the difficulty level of any given test trial. For example, in the domain of recognition memory, Kellen and Klauer (2015) focused their testing of conditional-independence on study-strength manipulations that were not identifiable during the test phase. In each test trial, participants were shown a word, without any indication of whether it might have been studied once or thrice (see also Province & Rouder, 2012). In the case of Ricker et al.'s paradigm, these external features are obviously present, as they correspond to difference between the two color alternatives.

A New Critical Test: Biased Confidence

The issues found with previous work using either binary-response or confidence-rating ROCs indicate the need for an alternative comparison approach that:

1. does not assume selective influence in response-bias manipulations,

2. does not impose deterministic mappings on confidence judgments, and

[4]More specifically, the discrete-state model can predict piecewise linear ROCs that can capture the finite ROC data points collected in a given experiment.

3. does not compromise the assumption of conditional independence.

These requirements are achieved by the critical test proposed here, which is based on the notion of biased confidence judgments (Balakrishnan, 1999). In order to understand what this bias is, let us first consider the continuous resource model: As illustrated in Figure 2, confidence judgments result from the comparison between the latent-strength of a test item and a set of ordered confidence criteria $\boldsymbol{\tau}$. The position of each criterion relative to the latent distributions determines the likelihood of each item type given a certain response and confidence level. The introduction of a response bias, for example towards responding "`change`", introduces the *possibility* of biased confidence judgments, especially when confidence is at a minimum, and minimum confidence covers a narrow range of strength values (Balakrishnan, 1999).[5] This possibility follows from the continuous model's core notion that confidence judgments are based on the segmentation of a latent-strength scale by confidence criteria. In the specific example given in Figure 2 (lower row, right panel), a minimum-confidence "`change`" response is more likely to occur in a Same trial. This confidence level can be said to be biased as the respondent would improve their accuracy if they simply reassigned all their minimum-confidence "`change`" responses to minimum-confidence "`same`" instead (i.e., bundled all their minimum-confidence responses on the "`same`"-side of the scale; see Balakrishnan, 1999).

In contrast, the discrete-state account *precludes* the possibility of biased confidence judgments. Confidence judgments result from the mapping of the different discrete-states onto a confidence scale, with the mapping of memory- and guessing-based responses being established by parameters $\boldsymbol{\xi}$ and $\boldsymbol{\gamma}$ respectively (for a discussion, see Klauer & Kellen, 2010). For instance, consider the probability of

[5] In other words, the observation of a response bias is *necessary but not sufficient* for the observation of biased confidence judgments.

response "`change`" made with minimum confidence:

$$\begin{aligned} P(\text{“}\texttt{change}_{\text{min}}\text{”} \mid \text{Change trial}) &= m \cdot \xi_{\text{min}} + (1-m) \cdot (1-g) \cdot \gamma_{\text{min}}, \\ P(\text{“}\texttt{change}_{\text{min}}\text{”} \mid \text{Same trial}) &= (1-m) \cdot (1-g) \cdot \gamma_{\text{min}}. \end{aligned}$$

Because the former probability cannot be smaller than the latter, confidence judgments *cannot* be biased (see Figure 2, top row, right panel). This inability to predict biased confidence judgments stems from the core notion within the discrete-state theory that, in the absence of stored information on the target stimulus, responses are invariably based on the same guessing process (i.e., conditional independence).

Continuous and discrete-state models can be compared by testing for the presence of biased confidence judgments. The discrete-state model only permits the inequalities

$$\begin{aligned} P(\text{“}\texttt{change}\text{”}, \texttt{conf} = i \mid \text{Change trial}) \geq P(\text{“}\texttt{change}\text{”}, \texttt{conf} = i \mid \text{Same trial}), \\ P(\text{“}\texttt{same}\text{”}, \texttt{conf} = i \mid \text{Same trial}) \geq P(\text{“}\texttt{same}\text{”}, \texttt{conf} = i \mid \text{Change trial}), \end{aligned}$$

for all i among possible confidence levels, whereas the continuous model imposes no such constraint. What this means is that the set of permissible outcomes of the discrete-state model is a subset of the permissible outcomes of the continuous model. Our task then is to attempt to reliably observe cases in which these inequalities are violated; i.e., try to obtain data belonging to the discrete-state model's set of forbidden outcomes. Given our understanding of the continuous model, we expect the occurrence of forbidden outcomes to be most likely when (a) individuals show a clear bias towards one binary response, and when (b) individuals seldom make *minimum-confidence* judgments (see also Balakrishnan, 1999). These expectations are illustrated in Figure 2, which shows the occurrence of biased confidence judgments for minimum-confidence "`change`" judgments when there is a bias towards responding "`change`". In light of these expectations, we will focus our analyses on *minimum-confidence* judgments.

Experiment 1

We conducted a change-detection task experiment in which we attempted to observe biased confidence judgments. To ensure that the test was applied in conditions where items are expected to be stored with high fidelity (in line with the discrete-state model assumptions), we relied on highly dissimilar colors. Similar to previous studies (e.g., Rouder et al., 2008), biases in binary responses were encouraged by manipulating the proportion of Change trials across blocks (75% vs. 25%). All data, scripts, and materials are available on the Open Science Framework: https://osf.io/es2rw

Participants

A total of 44 participants took part in the experiment. Our plan was to collect at least 40 participants, in order to roughly match the sample sizes used in previous critical tests comparing continuous and discrete-state models (e.g., 45 participants in Kellen & Klauer, 2014). We slightly overshot our intended target number as data collection was performed by a research assistant in a day-wise manner with as many participants as possible per day (i.e., on the second to last day of data collection we had not reached 40 participants, so we decided to continue for one more day).

The average age of our participants was 23.8 years, ranging from 18 to 29 (SD = 3.7). In exchange for their participation, participants received CHF 15 or course credit. Each session took about 50 minutes. All participants reported having normal or corrected-to-normal vision and normal color vision.

Stimuli and Apparatus

Our stimuli and presentation generally followed Donkin et al. (2013). We used a set of ten highly dissimilar colors (white, black, red, blue, green, yellow, orange, cyan, purple, and dark-blue-green). These colors were taken from Table 5 of Donkin et al. (colors with suffix "−1"). Importantly, note that these colors are expected to yield "large" Change trials (see Donkin et al., p. 891). The discrete-state model's assumption that memory-based judgments are always accurate is assumed to be reasonable here

(see Footnote 2; see also Nosofsky & Donkin, 2016). Stimuli were presented within a light gray rectangle of approximately 9.8×7.3 degree visual angle. Stimuli were 0.75×0.75 degrees in size. Participants were seated approximately 60 cm away from the screen and no chin rest was used. The position of each stimulus was chosen randomly with the constraint of a minimal distance of 2 degrees from other stimuli and the screen center (measured from the center of the stimuli).

Procedure

The experiment was comprised of a practice block with 20 trials using a confidence-rating scale, followed by eight blocks with 52 trials each also using a confidence-rating scale, and one last block of binary-response trials. In the practice block and the final binary-response block, half of the trials were Change trials, whereas in the remaining blocks using confidence ratings they were either 75% or 25% (4 blocks with each proportion). The biased blocks were randomized, with the constraint that the same proportion of Change trials did not occur more than twice in a row. Before each block, participants were informed about the percentages of Change and Same trials. In each trial, the percentages of Change trials as well as the percentage of Same trials were displayed in the labels shown above the confidence-rating scale (see figure in the Supplemental Materials).

Each trial followed the structure outlined in Figure 1: Each trial started with a fixation cross that was presented for 1,000 ms. An array of five square stimuli was then presented for 500 ms, followed by a blank screen for 500 ms. After the blank screen, a multicolored checkerboard-like mask was presented at each stimulus location for 500 ms. The test phase of each trial was self-paced: A test item was presented at a random stimulus location. Its color was either the same that was presented at the beginning of the trial at this location (Same trial), or a color that was shown at the beginning of the trial, but at a different location (Change trial). Participants were then asked to decide whether they are in a Change trial or a Same trial, while simultaneously stating their confidence by clicking on a response button on an eight-point confidence rating scale.

For each half of the scale, confidence was represented by a plus sign (+) increasing in size from low to high confidence. A verbal label above each half of the scale clearly indicated the binary choice (i.e., "`change`" versus "`same`"; the actual German words used were "`unterschiedlich`" versus "`gleich`"). Participants did not receive feedback on their performance.

Results

Manipulating the proportion of Change trials succeeded in affecting individuals' binary-response bias, although they were generally biased towards responding "`change`" across blocks. In blocks with 75% and 25% Change trials, the average proportion of "`change`" responses was .78 and .40, respectively. In the last block with 50% Change trials and binary judgments, the average proportion of "`change`" responses was .62. Overall, these proportions indicate a general bias toward "`change`" responses irrespective of experimental condition, although the base-rate manipulation appears to have produced an effect in the expected direction. This impression was corroborated using signal-detection bias measure c, which turned out to be more strict in 75% change condition (mean = 0.58, SD=0.33) than in the 25% Change condition (mean = 0.17, SD = 0.39; Wilcox's $W = 941$, $p < .0001$). The left panel of Figure 4 shows the confidence-rating ROC functions obtained with the grouped confidence ratings, for each of the bias conditions. Note that the point obtained in the binary-response condition falls very close to the ROCs. The ROCs take on concave and symmetrical shapes, as typically found in this type of data.[6] Given these results, our hope to observe biased confidence judgments is limited to the 75% Change condition. After all, this is the only condition in which we observe a strong response-bias effect, with minimum-confidence "`change`" responses being seldom made (see also Balakrishnan, 1999). For the sake of completeness, Figure 5 (first and second column) illustrates the results obtained in the 25% Change condition, but we will refrain any further discussion (for further details, see the Supplemental Materials).

As shown in Figure 5 (top row, third panel), the observed proportions of

[6] As a reminder, the discrete-state model can accommodate curved confidence-rating ROC data.

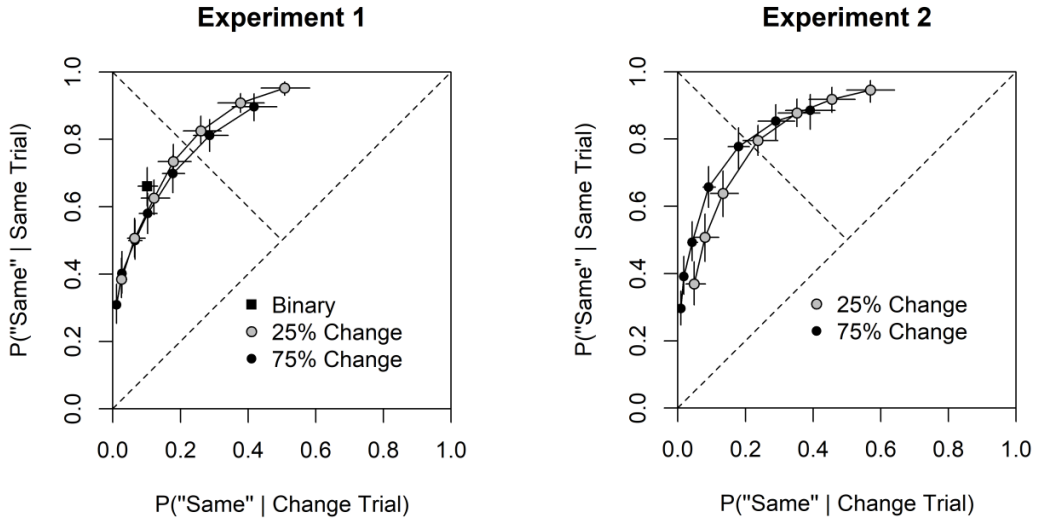

Figure 4. Circles show the Receiver Operating Characteristic (ROC) functions obtained with the grouped data from both bias conditions. The square shows the results from the binary condition. The 95% confidence intervals associated with each point were obtained via non-parametric bootstrap.

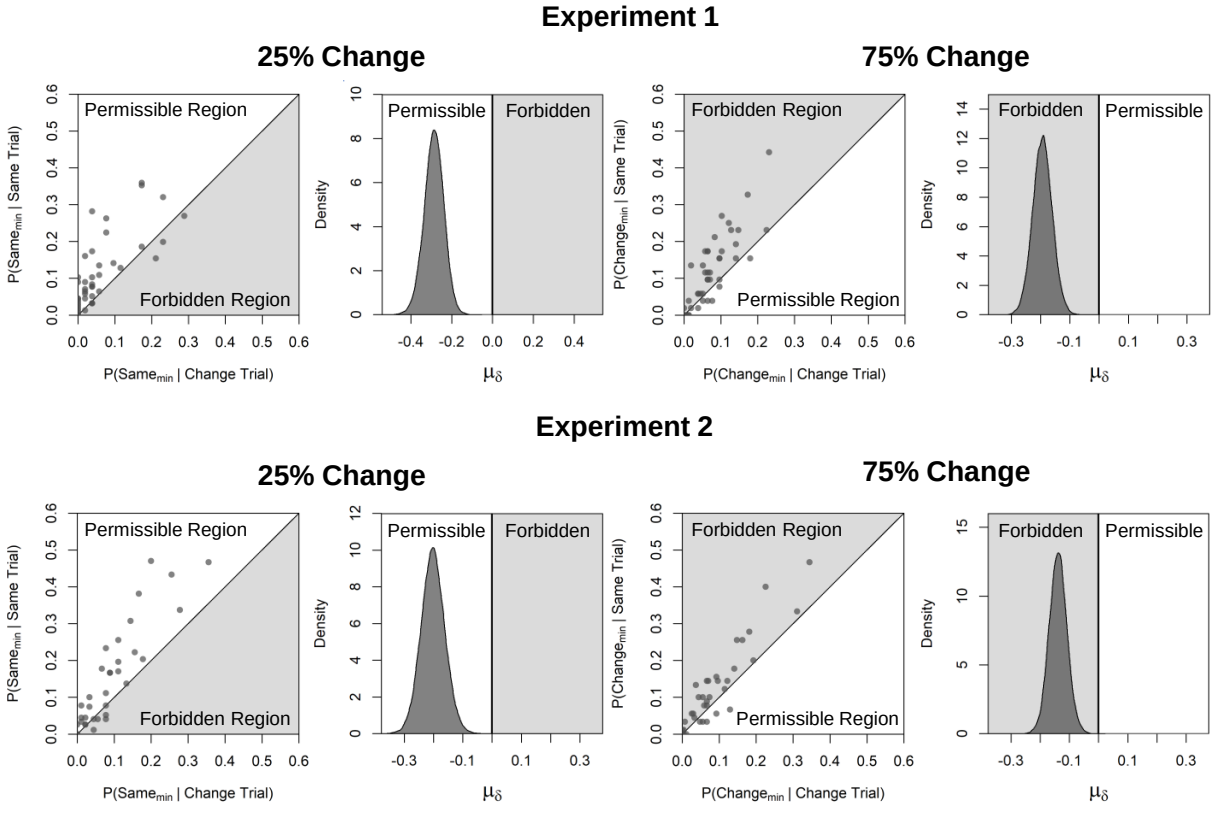

Figure 5. The scatterplots show the observed individual proportions of minimum confidence "`same`" responses (in the 25% Change condition) or "`change`" responses (in the 75% Change condition) in Change and Same trials. The density plots show the group-level posterior distribution of the difference parameter μ_δ. In all plots, the shaded areas correspond to the forbidden regions of the discrete-state model.

minimum-confidence "`change`" responses in the 75% Change condition were generally greater in Same trials, indicating the presence of a bias in confidence that is forbidden under the discrete-state model. Specifically, P("`change`$_{\text{min}}$" | Same trial) was greater/equal/smaller than P("`change`$_{\text{min}}$" | Change trial) in 68%, 11%, and 20% of the individual cases, respectively. At the group level, P("`change`$_{\text{min}}$" | Same trial) and P("`change`$_{\text{min}}$" | Change trial) were .12 and .07, respectively.

The individual minimum-confidence responses were fitted with a joint-binomial hierarchical-Bayesian model. In a Bayesian framework, the uncertainty regarding model parameters is represented via probability distributions, with *prior* distributions being updated into *posterior* distributions (via Bayes' theorem) in light of the data. The key component of the joint binomial model adopted were the probabilities P("`change`$_{\text{min}}$" | Same trial) and P("`change`$_{\text{min}}$" | Change trial), whose difference was captured by an effect-size parameter δ with mean μ_δ (for details, see the Supplemental Materials on Open Science Framework). The range of predictions permitted under the discrete-state model correspond to a μ_δ that is greater or equal to zero, indicating that the minimum-confidence "`change`" responses are more likely in Change trials than in Same trials. The discrete-state model's 'forbidden region', which is expected under the continuous resource model, corresponds to a negative μ_δ, indicating that such responses are more likely in Same trials than in Change trials. The visual inspection of the confidence-rating proportions in Figure 5 was corroborated by the joint binomial model's posterior parameter estimates, which yielded a posterior mean of μ_δ of -0.19, with a 95% credibility interval of [-0.26, -0.13] (see Figure 5, top row, right-most panel).

The evidence for the discrete-state model's forbidden predictions relative to its permitted predictions can be quantified by means of a Bayes Factor (BF), which in the present case can be computed via the following ratio of posterior probabilities (Klugkist, Kato, & Hoijtink, 2005):

$$\text{BF} = \frac{P_{\text{post}}(\mu_\delta < 0)}{P_{\text{post}}(\mu_\delta > 0)}.$$

Values larger than 1 support the presence of biased confidence judgments, which are not

permitted under the discrete-state model, whereas values between 0 and 1 indicate support for the absence of such bias. For the 75% change trials condition, we obtained a Bayes Factor of roughly 60,000 (as none of the 60,000 samples was $\mu_\delta > 0$), indicating strong support for the presence of biased confidence ratings.

One potential objection to the analysis reported so far is that, whereas many of the minimum confidence responses fall into the 'forbidden region' of the discrete-state model, the magnitude of these violations might be comparatively small and still compatible with a discrete-state model when taking sampling variability into account. In order to corroborate the obtained Bayes factor, we conducted a *posterior-predictive test* (see Gelman & Shalizi, 2013; Klauer, 2010). Specifically, we generated one set of synthetic data, $\boldsymbol{x}_{\text{synth}}$, per posterior parameter sample $\hat{\theta}_{\text{post}}$ from a joint-binomial model constrained to follow the predictions of the discrete-state model (i.e., μ_δ was constrained to be non-negative). We then computed the misfit of the true data, $\mathcal{L}(\boldsymbol{x}_{\text{obs}}, \hat{\theta}_{\text{post}})$, as well as the misfit of the synthetic data, $\mathcal{L}(\boldsymbol{x}_{\text{synth}}, \hat{\theta}_{\text{post}})$, according to the expectations derived from the posterior parameter sample. The latter can be understood as the expected amount of misfit if the only source of noise is sampling variability (and the true data generating process respects the constraints of the discrete-state model). We used these quantities to compute Bayesian p-values by estimating the probability of the misfit of the real data being larger than of the synthetic data; i.e., $P(\mathcal{L}(\boldsymbol{x}_{\text{obs}}, \hat{\theta}_{\text{post}}) > \mathcal{L}(\boldsymbol{x}_{\text{synth}}, \hat{\theta}_{\text{post}}))$. Small p-values indicate a poor fit of the model to the data. We obtained a Bayesian p-value of .003, which indicates that the observed data is very much at odds with the discrete-state model's range of predictions, even when taking sampling variability into account. Importantly, when applying the same procedure to a joint-binomial model constrained to produce biased confidence judgments, the obtained Bayesian p-value was .52, consistent with the fact that a confidence bias is generally present in the data.

Discussion

Although our results provide clear evidence against the discrete-state account, a number of concerns can be raised. First, our focus on minimum-confidence judgments might appear somewhat post-hoc and even self-serving. After all, we focused our analyses on the minimum-confidence judgments coming from the 75% Change condition, based on the fact that it was the only condition in which we observed a clear response-bias effect in the expected direction. Instead, shouldn't one engage in joint evaluation of the evidence for biased confidence judgments across all condition? No, not at all. The first thing one should keep in mind is the nested (and thus asymmetric) relationship between the two theoretical accounts in the present experimental setup: One account includes the *possibility* of biased confidence ratings, whereas the other one does not. The failure to observe response biases or biased confidence ratings only indicates that both accounts are sufficient (i.e., the data are not diagnostic).[7] In contrast, the observation of biased confidence judgments is highly diagnostic, as it cannot be accounted by a variety of discrete-state models (as will be shown later on). Given this state of affairs, our focus should be directed towards the experimental conditions that seem most promising, namely the minimum-confidence judgments in a condition showing strong response biases (see Balakrishnan, 1999). The same strategy is found in other critical tests, where the focus is placed on individuals or groups showing specific preferences or successful study-strength effects (see Birnbaum, 2008; Kellen & Klauer, 2015; Kellen, Steiner, Davis-Stober, & Pappas, 2020). A joint analysis including non-diagnostic data would be a counter-productive move, akin to tempering one's inferences coming from the observation of black swans by keeping tabs on all the white ones encountered along the way.

Another concern is that participants might not have fully understood the large differences between colors, which might have compromised their performance. This

[7]Note that sufficiency results are relevant when dealing with two theoretical accounts that differ in terms of their "ontological" complexity, such as single-process versus two-process accounts. Specifically, one can use them to make the case that there is no need to assume two-process accounts given that their single-process counterparts survived every attempt to reject them (for a discussion, see Stephens et al., 2018). However, this is not the case here.

possibility is not implausible, especially given our failure to observe a bias towards "`same`" responses in the 25% Change condition. In light of these concerns, it is advisable to conduct a follow-up (pre-registered) experiment in which we explicitly try to minimize potential misunderstandings and are more likely to observe a response bias in the 25% Change condition.[8]

Experiment 2

We implemented two major changes in our experimental design: First, we manipulated the proportion of Change and Same trials between subjects. The goal was to increase the possibility of observing a bias towards "`same`" responses in the 25% Change condition. Second, to avoid any misunderstanding regarding the large color differences in our study, participants first engaged in a training block in which they received feedback after every response. Participants also received feedback at the end of every block.

Participants

A total of 73 participants took part in the experiment. We aimed for a total of 80 participants but did not reach that goal by a predefined date deadline. This resulted in 36 participants in the 75% change trial condition and 37 participants in the 25% change trial condition. The average age of our participants was 24.5 years, ranging from 18 to 35 (SD = 3.90). In exchange for their participation, they received CHF 15 or course credit. Experimental sessions took about 50 minutes. All participants reported having normal or corrected-to-normal vision and normal color vision.

Procedure

The experiment started with a practice block of 40 trials in which only binary responses were requested, followed by ten test blocks, also with 40 trials, in which responses were given using a confidence-rating scale. In the practice block, half of the

[8]The pre-registration can be viewed at https://osf.io/4gh9e. The pre-registered data collection and analysis plan can be found in the Supplemental Materials.

trials were Change trials; Also, participants received feedback after each trial in the form of a green checkmark '✓' (correct response) or a red '×' (incorrect), presented at the center of the screen. In the remaining blocks, the proportion of Change trials were either 75% or 25%, depending on the participant's condition. In contrast with Experiment 1, this proportion stayed the same throughout the experiment (i.e., the proportion of Change trials was manipulated between subjects). Participants did not receive feedback after each trial anymore. However, after each block, they were reminded of the actual proportion of Change trials in the experiment together with their proportion of "`change`" responses in that block (see Dube & Rotello, 2012).

Results

In conditions with 75% and 25% Change trials, the average proportion of "`change`" responses was .77 and .34, respectively. In terms of response-bias measure c, we found a (weak) bias towards "`same`" responses in the 25% change condition (mean = −0.08, SD = 0.34, $W = 482$, $p = .05$)[9] and a (stronger) bias towards "`change`" responses in the 75% change condition (mean = 0.47, SD = 0.36, $W = 3$, $p < .0001$). Figure 4 shows the ROC functions obtained with the grouped confidence ratings from both bias conditions. Once again, they show the expected curvilinear, symmetrical shape.

As shown in Figure 5 (lower row, right panels) we again find evidence for biased confidence judgments in the 75% condition. The observed proportions of minimum-confidence "`change`" responses were generally greater in Same trials, indicating the presence of a bias that is not permitted under a discrete-state model. Specifically, P("$\texttt{change}_{\text{min}}$" | Same trial) was greater/equal/smaller than P("$\texttt{change}_{\text{min}}$" | Change trial) in 75%, 5%, and 20% of the individual cases, respectively. At the group level, P("$\texttt{change}_{\text{min}}$" | Same trial) and P("$\texttt{change}_{\text{min}}$" | Change trial) were .12 and .09, respectively.

The individual minimum-confidence responses in the 75% condition were fitted

[9]Given that the bias towards "`same`" responses in the 25% Change condition is relatively weak, making the observation of biased confidence judgments unlikely (see also Footnote 6). And indeed, no bias was found. We decided to omit these analyses from the main text, and report them in the Supplemental Materials. In any case, we report the relevant data in Figure 5.

with the same joint-binomial hierarchical-Bayesian model used in Experiment 1. As a reminder, the discrete-state account expects the difference parameter μ_δ to be greater or equal to zero. Parameter estimates yielded a posterior mean of μ_δ of -0.14, with a 95% credibility interval of [-0.20, -0.08] (see Figure 5). We obtained a Bayes Factor of roughly 30,000 (only two of the 60,000 samples were $\mu_\delta > 0$), indicating again strong support for the presence of biased confidence ratings. This result was again corroborated by means of posterior predictive tests. For the model that is constrained to follow the predictions of the discrete-state account, we obtained a Bayesian p-value of .002. For the model that can produce biased confidence judgments, we obtained a Bayesian p-value of .40.

Establishing the Scope of the Critical Test

At this point, it is not clear whether the implications of our test results are circumscribed to an overly simplistic set of models, especially in the case of the discrete-state model. This discussion is especially relevant given that both of our studies only used five-item arrays, a number that does not surpass many people's working-memory capacity (Cowan, 2001). This means that processes other than guessing are playing a major role. The purpose of this section is to provide some clarification on this matter, and show that the model predictions discussed above hold across a number of possible extensions and/or modifications. We will also discuss one discrete-state model variant that can account for biased confidence judgments and show that it outperforms a "pure" continuous counterpart.

Adopting Non-Gaussian Distributions in the Continuous Model

The illustration of the continuous model given in Figure 2 assumes that memory-strength distributions are Gaussian with equal variance. It is reasonable to ask whether the ability to predict biased confidence ratings requires specific distributional assumptions. *It does not.* The possibility of biased confidence only requires the ability to establish a pair of confidence criteria along an interval of latent values in which one stimulus type is more likely than the other. Note that differences in likelihood are implied by the mere observation of above-chance performance, as it implies that values

larger than the binary criterion τ are more likely under Same trials, and values below τ are more likely under Change trials.[10]

Relaxing the Discrete-State Model's Confidence Mapping

The discrete-state model can account for the observed biased confidence judgments if we relax the way in which each discrete state is mapped onto the confidence scale (e.g., Malmberg, 2002). So far, we have assumed that the memory state reached with probability m is always mapped onto the levels of the confidence-rating scale associated with one of two binary responses (see also Klauer & Kellen, 2010, 2015; Kellen, Singmann, Vogt, & Klauer, 2015; Province & Rouder, 2012). Note that by doing so we are not introducing any additional constraints to the model – we are merely retaining the assumption made in the case of binary choices.

We do not see this extension as a convincing way to salvage discrete-state models. Specifically, we do not find any justification for not imposing the constraints already in place when dealing with binary choices. Also, such a relaxation implies that individuals are willing to go against their memories in order to respond in conformity with the biases promoted by the experimental manipulation (e.g., responding 'same' in a block with 25% Change trials despite detecting a change), while simultaneously being willing to respond counter to the same biases when *guessing* (e.g. guessing 'change' in a block with 25% Same trials). Such an account is directly at odds with the idea that response biases affect first and foremost guesses (see Erdfelder, Küpper-Tetzel, & Mattern, 2011; Krantz, 1969; Luce, 1963). Moreover, we note that by assuming a relaxed mapping, proponents of discrete-state models can no longer resort to ROC shapes as a source of

[10]Note that the assumption that ROC functions are concave (which includes linearity as a boundary case) implies that the likelihood ratio is monotonic. Any point of the ROC with slope larger/smaller than 1 indicates that the value of binary-response criterion τ is more likely under the latent distribution associated with Same/Change trials (for details, see Kellen, Winiger, Dunn, & Singmann, 2019). To obtain biased confidence ratings, one only needs to place the confidence criteria associated with a "`change`"/"`same`" response along a range of values in which the latent memory-strength values are more likely under the distribution associated with Same/Change trials.

empirical support, given that the models no longer make clear predictions.[11]

Swap Errors in the Discrete-State Model

One limitation of our discrete-state model is that it omits the possibility of participants incorrectly associating one of the other studied colors with the test position — a so-called *swap error* (Bays, Catalao & Husain 2009; Wheeler & Treisman, 2002). Let w denote the conditional probability of a swap error, given that the correct color was not remembered. In the case of Same trials, swap errors will always lead to a "`change`" response, as the color has to be different than the one presented at test. In contrast, in Change trials, the probability of the remembered color mismatching the one at test is $\frac{3}{4}$ (see Donkin, Tran, & Le Pelley, 2015). The equations for minimum-confidence "`change`" judgment are then:

$$P(\text{“}\texttt{change}_{\text{min}}\text{”} \mid \text{Change trial}) = m \cdot \xi_{\text{min}} + (1-m) \cdot w \cdot \frac{3}{4} \cdot \xi_{\text{min}} + (1-m) \cdot (1-w) \cdot (1-g) \cdot \gamma_{\text{min}},$$

$$P(\text{“}\texttt{change}_{\text{min}}\text{”} \mid \text{Same trial}) = (1-m) \cdot w \cdot \xi_{\text{min}} + (1-m) \cdot (1-w) \cdot (1-g) \cdot \gamma_{\text{min}}.$$

Based on these equations, we can see that a biased confidence rating is only expected when $\frac{m}{(1-m)} < \frac{w}{4}$, which only holds when $m \leq .20$. The upper boundary $m = .20$ requires that $w = 1$, which implies that there are no guessing-based responses, only swap errors. The required m values are extremely low — with a set size of five items, they would imply a discrete capacity (with successful binding) of at most one item. Such a characterization is not consistent with the overall performance observed in both experiments (see the ROC plots in Figure 5).

Color-Position Binding Failures

Now, let us consider an extended model according to which a person can also make a judgment based on the fact that they remember the color presented at test from

[11] It is worth pointing out that Malmberg's (2002) case for a relaxed mapping of memory states hinges on the imposition of additional constraints on the mapping of guesses. Specifically, Malmberg assumed that guesses are distributed uniformly across confidence-rating scale. Without this constraint, one no longer needs to relax the mapping of memory-based responses to predict curvilinear ROCs (see Klauer & Kellen, 2010, 2015).

earlier in the trial but not its exact location:[12]

$$P(\text{“}\texttt{change}_{\text{min}}\text{”} \mid \text{Change trial}) = m \cdot \xi_{\text{min}} + (1-m) \cdot c \cdot z \cdot \omega_{\text{min}} + (1-m) \cdot (1-c) \cdot (1-g) \cdot \gamma_{\text{min}},$$

$$P(\text{“}\texttt{change}_{\text{min}}\text{”} \mid \text{Same trial}) = (1-m) \cdot c \cdot z \cdot \omega_{\text{min}} + (1-m) \cdot (1-c) \cdot (1-g) \cdot \gamma_{\text{min}}.$$

with c denoting the probability that the color at test is remembered (but not its location), z as the probability that this color is *not* attributed to the test item location, and ω_{min} corresponding to the probability of mapping this judgment onto a minimum-confidence “`change`” response. Once again, it is easy to see that $P(\text{“}\texttt{change}_{\text{min}}\text{”} \mid \text{Change trial})$ cannot be smaller than $P(\text{“}\texttt{change}_{\text{min}}\text{”} \mid \text{Same trial})$, therefore excluding the possibility of biased confidence.

Assuming Varying Discrete-State Probabilities

One way to accommodate biased confidence ratings is to allow for probability m to differ between Change and Same trials (i.e., establish m_{change} and m_{same} parameters instead of a single m parameter). For instance, the confidence bias observed in the 75% Change condition would be accounted for if $m_{\text{change}} > m_{\text{same}}$. One explanation for these probability differences is that the differential expectations induced by base-rate manipulation affects sensory processing (e.g., Summerfield & de Lange, 2014). Although we cannot provide a clear-cut evaluation of this hypothesis using the present data, we can nevertheless point out that it yields predictions that do not seem to pan out in when inspecting the binary-response ROC data coming from change-detection tasks. Specifically, the discrete-state model no longer expects *linear and symmetric* binary-response ROCs when m_{change} and m_{same} are allowed to differ across response-bias conditions. To illustrate this point, we considered three possibilities:

1. m_{change} and m_{same} vary in opposite directions across response-bias conditions, such that $m_{\text{change}} > m_{\text{same}}$ when there is a bias towards “`change`” responses and $m_{\text{change}} < m_{\text{same}}$ when there is a bias towards “`same`” responses.

[12] As a reminder, note that the color tested in Change trials was always presented in the study array of the same trial in one of the other locations.

2. m_{change} varies across response-bias conditions, but m_{same} remains fixed (the inequalities stated above are also expected).

3. m_{same} varies across response-bias conditions, but m_{change} remains fixed (the inequalities stated above are also expected).

As shown in the left panel of Figure 6, the first possibility results in *curved* binary-response ROCs, which mimic the predictions of the continuous model illustrated in Figure 3. The prediction is somewhat implausible given that binary-response ROCs obtained in change-detection paradigms are generally well captured by a linear and symmetric function (e.g., Donkin et al., 2014; Rouder et al., 2008). If this first possibility held, then one would expect the extant ROC data to generally favor the continuous model. In turn, the center and right panels of Figure 6 show that the second and third possibilities would lead to *asymmetric* binary-response ROCs. These asymmetries are at odds with the symmetric functions predicted by both discrete-state and continuous models (see Figure 3). They are also somewhat implausible given that to the best of our knowledge, ROC data coming from change-detection tasks have generally been well accounted for by models assuming symmetrical functions (whether they are continuous or discrete). One counterargument is that it is entirely possible for a small asymmetry to be masked by the noise usually found in binary-response ROC data (for a review, see Kellen, Klauer, & Bröder, 2013). But then again, this noise would have to somehow leave the prediction of biased confidence ratings unaffected.

Allowing for fallible memory representations in the discrete-state model

Our discrete-state model assumes that items stored in VWM are represented with high fidelity, such that responses based on them are always accurate. It is not difficult to see how this assumption can be unreasonable in certain circumstances. For example, it implies the ability to detect the smallest differences between colors. Our first reaction to such criticism is to note that both experiments reported here relied on distinct colors, which should reduce the probability of memory-based errors. Moreover, to dismiss any concerning regarding the possibility of individuals misunderstanding the large difference

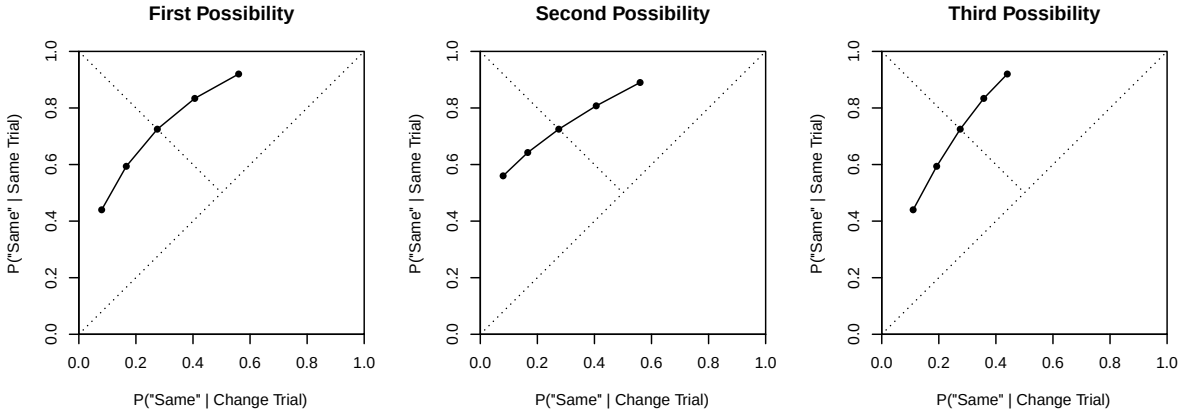

Figure 6. Illustrations of binary-response ROCs for three possibilities of varying parameters m_{change} and m_{same}. In the left panel, the parameter values are $m_{\text{change}} = [.600, .525, .450, .375, .300]$, $m_{\text{same}} = [.300, .375, .450, .525, .600]$, and $g = [.20, .35, .50, .65, .80]$, with each entry corresponding to one response-bias condition, going from a bias towards "`change`" to a bias towards "`same`". In the center panel, m_{same} is fixed to .45, whereas in the right panel, m_{change} is fixed to .45

between the colors, participants in Experiment 2 were given trial-by-trial feedback in an initial training phase.

But the key question remains: Could a somewhat fallible representation provide the discrete-state model with the ability to account for biased confidence judgments? The answer is *no*. To show this, we have to extend the model so that it includes the possibility of incorrectly remembering the color at a given position, while preserving the conditional-independence assumption (Rouder & Morey, 2009). Specifically, let $m^\star$ denote the probability that a remembered color in VWM is *correctly* determined to be different from the test item in a Change trial. Also, let $m^{\star\star}$ denote the probability that a remembered color in VWM is *incorrectly* determined to be different than the test item in a Same trial. Based on this extension, the equations for minimum-confidence "`change`" judgment correspond to:

$$\begin{aligned} P(\text{“}\texttt{change}_{\text{min}}\text{”} \mid \text{Change trial}) &= m \cdot m^\star \cdot \xi_{\text{min}} + (1-m) \cdot (1-g) \cdot \gamma_{\text{min}}, \\ P(\text{“}\texttt{change}_{\text{min}}\text{”} \mid \text{Same trial}) &= m \cdot m^{\star\star} \cdot \xi_{\text{min}} + (1-m) \cdot (1-g) \cdot \gamma_{\text{min}}. \end{aligned}$$

It follows that biased minimum-confidence judgments are only expected when the probability of an accurate memory representation in VWM in a Change trial, $m^\star$, is *smaller* than the probability $m^{\star\star}$ of an inaccurate memory representation in a Same

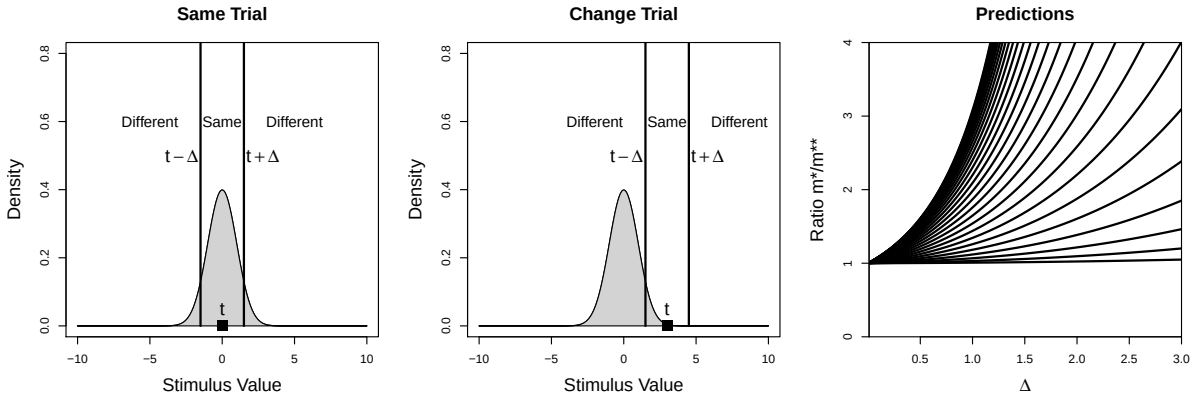

Figure 7. Example illustration of graded, fallible memory representation and its predictions. The densities characterize the fallible memory representations, whereas t corresponds to value of the test item.

trial; i.e., $\frac{m^\star}{m^{\star\star}} < 1$. What this means is that the memory representations provided by the discrete-state model would have to predict *below-chance* accuracy, which is not to be expected under any reasonable circumstances. To show this, let us consider a case in which both $m^\star$ and $m^{\star\star}$ are determined by the comparison of noisy representations and test items using a pair of perceptual thresholds. More specifically, let us assume that the noisy representation x of a given studied item is characterized by the normal distribution illustrated in Figure 7. When a test item T with value t is presented, the color is deemed to be the same as the one stored in VWM if x falls within the interval $[t - \Delta, t + \Delta]$, and different otherwise. The right panel of Figure 7 shows the ratio $\frac{m^\star}{m^{\star\star}}$ under different values of Δ and t. This ratio is larger than 1 across all the different values considered, therefore failing to show the condition necessary for the model to predict biased confidence judgments.

Casting Working-Memory Slots Through Mixture Signal-Detection Modeling

Finally, a discrete-state model extension that can account for biased confidence judgments in a plausible way assumes that people's responses consist of a mixture of guesses and memory-based judgments, the latter being based on a comparison between graded and fallible memory representations with response criteria (e.g., Nosofsky & Gold, 2018; Xie & Zhang, 2017; see also Keshvari, van den Berg, & Ma, 2013; Zhang &

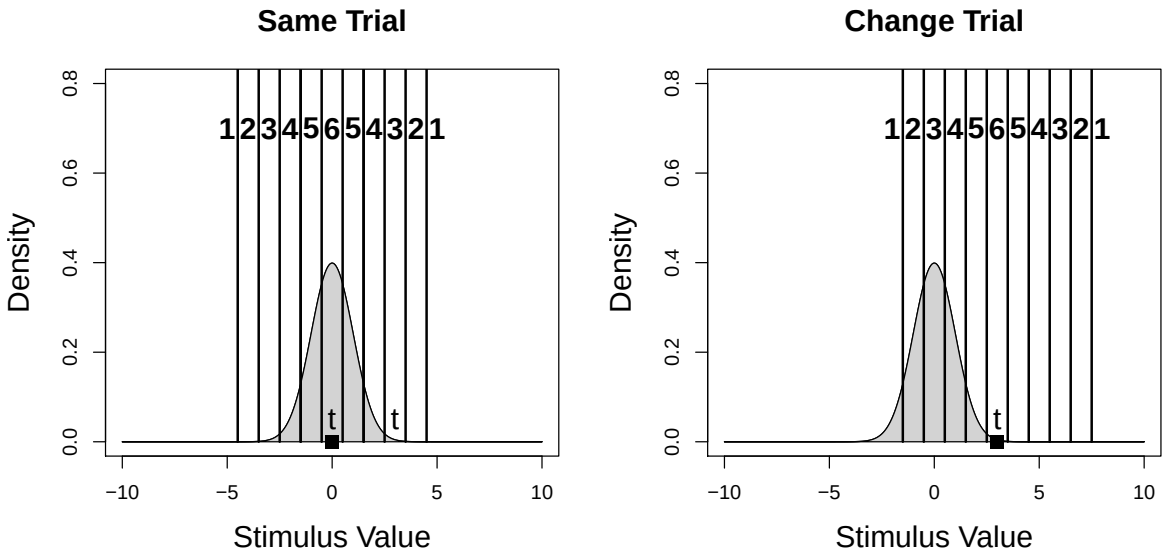

Figure 8. Example illustration of memory-based judgments according to the mixture model. The densities characterize the graded, fallible memory representations. The different vertical lines correspond to the confidence criteria, whereas t corresponds to the test item. Similar to Figue 1, the six-point confidence scale illustrated here goes from `1: very sure change` to `6: very sure same`.

Luck, 2008). As illustrated in Figure 8, memory-based confidence judgments are based on a comparison between the memory representation and the item presented at test. Specifically, the difference between the two is compared with pairs of confidence criteria. Confidence judgments are determined by the "tightest" pair of criteria that includes the difference. For instance, low-confidence "`change`" judgments (e.g., rating `4` in Figure 8) are expected when the difference is relatively small.

Despite some minor differences, the characterization of memory-based responses given above ends up being pretty much the same as the continuous account's (see Figure 2; see also DeCarlo, 2013). Therefore, it is reasonable to question whether it is even necessary to consider a mixture between guesses and memory-based judgments. In other words, *would a "pure" continuous account be enough, given that the memory component in the mixture is what is doing the leg work?* Although the present data do not provide us with the means to implement a critical test on the need for guessing-based confidence judgments, a tentative assessment can be obtained through parametric modeling. The approach we pursued here consisted of fitting a mixture model to individual ROC data coming from each base-rate condition (details can be

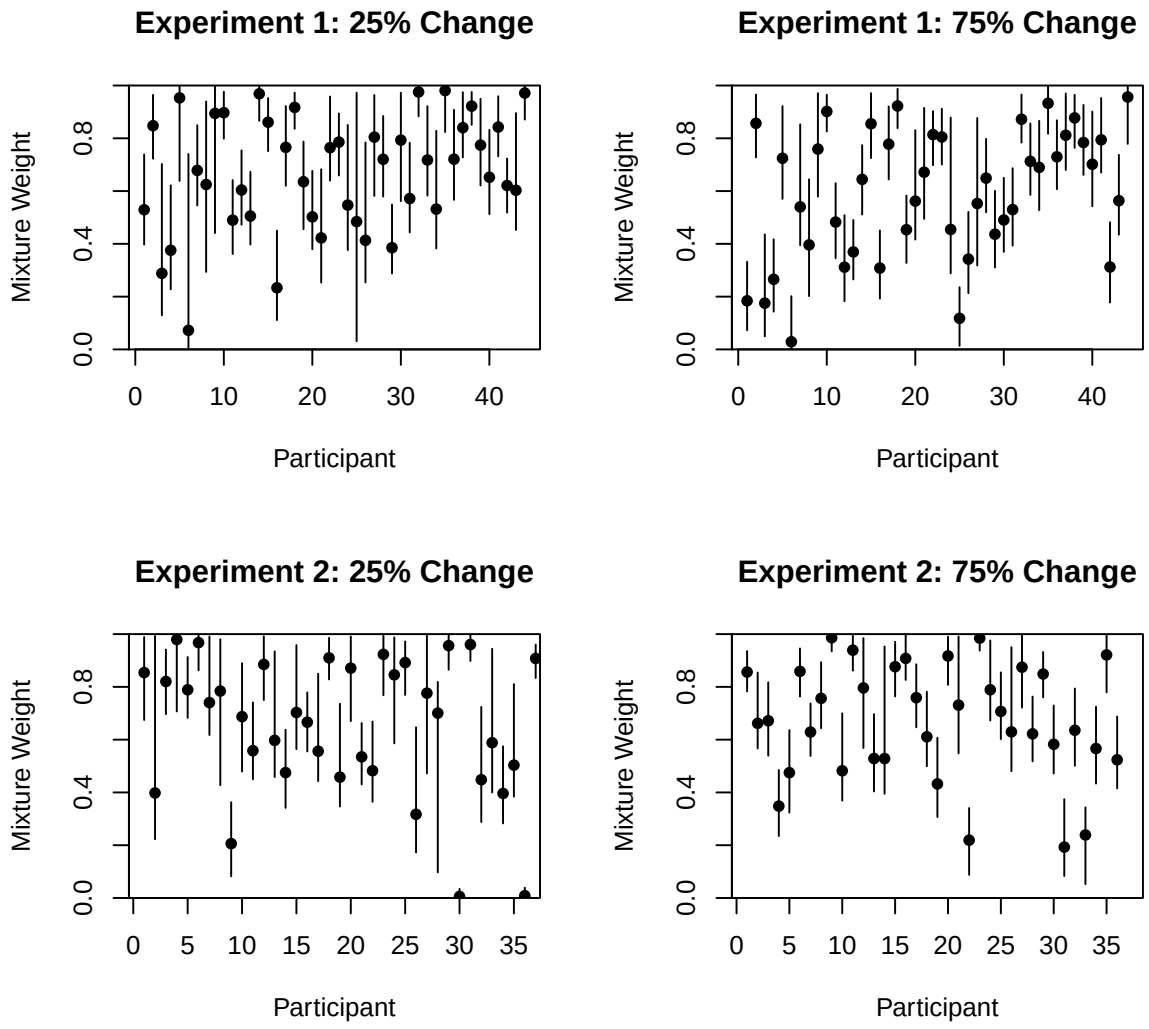

Figure 9. Median posterior estimates of the mixture weights and their respective 95% credibility intervals. A mixture weight of 1 would indicate that responses are solely determined by the memory component and there is no contribution from the guessing component.

found in the Supplemental Materials).[13] Using Bayesian methods, we evaluated the posterior estimates of the *mixture weight* quantifying the proportion of memory-based judgments (and its complement the proportion of guesses). If guessing-based responses make a non-negligible contribution, then the posterior distributions of individuals'

[13]The mixture model used assumes that the memory representation in Same trials follows a Normal distribution with mean 0 and variance 1. In turn, the representation of Change also follows a Normal distribution, this time with mean μ and variance 1. The confidence criteria were free to vary (aside from their order constraints). Guessing-based confidence judgments is captured by an unconstrained multinomial distribution. Finally, the model has a mixture-weight parameter ω.

mixture weights tend be concentrated on values away from 1.[14] The posterior estimates reported in Figure 9 show that even though there is considerable uncertainty, the mixture weights were often concentrated along values away from 1. This result suggests that guessing-based responses play a non-negligible role in describing the data (see also Nosofsky & Gold, 2018).[15]

Conclusion

The present work showed that a number of discrete-state models of VWM exclude the possibility of biased confidence ratings. In two experiments, we were able to observe biased "`change`" judgments, a *behavioral signature* that speaks directly against these models. These results provide us with a clear standard for evaluating the sufficiency and necessity of certain theoretical features (e.g., the nature of memory representations in VWM). Subsequent analyses show that a mixture account assuming pure guesses along with graded memory representations provides a good account of the results (e.g., Nosofsky & Gold, 2018). Altogether, these results contribute to the behavioral foundation of mixture modeling in VWM. These results also have strong implications on the interpretation of previous model comparisons. For instance, it no longer seems reasonable to consider model comparisons supporting a discrete-state account, when such results are predicated on models that fail the present critical test (e.g., Donkin et al., 2014; Rouder et al., 2008).

More broadly, the present work demonstrates the potential of critical tests in the comparison of formal theories (Birnbaum, 2011; Kellen & Klauer, 2014, 2015). Researchers should keep the advantages of this approach in mind: First, it often allows for the direct testing and dismissal of broad classes of models; i.e., strong inference (Platt, 1964). Second, it shifts the focus away from global model-performance statistics,

[14] The mixture model considered here has more parameters (16) than there are degrees of freedom (15), which means that not all of its parameters are identifiable. The lack of identifiability can sometimes be a problem, even in the context of Bayesian parameter estimation (see Spektor & Kellen, 2018). Fortunately, the mixture-weight parameter that we are interested in is not severely compromised given that it plays a major role in establishing the model's ability to describe data, a role that cannot be generally fulfilled by the other model parameters.

[15] For reference, the rank-correlation of the median posterior mixture weights across base-rate conditions in Experiment 1 (remember that this was a within-subjects manipulation) was 0.84.

which can be somewhat opaque, and places it entirely on the specific behavioral patterns that have a clear diagnostic value. Third, it contributes to the development of a corpus of behavioral results that any candidate theory needs to be able to accommodate (see Oberauer et al., 2018).

To finalize, a clarification: The focus on specific portions of the data advocated by the critical-test approach discussed here can be seen as being antagonistic towards the development of accounts that include as many sources of data as possible (e.g., categorical responses, reaction times; Donkin et al., 2013). This is inaccurate. Both approaches are complementary, serving different goals and criteria (for a discussion, see Kellen, 2019): Critical tests provide sharp, localized evaluations of theories, whereas the development of increasingly-encompassing models tests the ability of certain key theoretical notions to provide a competent and consistent characterization of the different types of data, and how these can differ across people and conditions. This complementarity is demonstrated by our parametric mixture-model fits, which show that among the theoretical accounts that happen to provide a plausible characterization of biased confidence judgments, one can find support for the presence of pure guesses.

References

Ashby, F. G. (1983). A biased random walk model for two choice reaction times. *Journal of Mathematical Psychology*, *27*(3), 277–297. doi:10.1016/0022-2496(83)90011-1

Balakrishnan, J. D. (1999). Decision processes in discrimination: Fundamental misrepresentations of signal detection theory. *Journal of Experimental Psychology: Human Perception and Performance*, *25*, 1189–1206.

Bayliss, D. M., Jarrold, C., Gunn, D. M., & Baddeley, A. D. (2003). The complexities of complex span: Explaining individual differences in working memory in children and adults. *Journal of Experimental Psychology: General*, *132*(1), 71–92. doi:10.1037/0096-3445.132.1.71

Bays, P. M., Catalao, R. F. G., & Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision*, *9*(10), 7. doi:10.1167/9.10.7

Bays, P. M. & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science*, *321*, 851–854.

Birnbaum, M. H. (2008). New paradoxes of risky decision making. *Psychological Review*, *115*(2), 463–501. Publisher: American Psychological Association. doi:10.1037/0033-295X.115.2.463

Birnbaum, M. H. (2011). Testing theories of risky decision making via critical tests. *Frontiers in Psychology*, *2*, 315.

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. doi:10.1037/a0015279

Burns, M. S., Leveille, M. D., Dominguez, A. R., Gebhard, B. B., Huestis, S. E., Steele, J., ... Schiller, J. (2017). Measurement of gravitational time dilation: An undergraduate research project. *American Journal of Physics*, *85*(10), 757–762. doi:10.1119/1.5000802

Cowan, N. (2001). The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–114. doi:10.1017/S0140525X01003922

Daneman, M. & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review*, *3*(4), 422–433. doi:10.3758/BF03214546

DeCarlo, L. T. (2013). Signal detection models for the same–different task. *Journal of Mathematical Psychology*, *57*, 43–51.

Diederich, A. & Busemeyer, J. R. (2006). Modeling the effects of payoff on response bias in a perceptual discrimination task: Bound-change, drift-rate-change, or two-stage-processing hypothesis. *Perception & Psychophysics*, *68*(2), 194–207. doi:10.3758/BF03193669

Donkin, C., Nosofsky, R. M., Gold, J. M., & Shiffrin, R. M. (2013). Discrete-slots models of visual working-memory response times. *Psychological Review*, *120*, 873–902.

Donkin, C., Tran, S. C., & Le Pelley, M. (2015). Location-based errors in change detection: A challenge for the slots model of visual working memory. *Memory & Cognition*, *43*(3), 421–431. doi:10.3758/s13421-014-0487-x

Donkin, C., Tran, S. C., & Nosofsky, R. (2014). Landscaping analyses of the ROC predictions of discrete-slots and signal-detection models of visual working memory. *Attention, Perception, & Psychophysics*, *76*(7), 2103–2116. doi:10.3758/s13414-013-0561-7

Dube, C. & Rotello, C. M. (2012). Binary ROCs in perception and recognition memory are curved. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *38*(1), 130–151. doi:10.1037/a0024957

Duhem, P. M. M. (1954). *The aim and structure of physical theory*. Princeton, NJ: Princeton University Press.

Dunsford, I., Bowley, C. C., Hutchison, A. M., Thompson, J. S., Sanger, R., & Race, R. R. (1953). Human Blood-Group Chimera. *Br Med J*, *2*(4827), 81–81. doi:10.1136/bmj.2.4827.81

Erdfelder, E. & Buchner, A. (1998). Comment: Process-dissociation measurement models: Threshold theory or detection theory? *Journal of Experimental Psychology: General*, *127*(1), 83–96. Publisher: American Psychological Association. doi:10.1037/0096-3445.127.1.83

Erdfelder, E., Küpper-Tetzel, C. E., & Mattern, S. D. (2011). Threshold models of recognition and the recognition heuristic. *Judgment and Decision Making*, *6*(1), 7–22.

Falmagne, R. J. (1985). Normative theory and the human mind. *Behavioral and Brain Sciences*, *8*(4), 750–751. Publisher: Cambridge University Press. doi:10.1017/S0140525X00046070

Gelman, A. & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, *66*(1), 8–38.

Green, D. E. (1966). *Signal detection theory and psychophysics*. New York, NY: Wiley.

Kellen, D. (2019). A Model Hierarchy for Psychological Science. *Computational Brain & Behavior*, *2*(3), 160–165. doi:10.1007/s42113-019-00037-y

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*, 1795–1804.

Kellen, D. & Klauer, K. C. (2015). Signal detection and threshold modeling of confidence-rating ROCs: A critical test with minimal assumptions. *Psychological Review*, *122*, 542–557.

Kellen, D. & Klauer, K. C. (2018). Elementary Signal Detection and Threshold Theory. In *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience* (pp. 1–39). American Cancer Society. doi:10.1002/9781119170174.epcn505

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. doi: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. Publisher: Hogrefe Publishing. doi:10.1027/1618-3169/a000272

Kellen, D., Steiner, M., Davis-Stober, C., & Pappas, N. (2019). *Modeling Choice Paradoxes Under Risk: From Prospect Theories to Sampling-Based Accounts*. PsyArXiv. doi:10.31234/osf.io/qvcbk

Kellen, D., Winiger, S., Dunn, J. C., & Singmann, H. (2019). *Testing the Foundations of Signal Detection Theory in Recognition Memory*. PsyArXiv. doi:10.31234/osf.io/p5rj9

Keshvari, S., Berg, R. v. d., & Ma, W. J. (2013). No Evidence for an Item Limit in Change Detection. *PLOS Computational Biology*, *9*(2), e1002927. doi:10.1371/journal.pcbi.1002927

Klauer, K. C. (2010). Hierarchical multinomial processing tree models: a latent-trait approach. *Psychometrika*, *75*, 70–98.

Klauer, K. C. & Kellen, D. (2010). Toward a complete decision model of item and source recognition: a discrete-state approach. *Psychonomic Bulletin & Review*, *17*, 465–478.

Klugkist, I., Kato, B., & Hoijtink, H. (2005). Bayesian model selection using encompassing priors. *Statistica Neerlandica*, *59*, 57–69.

Krantz, D. H. (1969). Threshold theories of signal detection. *Psychological Review*, *76*(3), 308–324. Publisher: American Psychological Association. doi:10.1037/h0027238

Luce, R. D. (1963). Detection and recognition. *Handbook of mathematical psychology*, 103–189.

Luck, S. J. & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*, 279–281.

Malmberg, K. J. (2002). On the form of ROCs constructed from confidence ratings. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(2),

380–387. Place: US Publisher: American Psychological Association. doi:10.1037/0278-7393.28.2.380

Nosofsky, R. M. & Donkin, C. (2016, October). Qualitative contrast between knowledge-limited mixed-state and variable-resources models of visual change detection. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(10), 1507–1525. Publisher: American Psychological Association. doi:10.1037/xlm0000268

Nosofsky, R. M. & Gold, J. M. (2018). Biased guessing in a complete-identification visual-working-memory task: Further evidence for mixed-state models. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(4), 603–625. Publisher: American Psychological Association. doi:10.1037/xhp0000482

Oberauer, K., Lewandowsky, S., Awh, E., Brown, G. D. A., Conway, A., Cowan, N., ... Ward, G. (2018). Benchmarks for models of short-term and working memory. *Psychological Bulletin*, *144*(9), 885–958. doi:10.1037/bul0000153

Platt, J. R. (1964). Strong Inference. *Science*, *146*(3642), 347–353.

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.

Quine, W. V. O. (1963). Two dogmas of empiricism. In *From a logical point of view* (pp. 20–46). New York: Harper & Row.

Rademaker, R. L., Tredway, C. H., & Tong, F. (2012). Introspective judgments predict the precision and likelihood of successful maintenance of visual working memory. *Journal of Vision*, *12*(13), 21–21. Publisher: The Association for Research in Vision and Ophthalmology. doi:10.1167/12.13.21

Ratcliff, R. (1981). A theory of order relations in perceptual matching. *Psychological Review*, *88*(6), 552–572. Publisher: American Psychological Association. doi:10.1037/0033-295X.88.6.552

Ratcliff, R., Clark, S. E., & Shiffrin, R. M. (1990). List-strength effect: I Data and discussion. *Journal of Experimental Psychology: Learning, Memory, and*

Cognition, *16*(2), 163–178. Publisher: American Psychological Association. doi:10.1037/0278-7393.16.2.163

Ricker, T. J., Thiele, J. E., Swagman, A. R., & Rouder, J. N. (2017). Recognition decisions from visual working memory are mediated by continuous latent strengths. *Cognitive Science*, *41*, 1510–1532.

Rouder, J. N. & Morey, R. D. (2009). The nature of psychological thresholds. *Psychological Review*, *116*(3), 655–660. Publisher: American Psychological Association. doi:10.1037/a0016413

Rouder, J. N., Morey, R. D., Cowan, N., Zwilling, C. E., Morey, C. C., & Pratte, M. S. (2008). An assessment of fixed-capacity models of visual working memory. *Proceedings of the National Academy of Sciences*, *105*, 5975–5979.

Rozeboom, W. W. (1970). 2. The Art of Metascience, or, What Should a Psychological Theory Be? In J. Royce (Ed.), *Toward Unification in Psychology* (pp. 53–164). Toronto: University of Toronto Press. doi:10.3138/9781487577506-003

Rozeboom, W. W. (2008). The problematic importance of hypotheses. *Journal of Clinical Psychology*, *64*(9), 1109–1127. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jclp.20522. doi:10.1002/jclp.20522

Shiffrin, R. M. & Nobel, P. A. (1997). The art of model development and testing. *Behavior Research Methods, Instruments, & Computers*, *29*(1), 6–14. doi:10.3758/BF03200560

Shiffrin, R. M. & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*(2), 145–166. doi:10.3758/BF03209391

Spektor, M. S. & Kellen, D. (2018). The relative merit of empirical priors in non-identifiable and sloppy models: Applications to models of learning and decision-making. *Psychon Bull Rev*, 22.

Stephens, R. G., Dunn, J. C., & Hayes, B. K. (2018). Are there two processes in reasoning? The dimensionality of inductive and deductive inferences. *Psychological*

Review, *125*(2), 218–244. Publisher: American Psychological Association. doi:10.1037/rev0000088

Summerfield, C. & de Lange, F. P. (2014). Expectation in perceptual decision making: neural and computational mechanisms. *Nature Reviews Neuroscience*, *15*(11), 745–756. doi:10.1038/nrn3838

Süß, H.-M., Oberauer, K., Wittmann, W. W., Wilhelm, O., & Schulze, R. (2002). Working-memory capacity explains reasoning ability—and a little bit more. *Intelligence*, *30*(3), 261–288. doi:10.1016/S0160-2896(01)00100-3

van den Berg, R., Awh, E., & Ma, W. J. (2014). Factorial comparison of working memory models. *Psychological Review*, *121*(1). doi:10.1037/a0035234

van den Berg, R., Shin, H., Chou, W.-C., George, R., & Ma, W. J. (2012). Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences*, *109*(22), 8780–8785. doi:10.1073/pnas.1117465109

van den Berg, R., Yoo, A. H., & Ma, W. J. (2017). Fechner's law in metacognition: A quantitative model of visual working memory confidence. *Psychological Review*, *124*, 197–214.

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

Wheeler, M. E. & Treisman, A. M. (2002). Binding in short-term visual memory. *Journal of Experimental Psychology: General*, *131*(1), 48–64. Publisher: American Psychological Association. doi:10.1037/0096-3445.131.1.48

Wilken, P. & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision*, *4*, 1120–1135.

Xie, W. & Zhang, W. (2017). Dissociations of the number and precision of visual short-term memory representations in change detection. *Memory & Cognition*, *45*(8), 1423–1437. doi:10.3758/s13421-017-0739-7

Zhang, W. & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*(7192), 233–235. doi:10.1038/nature06860