Cognition 227 (2022) 105208

Contents lists available at ScienceDirect

Cognition

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

Metacognitive judgements of change detection predict change blindness

Adam J. Barnas [a,b], Emily J. Ward [a,*]

[a] Department of Psychology, University of Wisconsin-Madison, Madison, WI, USA
[b] Department of Psychology, University of Florida, Gainesville, FL, USA

ARTICLE INFO

Keywords:
Change blindness
Change detection
Visual awareness
Metacognition

ABSTRACT

People tend to think they are not susceptible to change blindness and overestimate their ability to detect salient changes in scenes. Yet, despite their overconfidence, are individuals aware of and able to assess the relative difficulty of such changes? Here, we investigated whether participants' judgements of their ability to detect changes predicted their own change blindness. In Experiment 1, participants completed a standard change blindness task in which they viewed alternating versions of scenes until they detected what changed between the versions. Then, 6 to 7 months later, the same participants viewed the two versions and rated how likely they would be to spot the change. We found that changes rated as more likely to be spotted were detected faster than changes rated as more unlikely to be spotted. These metacognitive judgements continued to predict change blindness when accounting for low-level image properties (i.e., change size and eccentricity). In Experiment 2, we used likelihood ratings from a new group of participants to predict change blindness durations from Experiment 1. We found that there was no advantage to using participants' own metacognitive judgements compared to those from the new group to predict change blindness duration, suggesting that differences among images (rather among individuals) contribute the most to change blindness. Finally, in Experiment 3, we investigated whether metacognitive judgements are based on the semantic similarity between the versions of the scene. One group of participants described the two versions of the scenes, and an independent group rated the similarity between the descriptions. We found that changes rated as more similar were judged as being more difficult to detect than changes rated as less similar; however, semantic similarity (based on linguistic descriptions) did not predict change blindness. These findings reveal that (1) people can rate the relative difficulty of different changes and predict change blindness for different images and (2) metacognitive judgements of change detection likelihood are not fully explained by low-level and semantic image properties.

1. Introduction

Our experiences and interactions with the natural world tend to be stable and uninterrupted. We rarely, if ever, expect our friend's shirt to change color or the wheel of a car to suddenly disappear in front of our eyes, and we assume that we would immediately and easily notice such occurrences. But in fact, when a visual scene is interrupted, such as when a large object passes in front of our view or even during a brief visual transient like a saccade or a blink, people can fail to see salient changes occurring within the scene, a phenomenon known as change blindness. Change blindness is usually demonstrated using a flicker paradigm (Rensink, O'Regan, & Clark, 1997, 2000), in which participants view a scene that cycles between an unmodified version and a modified version where some part of the scene has changed (separated by a brief disruption like a blank screen or a visual transient resembling a "mud splash"; O'Regan, Rensink, & Clark, 1999). The cycle continues as participants actively search for the change but remain unable to find it. Thus, this flicker paradigm provides a continuous measure of the difficulty of detecting the changing object (Levin, 2002).

Two features of change blindness make it an especially flexible and compelling phenomenon: (1) the changes can be large and occur even to centrally located aspects of the scene and (2) the changes can take several seconds or even minutes for people to detect. These two features make change blindness easy to demonstrate across different experimental paradigms (Levin & Simons, 1997; Levin & Simons, 2000; O'Regan et al., 1999; Rensink et al., 1997; Simons & Rensink, 2005) and in real world settings (e.g., Simons & Levin, 1998).

Despite the magnitude (in size and duration) of change blindness, people are generally unaware that they experience change blindness at all, resulting in "change blindness blindness" (Beck, Angelone, & Levin,

* Corresponding author at: Department of Psychology, University of Wisconsin-Madison, 1202 West Johnson Street, Madison, WI 53706, USA.
E-mail addresses: abarnas@ufl.edu (A.J. Barnas), ejward@wisc.edu (E.J. Ward).

https://doi.org/10.1016/j.cognition.2022.105208
Received 14 August 2021; Received in revised form 11 June 2022; Accepted 14 June 2022
Available online 2 July 2022

2004; Beck, Levin, & Angelone, 2007; Levin, Drivdahl, Momen, & Beck, 2002; Levin, Momen, Drivdahl, & Simons, 2000; Scholl, Simons, & Levin, 2004). Change blindness blindness is robust and has been shown to occur for a range of durations in between changes (e.g., one minute or one hour), for varying number of objects in the scene (i.e., scene complexity), and even when participants are reminded about the role of memory and attention in change detection (Beck et al., 2007; Levin et al., 2002).

In contrast to the flicker paradigm, which is a continuous measure of change blindness, studies of change blindness blindness have usually used a one-shot change detection task. In a one-shot change detection paradigm, developed by Phillips (1974), participants view one repetition of a pre-change and post-change image with a brief disruption inserted in between the images. Participants have a single opportunity to identify the changing object. The one-shot change detection technique has been widely used in visual working memory research where participants are required to temporarily hold items (typically simple stimuli consisting of colored squares or objects) in memory over the duration of the disruption (Gaspar, Neider, Simons, McCarley, & Kramer, 2013; Luck & Vogel, 1997; Vogel, Woodman, & Luck, 2001). Performance in the one-shot change detection technique is measured by the accuracy of detecting the change.

1.1. Does overconfidence mask awareness of relative difficulty?

In one such example of change blindness blindness (Levin et al., 2000), participants were told about four scenarios in three videos that were designed to evoke change blindness (e.g., Levin & Simons, 1997; Simons & Levin, 1998) when an unexpected change occurred across shots (i.e., a plate changing color, a scarf disappearing, or an actor switching with another actor). For each of the four scenarios, participants were asked to imagine watching the videos and were then presented with video stills showing pre-change and post-change views. The experimenter pointed out the change and asked participants whether they thought they would have noticed it in the video. A majority of participants (83%) indicated that they would notice these changes. But in fact, only a handful of participants (11%) in the earlier studies actually detected the same changes. Participants' overconfidence in their own noticing ability extends to their assessment of others' noticing ability as well: when participants were asked whether *other people* would notice the changes, there was no significant difference between the high ratings of their own ability and those for others (Levin et al., 2000; See also Ortega, Montañes, Barnhart, & Kuhn, 2018). These results suggest that participants' change detection metacognition—how participants *think* they will perform in a situation where change blindness may occur—greatly overestimates actual change detection performance, even in situations where participants are made explicitly aware of the object that was changing.

Another possibility is that people are aware that some changes are easier to detect than others. In this case, participants' change detection metacognition may still overestimate actual performance, but participants may be able to predict which changes would lead to the longest change blindness duration. There are several factors that may have resulted in the discrepancy shown by Levin et al. (2000) between what participants *think* they will notice and what participants actually notice. First, this earlier work used a limited set of (four) scenarios and a dichotomous response ("yes" or "no" regarding noticing). Second, there was a frame of reference change across shots in addition to a target object change. Finally, participants were explicitly told what the change was in each scenario before they indicated whether they thought they would notice it or not. Thus, metacognition in this study may reflect participants' ability to detect the *specific* object that is changing.

Change blindness can occur in many different scenarios and, specifically, for many *different* objects with different image properties. Some low-level image properties, such as the size, eccentricity, and visual salience of the change, do not tend to be predictive of change blindness duration (Sareen, Ehinger, & Wolfe, 2015; Stirk & Underwood, 2007; cf. Pringle, Irwin, Kramer, & Atchley, 2001), whereas high-level visual properties, such as scene-schema consistency (whether a change is consistent or inconsistent with a scene) and meaningfulness (the importance of the change to the context of the scene), are predictive of change blindness duration. Specifically, changes that are semantically inconsistent (e.g., a toothbrush appearing on a desk) compared to semantically consistent (e.g., a pencil appearing on a desk) are detected faster and more accurately (Hollingworth & Henderson, 2000; LaPointe, Lupiáñez, & Milliken, 2013; Stirk & Underwood, 2007; cf. Davenport & Potter, 2004 and Türkan, İyilikci, & Amado, 2021). Furthermore, changes that are of central interest (high meaningfulness) compared to marginal interest (low meaningfulness) are also detected faster and more accurately (O'Regan et al., 1999; Pringle et al., 2001; Rensink et al., 1997). If these various factors contribute to change blindness, are they also available as input to metacognitive processes, such that people may be able to judge which changes will be more difficult to detect than others?

1.2. Using metacognition to evaluate awareness

Metacognitive judgements can help clarify instances where behavior does not—or may not—correspond to awareness. Metacognition depends on both low-level visual signals and stimulus awareness: signals from low-level visual processes (e.g., target change detection) provide the signals for high-level cognitive processes (e.g., metacognitive judgements) in a bottom-up hierarchical structure within a signal detection theoretic framework (e.g., Maniscalco & Lau, 2012; Snodgrass, Bernat, & Shevrin, 2004). Because the ability to accurately reflect on performance likely requires explicit information processing, the contents of metacognitive processes are often assumed to reflect the contents of consciousness (Kunimoto, Miller, & Pashler, 2001), although there are some situations in which metacognition and awareness are dissociable (Jachs, Blanco, Grantham-Hill, & Soto, 2015; McAnally, Morris, & Best, 2017; Scott, Dienes, Barrett, Bor, & Seth, 2014).

In the domain of visual awareness, metacognition is typically measured by asking participants to rate their confidence in their performance on a challenging visual awareness paradigm on a trial-by-trial basis, such as rating their confidence that they saw (or did not see) a target presented at perceptual threshold after each trial (e.g., Jachs et al., 2015; Kunimoto et al., 2001; See also Fleming & Lau, 2014). It can also be measured by asking participants to consider a certain scenario and asking them to think about how they would perform, such as displaying the pre-change and post-change versions of a scene from a change blindness scenario and asking participants whether they thought they would notice the change (Levin et al., 2000). Participants with good knowledge about the accuracy of their perceptual abilities will be more confident about correct trials than incorrect trials, demonstrating high metacognitive performance (Fleming, Weil, Nagy, Dolan, & Rees, 2010; Metcalfe & Shimamura, 1994), whereas participants who lack good knowledge about the accuracy of their abilities will not show any relationship between their confidence and accuracy, demonstrating low metacognitive performance.

Assessing what people are and are not aware of in visual scenes can be a challenge because performance and awareness can be dissociated in two ways. First, although explicit knowledge or conscious processing can lead to accurate performance, accurate performance itself cannot be taken as evidence of awareness, such as cases of unconscious priming, statistical learning, and ensemble perception, or even just guessing. Second, although a lack of awareness for a stimulus can lead to poor or chance-level performance, poor performance itself cannot be taken as evidence for a lack of awareness. This second dissociation is clear when comparing the "full-report" assessment of iconic memory (in which participants perform poorly when asked to report *all* letters from a briefly presented letter array) and the "partial-report" assessment (in which participants accurately report a subset of letters from the array;

Sperling, 1960). At face value, the full-report performance suggests that participants had limited awareness of the letters, but the partial-report indicates they were aware of all the letters but could only access some of them before the representation faded.

The dissociation between performance and awareness is potentially a concern in change blindness because poor performance could be due to a failure of comparison rather than a failure of awareness (Hollingworth, 2003; Mitroff, Simons, & Levin, 2004; Scott-Brown, Baker, & Orbach, 2000; Varakin, Levin, & Collins, 2007), in which case participants are unable to compare the unmodified and modified versions of the scenes due to working memory limits but perceive both versions in rich detail. In this situation, participants' metacognition—how confident they are that they successfully detected a change—can be used as an index of the contents of consciousness. For example, when making confidence judgements about their performance in a change detection paradigm with briefly presented stimulus arrays to measure sensory memory (such as iconic memory and fragile visual short-term memory), participants showed equally high metacognition for sensory memory as for more deliberate and explicit working memory (Vandenbroucke et al., 2014). This suggests that people are as aware of their failures of awareness as they are their failures of working memory.

Metacognitive judgements about failures of visual awareness have been used successfully to make inferences about the content of consciousness when measured through confidence ratings after each trial on tasks using challenging but homogenous stimuli, such as Gabor filters presented at threshold (Jachs et al., 2015). In comparison, change blindness experiments typically include a heterogenous set of images, presented at full contrast for several seconds, while participants are fully aware that something obvious is changing right before their eyes. Their failure of awareness persists until they correctly identify the change. Can metacognitive judgements be used to predict failures of awareness of such different magnitude, variability, and content?

1.3. The current study

Here, we systematically test whether participants can predict their own change blindness using a large set of images that vary with respect to scene and change type, but not frame of reference (such as between different scene cuts in a video, like the scenarios used in Levin et al., 2000), which may inadvertently cause multiple objects to change. We measured participants' change blindness duration using a standard change blindness paradigm (O'Regan et al., 1999), in which each image cycles between an unmodified version and a modified version and participants know *something* will change, but do not know what. We used the "mud splash" change detection paradigm (O'Regan et al., 1999) instead of the one-shot paradigm to measure change detection performance because the mud splash paradigm provides continuous measures of change blindness and the difficulty of detecting the change. This allows us to determine whether particular predictor and performance variables are significantly associated with one another. We then brought the participants back several months later, and similar to the approach in Levin et al. (2000), we showed them the unmodified and modified versions with the change highlighted and asked them to judge how likely they thought they were to detect the change in each of the images using a 5-point Likert scale.[1] This subjective rating of performance served as a measure of metacognition for the deliberate change detection task.

In Experiment 1, we show that participants' metacognitive judgements of change detection significantly predicted their change blindness duration (as well as change blindness in others in Experiment 2), such that changes rated as likely to be spotted were detected faster than changes rated as unlikely to be spotted. Moreover, low-level image features (like the size and eccentricity of the change) did not modulate the relationship between metacognitive judgements of change detection and change blindness duration. Finally, in Experiment 3, we show that the semantic similarity (based on linguistic descriptions) between the unmodified and modified scenes significantly predicted metacognitive judgements of change detection but also did not modulate the relationship between metacognitive judgements and change blindness duration.

2. Experiment 1: predicting change blindness with metacognitive judgements

We investigated whether participants' metacognitive judgements of change detection predicted their own change blindness duration. Participants completed two tasks. First, they completed a standard change blindness task that used the "mud splash" change detection technique (O'Regan et al., 1999). The mud splash technique relies upon several local disruptions of the scene's visual continuity (instead of a global disruption in the scene's visual continuity produced by, for example, the flicker change detection technique). We measured time to detect the change (i.e., change blindness duration) for a set of images. Next, the same participants were re-contacted after 6 to 7 months and recruited to the second task, in which they viewed the same set of images from the change blindness task and rated how likely they would be to spot the change in each of the images. Note that the 6 to 7 month interval was chosen to try to reduce participants' reliance on their own past objective change detection performance as the basis for their subjective ratings (Lau & Passingham, 2006), while still retaining a sufficient number of participants. That is, by 6 or 7 months, participants may have remembered the scenes (Standing, 1973; Mandler & Ritchey, 1977), but they likely forgot how quickly they detected the change on a particular trial (which can happen almost immediately, such as in choice blindness tasks; e.g., Johansson, Hall, Sikström, & Olsson, 2005). We assessed how well participants' change detection likelihood ratings predicted their change blindness duration.

2.1. Method

2.1.1. Participants

For both tasks, participants were recruited online via Amazon Mechanical Turk (MTurk). (See Buhrmester, Talaifar, & Gosling, 2018 and Crump, McDonnell, & Gureckis, 2013 for reviews of this subject pool's reliability.) The study was approved by the University of Wisconsin-Madison Institutional Review Board and all participants provided informed consent prior to the start of the task. Here and in all subsequent experiments, eligible participants were age 18 to 70 years, were located within the United States, had at least 100 MTurk tasks approved, and had a task acceptance rate of at least 95%. Participants were paid $7.25/h as compensation for their participation.

For the change blindness task, 450 participants were recruited. Participants who failed to demonstrate adequate task performance were excluded from subsequent analyses: 13 participants failed an attention-check trial and an additional 35 participants failed to click on the change in every experimental trial. Thus, we measured change blindness in a total of 402 participants.

For the change detection likelihood ratings task, we re-contacted these 402 participants 6 to 7 months after their initial participation in the change blindness task and invited them to participate in a follow-up task (i.e., rating the images). A total of 243 people participated (return rate of 60.0%), although 27 of these participants failed two attention-

[1] We avoided asking participants to predict how long it would take them to detect the change because, in general, people are bad at estimating time. For example, prior work has demonstrated that most estimates of task duration are inaccurate and easily biased (Block & Zakay, 1997; Halkjelsvik & Jørgensen, 2012; Roy, Christenfeld, & McKenzie, 2005), and that people tend to overestimate the duration of shorter tasks and underestimate the duration of longer tasks (Roy & Christenfeld, 2008). Moreover, the duration of change blindness varies across both participants and images, so it is not clear what the most appropriate scale anchors would be.

check trials. Thus, we obtained both change blindness durations and change detection likelihood ratings from a total of 216 participants (effective return rate of 53.7%; $M_{age} = 37.37$ years, $SD_{age} = 10.80$ years).

2.1.2. Apparatus

For both tasks in the experiment, stimulus presentation and data collection were completed in participants' web browsers, so viewing distance, operating system, web browser, and screen resolution varied across participants. For the change blindness task, custom scripts written using a combination of PHP and jsPsych (de Leeuw, 2015) controlled all aspects of the task. For the change detection likelihood ratings task, the experiment was created in PsychoPy (Peirce et al., 2019) and converted to the necessary HTML and JavaScript files to be run online.

2.1.3. Stimuli

Both tasks in this experiment drew from 482 total image pairs that were compiled from four sets of stimuli used in previous experiments of change blindness (Ehinger, Allen, & Wolfe, 2016; Ma, Xu, Wong, Jiang, & Hu, 2013; Rensink et al., 1997; Sareen et al., 2015; we direct readers to these papers to learn how each stimulus set was developed). Each pair contained an unmodified and modified version of the image, and the modification—henceforth, the change—could be a change to an object property, such as its color ($n = 93$), size ($n = 15$), or location ($n = 35$), the appearance/disappearance of an object ($n = 335$), or the replacement of an object with a different object ($n = 4$). To account for variability in viewing distance and screen resolution, the dimensions of all stimuli used in every task are reported in pixel values. The average size (as a percent of the total image area) and eccentricity (relative to the center of the image) of the changing object was 1.5% ($SD = 3.2\%$) and 192 pixels ($SD = 85$ pixels), respectively.

For the change blindness task, all images were scaled to fit within an imaginary box that measured 700 pixels × 500 pixels and presented on a white background. A few image pairs had a different aspect ratio but were resized to fit maximally within the boundaries of this box. A mud splash change detection technique adapted from O'Regan et al. (1999) was applied to the images. This technique relies upon several local disruptions of the scene's visual continuity, which create a large number of transients that compete with the transient produced by the change. This competition prevents attention mechanisms from being automatically drawn to the change because the mud splashes act as "decoys" and attract attention to locations other than the location of the change. GNU Octave (Bateman, Eaton, Wehbring, & Hauberg, 2015) using Psychophysics Toolbox extensions (Brainard, 1997; Pelli, 1997) was used to superimpose 6 mud splashes, each ranging between 38 and 128 pixels in either dimension and composed of a high-contrast checkerboard pattern, on both the unmodified and modified images in a pair. Mud splashes were randomly positioned on an image so as not to cover any part of the change nor other mud splashes. The sizes and locations of the mud splashes were consistent within a pair of images but varied between pairs of images (Fig. 1A).

For the change detection likelihood ratings task, all the same image pairs from the change blindness task were used. Now, the unmodified and modified images in each pair (without mud splashes) were presented side-by-side to allow participants to easily compare them with no need to search and no demand on their working memory. Images were scaled to 500 pixels × 357 pixels so that they could be presented next to one another. Additionally, a yellow bounding box (RGB: [255,255 0]; 5 pixels thick) surrounded the site of the object that changed between the two versions of the scene (Fig. 1B) and served to highlight the change that the participants were asked to judge, thus eliminating the possibility that participants would fail to see the change when making their metacognitive judgements, and also ensuring that the metacognitive judgements could not be attributed to a failure of comparison (see Hollingworth, 2003; Mitroff et al., 2004; Scott-Brown et al., 2000; Varakin et al., 2007).

2.1.4. Design & procedure

For the change blindness task, participants viewed 30 image pairs that were randomly selected without replacement from one of the four stimulus sets. Each image pair comprised one experimental trial, and trial order was randomized for each participant. Participants were provided with written instructions that stated that something would change somewhere in the image. Participants were instructed to press the space bar as soon as they spotted the change (providing a measure of response time) and to then click on the location of the change (providing a measure of accuracy). The experiment started with two practice trials at the beginning that provided participants a preview of the change blindness paradigm and one catch trial at the end, which consisted of the disappearance of a large central object (occupying approximately 35% of the image) and served as an attention check to ensure participants knew what the task was and could perform the task. A change was present in every image pair and each image pair was viewed by approximately 18 participants ($SD = 5$ participants, range = 2 to 28 participants).

Each trial began with a centrally presented fixation cross for 3 s, which signaled to participants the start of the trial (Fig. 1A). The fixation cross disappeared, leaving a white screen that was presented for 1 s. The change blindness paradigm then began and consisted of four alternating images in the following sequence: A, B′, B, A′. Unmodified and modified images without mud splashes (A and B) were presented for 2 s and unmodified and modified images with mud splashes (A′ and B′) were presented for 100 ms. The entire sequence repeated until participants pressed the space bar (indicating they had detected the change) or timed out after 60 s. When the space bar was pressed, the last image without mud splashes remained on the screen. At that point, participants were required to use the trackpad or mouse to click on the object that had changed. There was no time constraint to make this response. After participants clicked on the object, the experiment immediately proceeded to the next trial. Thus, response time and accuracy were recorded for every image pair. Response time was measured from the onset of the first image in the change blindness sequence (image A) to when a participant pressed the space bar.[2] To account for skewness in response time data[3] and to achieve as close as a normal distribution, we computed the log response time and report this variable as the change blindness duration. Accuracy was determined based on whether a participant's click landed within a bounding box that surrounded each change. The size and the placement of the bounding box was determined using two custom Python scripts. The first script computed a difference mask between the images in a pair (e.g., for defining the change between two versions of a change blindness stimulus). The second script drew a convex hull around the difference. The bounding box (or, rectangle) was then drawn around the convex hull based on its minimum and maximum vertices. Trials in which the click landed within the bounding box were considered accurate and trials in which the click landed outside the

[2] Of course, participants cannot detect a change within the first image presentation because nothing has yet changed. Subtracting 2 s (corresponding to the duration of the first image) from all responses would account for this, but because this operation would be applied to all data points, it would only affect the intercept of our model, rather than the relationship between change blindness duration and change detection likelihood ratings. Thus, we chose to measure response time from trial onset, consistent with other studies using flicker methods (e.g., Ehinger et al., 2016; Hollingworth & Henderson, 2000; Pringle et al., 2001; Sareen et al., 2015).

[3] The amount of skewness of a distribution is determined based on the skewness factor. A log transform can then be applied to data that exceed a specified skewness factor (e.g., ±3). We used the *moments* package (Komsta & Novomestky, 2015) in R to calculate the skewness factor for the raw RT data, which was 3.44. Using a liberal range of skewness of ±3, we log transformed the raw RT data and the resulting skewness factor was 1.37. Similar logarithmic transformations have been applied to skewed change detection RT data with more conservative criteria (i.e., ±1.5; see Pringle et al., 2001).

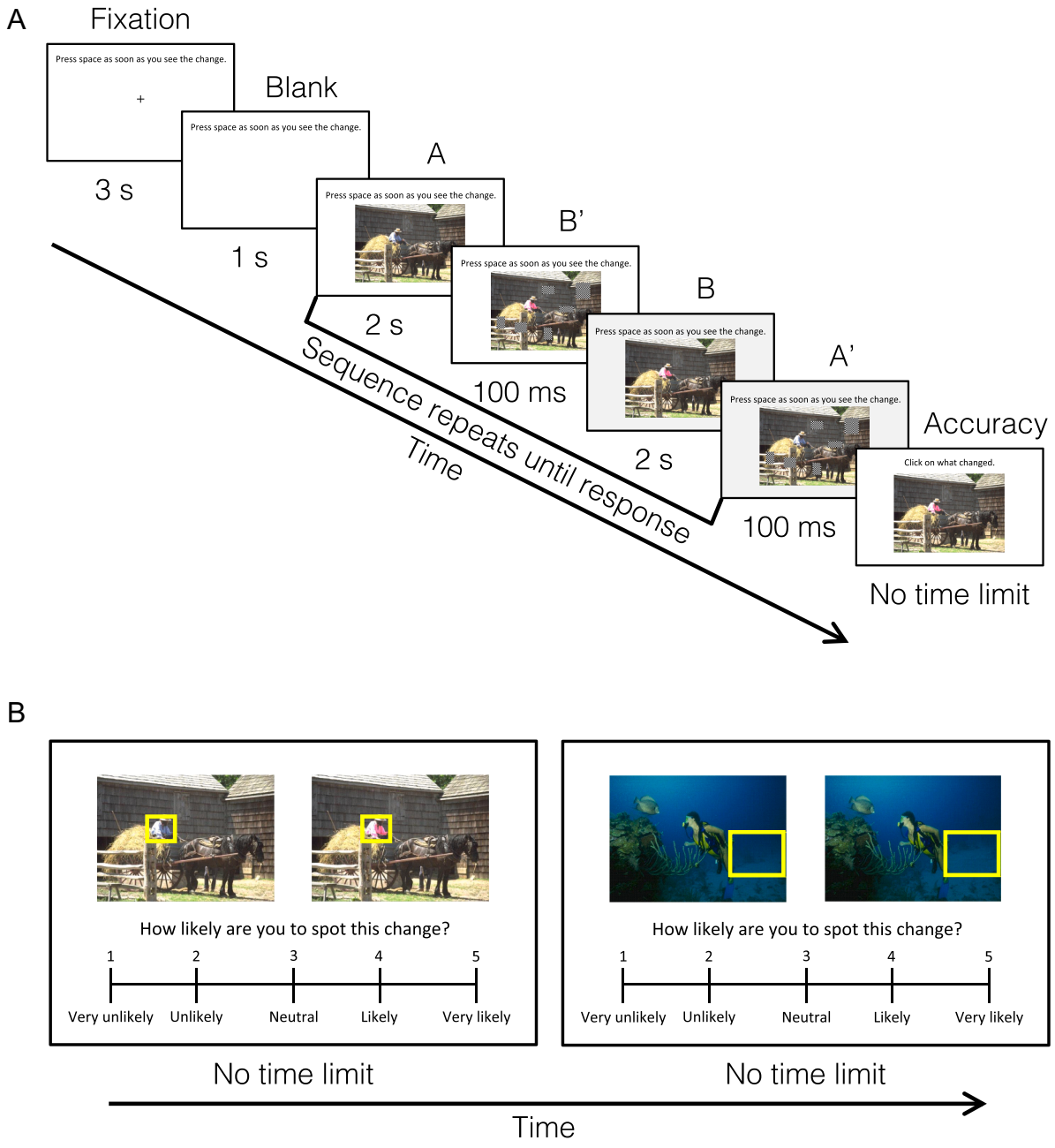

Fig. 1. A. Design and example trial for the change blindness task in Experiments 1 and 2. B. Design and example trials for the change detection likelihood ratings task in Experiments 1 and 2. *Note*: Images and text are not drawn to scale.

bounding box were considered inaccurate. This bounding box was not visible to participants during the task.

For the change detection likelihood ratings task, returning participants viewed the same 30 image pairs they viewed in the change blindness task. Each image pair comprised one experimental trial, and trial order was randomized for each participant. Participants were provided with written instructions that stated that the pair of pictures differed slightly and that the change would be outlined by a yellow rectangle and directed participants to rate how likely (on a 5-point Likert scale) they are to spot the change, with 1 being "very unlikely" to spot the change and 5 being "very likely" to spot the change. In each trial, the 5-point Likert scale was displayed below the image pair (Fig. 1B) and participants provided their rating. There was no time constraint and participants were told to take enough time to think about their ability to detect each change before responding. The experiment immediately advanced to the next trial once the participant made a response (numbers 1–5 on the participant's keyboard). Two catch trials at the end served as attention checks and consisted of the disappearance of a large central object and a small peripheral object (which needed to be rated as "likely" or "very likely" and "unlikely" or "very unlikely", respectively, for the participant's data to be included in further analyses). Each image pair was rated by an average of 10 participants (SD = 4 participants, range = 1 to 20 participants).

2.2. Results

For all 402 participants who completed the change blindness task, there were an average of 9 inaccurate trials (SD = 8 trials) per participant that were excluded (the number of *total* inaccurate trials excluded across these participants was 3443). In addition, of the 402 participants,

10 participants each had one trial excluded with no response or that timed out and 109 participants had an average of 1 trial ($SD = 0.28$) excluded in which the response time was ±3 standard deviations from the participant's mean response time (the number of *total* trials excluded for being ±3 SDs from the participant's mean across these participants was 116). Altogether, a total of 3569 trials were excluded, accounting for approximately 30% of all trials. These exclusions helped ensure that we measured change blindness *per se* rather than inattentiveness, and by considering only change blindness duration from accurate trials, we eliminated response bias that can be problematic in measures of metacognition that use binary (i.e., correct/incorrect) decisions (Galvin, Podd, Drga, & Whitmore, 2003). However, we also conducted all our primary analyses without any of these exclusions as a robustness check. In brief, our primary results were robust even when all trials and participants are included (See the Supplementary Materials), and we note results that were less robust without these exclusions

In the change blindness task, participants took approximately 10.03 s ($SD = 6.05$ s) on average to notice the changing object in the scene. In the change detection likelihood ratings task, the average rating across all trials was 3.06 ($SD = 1.33$). The percentage that each value on the rating scale was used across all trials is as follows: 1 = 16.3%, 2 = 20.4%, 3 = 21.0%, 4 = 26.0%, 5 = 16.4%.

To address our main research question, whether participants' metacognitive judgements of change detection ability predicted change blindness duration, we tested the association between change detection likelihood ratings and change blindness duration for each image on a trial-by-trial basis. Using the *lme4* package (Bates, Mächler, Bolker, & Walker, 2015) in R, we constructed a linear mixed-effects model for predicting change blindness duration with change detection likelihood ratings as a fixed effect and participant, image pair, and stimulus set as random intercepts. Including these random intercepts in our model allowed us to partial out the variance due to participant, image pair, and stimulus set. Here and in all subsequent models, interval predictors were standardized to eliminate potential collinearity and change detection likelihood ratings were analyzed as an ordinal fixed effect.

We found a small but highly significant negative relationship between change detection likelihood ratings and change blindness duration, $\beta = -0.18$, $p < .001$, 95% CI = [−0.25, −0.11], such that changes rated as likely to be spotted were detected faster than changes rated as unlikely to be spotted (Fig. 2A). This suggests that participants' metacognitive judgements of change detection ability significantly predict change blindness duration. Nonetheless, there were several other accounts of our data to rule out.

To ensure that metacognitive judgements of change detection ability predicted change blindness *per se,* rather than visual search times, we calculated change blindness duration as a function of change exposures (i.e., unmodified to modified, modified to unmodified, etc.). This is because participants may only have the opportunity to detect a change when these transitions between versions occur. Similar to our finding using change detection *time*, change detection likelihood ratings significantly predicted the number of change exposures, $\beta = -0.20$, $p < .001$, 95% CI = [−0.27, −0.13], such that changes rated as likely to be spotted were detected in fewer exposures than changes rated as unlikely to be spotted.

We also wanted to ensure that the relationship between change detection likelihood ratings and change blindness duration was not merely an artifact of the fact that variation in change blindness duration is a prerequisite for a correlation between duration and ratings. We calculated the change blindness standard deviation for each participant and included it as a covariate in our regression. This allowed us to measure our primary relationship of interest between change blindness duration and change detection likelihood while controlling for variation in change blindness times. We found that change blindness standard deviation was highly predictive of change blindness duration, $\beta = 0.48$, $p < .001$, 95% CI = [0.45, 0.51], which is reasonable because people with a lot of variability will likely have a longer mean duration than people without much variability. But critically, change detection likelihood ratings continued to significantly predict change blindness duration, $\beta = -0.17$, $p < .001$, 95% CI = [−0.24, −0.11], even when accounting for individual variability in change blindness duration. This means that the relationship is not an artifact of variation in detection time.

Although these participants completed the likelihood rating portion of the experiment 6 to 7 months after the change blindness task, if participants nonetheless remembered how good or bad their performance was for the change blindness task, their *performance memory* (and not their metacognition) could explain our results. This could be the case even if participants only remembered a handful of hard and easy changes. To rule out this possibility, we excluded the 6 hardest and 6 easiest trials for each participant. Change detection likelihood ratings continued to significantly predict change blindness duration even after excluding the 6 hardest and easiest trials for each participant, $\beta = -0.08$,

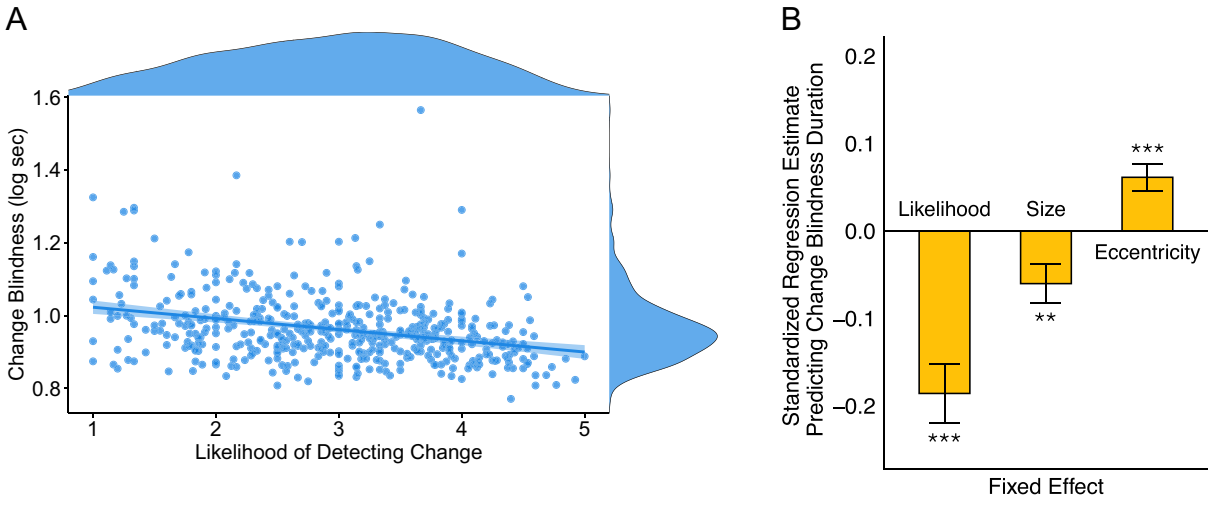

Fig. 2. A. Metacognitive judgements of change detection likelihood significantly predict change blindness duration (Experiment 1). The average likelihood rating and change blindness duration is plotted for each image pair and each dot represents an image pair. Density distributions are provided for metacognitive judgements of change detection likelihood and change blindness duration. B. Standardized regression estimates from the linear mixed-effects model predicting change blindness duration (Experiment 1), indicating that likelihood ratings, size, and eccentricity are significant predictors of change blindness duration. *Note:* $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

$p < .05$, 95% CI = [−0.15, −0.01]. This result indicates that the relationship between change detection likelihood ratings and change blindness duration is present for images that are not particularly hard or easy (and thus, performance on these is unlikely to be remembered by participants).

Next, we examined whether the strength of the relationship between change detection likelihood ratings and change blindness duration varied as a function of experiment half or by difficulty of the change. We assessed these two factors separately by including them in the linear mixed-effects model and testing for an interaction between likelihood ratings and the factor of interest on change blindness duration (i.e., is there a difference in slope based on the factor). We found that the relationship between change detection likelihood ratings and change blindness duration for each image was not significantly different ($\beta = -0.04$, $p = .475$, 95% CI = [−0.14, 0.07]) for the first half of trials ($r(474) = -0.32$, $p < .001$) as compared to the last half of trials ($r(474) = -0.32$, $p < .001$), indicating that participants' metacognitive accuracy did not improve as they made more ratings (and thus learned the distribution of change difficulty). Furthermore, the relationship between change detection likelihood ratings and change blindness duration was not significantly different ($\beta = -0.04$, $p = .504$, 95% CI = [−0.15, 0.07]) for the easiest images (i.e., the half of images with changes that were detected the fastest; $M_{easiest} = 8.29$ s, $SD_{easiest} = 3.08$ s; $r(239) = -0.13$, $p < .05$) as compared to the hardest images (i.e., the half of images with changes that were detected the slowest; $M_{hardest} = 11.83$ s, $SD_{hardest} = 7.62$ s; $r(238) = -0.23$, $p < .001$). These results suggest that participants' metacognition about change detection is not acquired over the course of performing judgement tasks and that participants' metacognition about change detection can be used to predict performance for many types of changes (easy or hard).

To test whether participants' change detection likelihood ratings simply draw from low-level image properties of the change, such as how big the change is or where the change occurs in the image, we included size (as a percent of the total image area) and eccentricity (relative to the center of the image) as additional predictors in the model. Both size ($\beta = -0.06$, $p < .01$, 95% CI = [−0.10, −0.02]) and eccentricity ($\beta = 0.06$, $p < .001$, 95% CI = [0.03, 0.09]) weakly but significantly predicted change blindness duration, such that changes located near the center of the image were detected faster than changes located in the periphery of the image, and larger changes were detected faster than smaller changes. Critically, however, change detection likelihood ratings remained significantly predictive of change blindness duration, $\beta = -0.19$, $p < .001$, 95% CI = [−0.25, −0.12] (Fig. 2B). This indicates that the ability of participants' metacognitive judgements of change detection to predict change blindness is not modulated simply by low-level image properties. In fact, size ($\beta = 0.13$, $p = .449$, 95% CI = [−0.21, 0.48]) and eccentricity ($\beta = -0.005$, $p = .958$, 95% CI = [−0.20, 0.19]) did not predict change detection likelihood ratings when included as fixed factors in a model (with participant, image pair, and stimulus set included as random intercepts) predicting change detection likelihood ratings directly.

Finally, we asked whether participants with the highest metacognitive performance were least susceptible to change blindness. For each participant, we measured individual "metacognitive performance" by finding the correlation between their change detection likelihood ratings and their change blindness durations (where a strong negative correlation indicates most accurate prediction, or higher metacognitive performance). Then, we correlated individual metacognitive performance and their mean change blindness duration. Surprisingly, we found a significant negative correlation, $r(214) = -0.14$, $p < .05$, indicating that participants with higher metacognitive performance experienced change blindness for *longer* than participants with lower metacognitive performance.

We also conducted several additional analyses on these data, which are available in the Supplementary Material.

2.3. Discussion

Rather than being indiscriminately overconfident in their ability to notice changes to visual scenes, participants are aware of the relative difficulty of the changes and their change detection metacognition—how likely participants *think* they are to detect different changes across a set of images—predicts their own change blindness for the images. These metacognitive judgements about change detection ability are not explained by low-level image properties, such as the size or eccentricity of the change, although these properties significantly predicted change blindness duration. This finding differs from previous work that found mixed or no effects for similar low-level image properties of the change on change blindness (Sareen et al., 2015; Stirk & Underwood, 2007), but replicate previous work that showed that the eccentricity of the change is predictive of change blindness duration (Pringle et al., 2001). Critically, these properties did not predict change detection likelihood ratings and the ratings remained predictive of change blindness even when we controlled for these image properties. The predictive power of metacognitive judgements does not seem to stem from experience making such judgements, since they were equally effective for both halves of the experiment. Finally, having high metacognition does not protect individuals from change blindness. In fact, those with the highest metacognition experienced change blindness for the longest. This could reflect an experimental strategy of carefulness or thoroughness of these participants. Alternatively, it may show that high metacognition is not beneficial in change blindness scenarios, even though it is predictive of change blindness.

Participants were re-contacted to complete the change detection likelihood ratings task 6 to 7 months after completing the change blindness task. A long gap in between tasks was chosen to try to reduce participants' reliance on their own past objective change detection performance as the basis for their subjective ratings (Lau & Passingham, 2006), but we did not ask participants whether they remembered having done the change blindness task. This means that we do not know if participants remembered (1) the scenes (and if so, which ones) or (2) their previous change detection performance. Because participants were asked to rate how likely they were to find the change when the scenes and change were clearly visible (thereby not taxing memory at all), the quality of their *scene memory* should not impact this judgment. However, if participants remembered how good or bad their performance was for the change blindness task, their *performance memory* (and not their metacognition) could explain our results. Previous research has demonstrated that memory for trial-wise performance is not especially robust (Johansson et al., 2005). More importantly, we ruled out the possibility that our findings stemmed from participants having remembered their performance on even a handful of hard and easy changes by demonstrating that metacognitive judgements predict change blindness when those trials are excluded.

3. Experiment 2: predicting change blindness with judgements from self vs. other

Experiment 1 showed that participants can rate which changes to visual scenes took them the longest to detect. The ability of these metacognitive judgements to predict change blindness could be due to the individual—participants rely on their memory for their own change detection performance—or could be due to the images—participants rely on the variability among changes in the images. In general, people are just as overconfident in *others*' ability to notice salient changes to scenes as they are about their own abilities (Levin et al., 2000). Do participants' own change detection likelihood ratings best predict their own change blindness duration, or are the ratings from others just as effective? Here, we collected change detection likelihood ratings from a new group of participants and tested whether the ratings from the new group were similarly effective at predicting the change blindness durations of the participants from Experiment 1. We also establish whether

accurate metacognitive judgements require actually experiencing change blindness or can be obtained from a new group of participants who necessarily have no memory for the actual experienced detectability of the scenes.

3.1. Method

This experiment was identical to the change detection likelihood ratings task of Experiment 1 (Fig. 1B), except 343 new participants were recruited in the same manner as in Experiment 1, none having previously completed any change blindness experiment in our lab. Participants who did not respond correctly on either attention-check trial were excluded (123 participants failed these checks), resulting in a final sample of 220 participants (M_{age} = 38.29 years, SD_{age} = 12.18 years). Each image pair was rated by an average of 14 participants (SD = 3 participants, range = 4 to 29 participants).

3.2. Results and discussion

The data from the change blindness task from Experiment 1 were used here (to restate: participants took approximately 10.03 s (SD = 6.05 s) on average to notice the changing object in the scene). In the change detection likelihood ratings task, the average rating across all trials for these new participants was 2.98 (SD = 0.71). The percentage that each value on the rating scale was used across all trials by these new participants is as follows: 1 = 18.8%, 2 = 21.8%, 3 = 19.3%, 4 = 23.1%, 5 = 17.0%.

To determine whether change detection likelihood ratings from a new group of participants were equally predictive of Experiment 1 (returning) participants' change blindness duration as their *own* change detection likelihood ratings, we constructed a linear mixed-effects model for change blindness duration with participant group (new or returning) and change detection likelihood ratings (and critically, the interaction between group and ratings). We also included size and eccentricity of the change as fixed effects to control for low-level image properties of the change, and participant, image pair, and stimulus set as random intercepts. Change detection likelihood ratings ($\beta = -0.05$, $p < .05$, 95% CI = [−0.1, −0.004]), size ($\beta = -0.10$, $p < .01$, 95% CI = [−0.17, −0.03]), and eccentricity ($\beta = 0.08$, $p < .01$, 95% CI = [0.03, 0.13]) of the change significantly predicted change blindness duration, replicating our finding from Experiment 1 (Fig. 3B). However, there was no interaction between participant group and change detection likelihood ratings, $\beta = 0.002$, $p = .941$, 95% CI = [−0.06, 0.07] (Fig. 3A), although the relationship between change blindness duration and likelihood ratings was numerically stronger for these new participants, $r(479) = -0.35$, $p < .001$, than for the returning participants, $r(479) = -0.31$, $p < .001$.

These results indicate that change detection likelihood ratings from an independent group of participants can predict change blindness duration in a separate group of people. Therefore, the ability of metacognitive judgements about change detection to predict change blindness is likely due to variability among changes in the images, rather than introspection about or memory for personal performance. Specifically, scene memory or performance memory cannot explain our results because participants in this experiment did not actually experience change blindness for these scenes (i.e., they necessarily did not participate in the change blindness task) yet provided likelihood ratings that predicted change blindness in others.

4. Experiment 3: predicting change blindness and metacognitive judgements with semantic similarity

Because change detection likelihood ratings from participants can predict change blindness in an independent group of participants, change detection metacognition likely relies on input from the changing scene itself. Change detection likelihood ratings remain a significant predictor of change blindness after low-level image properties like change size and eccentricity are controlled for, but it is possible that participants' judgements are based on a high-level property, such as the semantic similarity between the unmodified and modified scene. For example, a color change to a traffic light in a street scene would result in two versions of the scene that are semantically different and may be an easier change to detect, whereas a color change to a miscellaneous building in a street scene would result in two versions of the scene that are semantically similar and may be a more difficult change to detect. Here, to quantify semantic similarity of the changes, a group of new participants provided a pair of written descriptions for each scene version of the image pair and then a separate group of participants provided a rating of the similarity between the descriptions. Although people certainly know more about the scenes than what they write, this

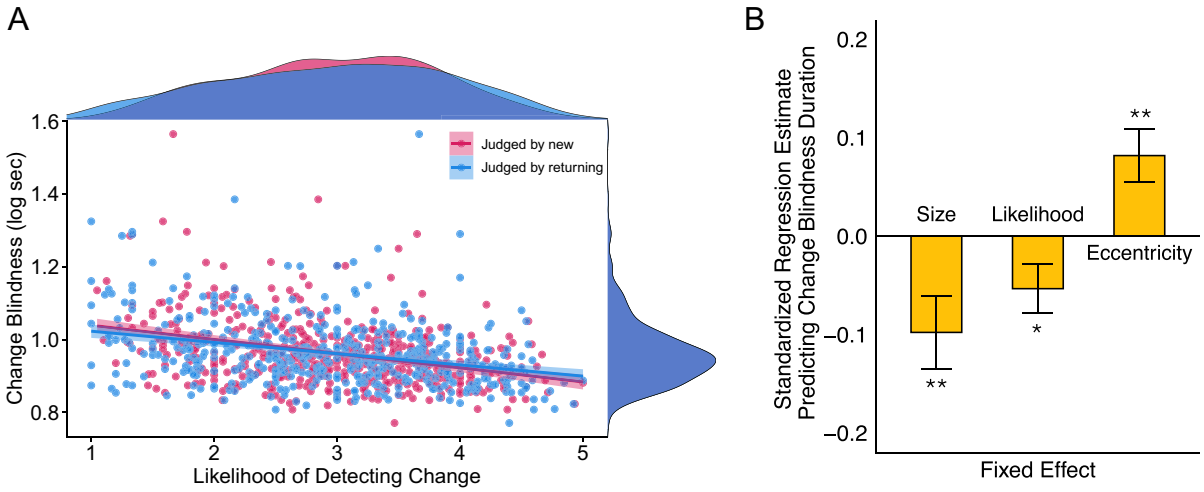

Fig. 3. A. No significant interaction between participant group and metacognitive judgements of change detection likelihood (Experiment 2). The average likelihood rating and change blindness duration is plotted for each image pair and each dot represents an image pair. Density distributions are provided for metacognitive judgements of change detection likelihood and change blindness duration. B. Standardized regression estimates from the linear mixed-effects model predicting change blindness duration (Experiment 2), indicating that size, likelihood ratings, and eccentricity are significant predictors of change blindness duration in others. *Note:* *$p < .05$, **$p < .01$, ***$p < .001$.

approach of rating written descriptions of scenes has the advantage of assessing semantic similarity of the change independent of any visual similarity. We analyzed whether semantic similarity (based on the similarity ratings of the linguistic descriptions) predicted metacognitive judgements of change detection and whether metacognitive judgements continued to predict change blindness when controlling for semantic similarity.

4.1. Method

4.1.1. Participants

For the written descriptions task, 323 new participants completed the task. Data from 97 participants were excluded from the initial sample due to poor descriptions and/or low effort (i.e., one-word descriptions) on either attention-check or main trials. This resulted in a final sample of 226 participants (M_{age} = 37.22 years, SD_{age} = 10.67 years).

For the description similarity ratings task, 472 new participants completed the task. Participants who failed an attention-check trial were excluded, which totaled 131 participants. Thus, we collected description similarity ratings from 341 participants (M_{age} = 38.40 years, SD_{age} = 11.08 years).

4.1.2. Apparatus and stimuli

For both stages of the experiment (providing descriptions and ratings the similarity between descriptions), stimulus presentation and data collection were completed in participants' web browsers, so viewing distance, operating system, web browser, and screen resolution varied across participants. Each component was created in PsychoPy (Peirce et al., 2019) and converted to the necessary HTML and JavaScript files to be run online.

We used the same 482 image pairs that were used in Experiment 1. The written descriptions task was identical to the change detection likelihood ratings task of Experiment 1 (Fig. 1B), except two textboxes (500 pixels × 250 pixels) replaced the rating scale and were located below the images (Fig. 4A). For the description similarity ratings task, two textboxes (also 500 pixels × 250 pixels) replaced the images and were presented side-by-side (Fig. 4B). The textboxes in the description similarity ratings task contained the descriptions collected in the written descriptions task.

4.1.3. Design and procedure

For the written descriptions task (Fig. 4A), participants were instructed to view the pair of images and write a description of each image as a whole, including the object inside the bounding box. We presented these images side-by-side to allow participants to easily compare them with no need to search and no demand on their working memory. We were also concerned that participants may not even notice the changing object if participants' attention was not drawn to it via the bounding box (after all, participants were instructed to provide descriptions of each image, not spend time searching for the changing object). More generally, if the participants providing the descriptions did not find the change themselves, then the task would not match the performance variable from the change blindness task, because

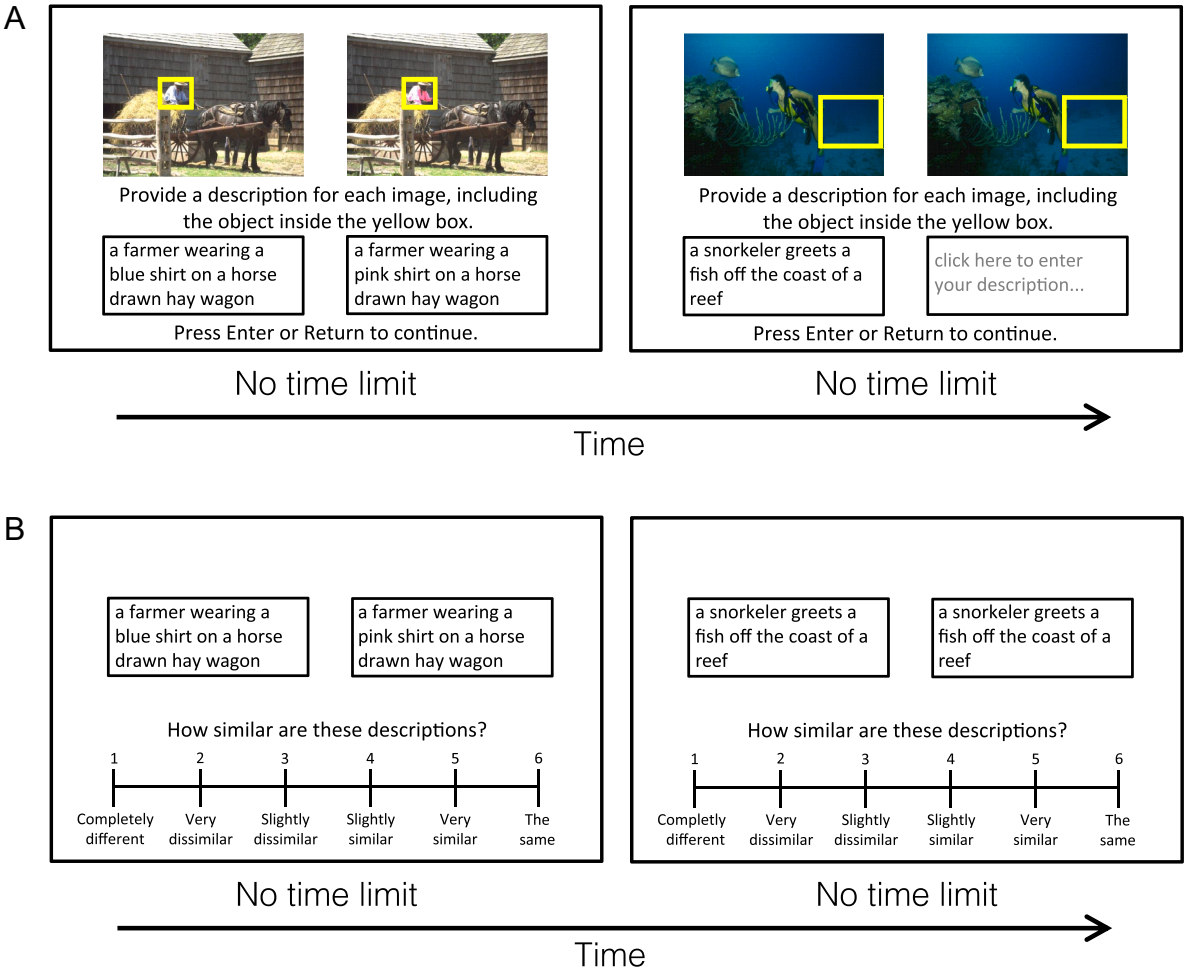

Fig. 4. A. Example trial for the written descriptions task in Experiment 3. B. Example trial for the description similarity ratings task in Experiment 3. *Note*: Images and text are not drawn to scale.

participants (almost always) found the change in that task.

Three examples were provided before the start of the task and participants were instructed to use the examples as models for their own descriptions. Participants provided descriptions for 12 randomly selected image pairs (without replacement) from within one of the four stimulus sets, in addition to two attention-check trials at the end of the task. Attention-check trials consisted of the disappearance of a large central object. There was no time constraint and participants were told to allow sufficient time to provide detailed descriptions. Participants were able to advance to the next trial once text was entered into both textboxes. Each image pair was described by an average of 6 participants (SD = 2 participants, range = 2 to 12 participants) and a total of 2667 descriptions were obtained. The descriptions for each pair of images in a stimulus set served as the stimuli for the description similarity ratings task, as follows.

For the description similarity ratings task (Fig. 4B), the new group of participants were instructed to read the pair of descriptions and rate how similar the written descriptions were to each other on a 6-point scale, with 1 being "completely different" and 6 being "exactly the same". Participants were told that the descriptions were provided by someone else who had viewed a pair of pictures that contained a small difference. Written descriptions appeared within separate textboxes and the similarity rating scale was displayed below. We opted for a 6-point scale so that participants did not have the option of providing a neutral midpoint response. Participants viewed 30 randomly selected pairs of descriptions (without replacement) from within one of the four stimulus sets. A participant could have rated multiple descriptions of the same image pair, and this occurred for 290 out of 341 participants. Across all 341 participants, each participant rated an average of 3 descriptions (SD = 3 descriptions) of the same image pair. Participants also completed one attention-check trial at the end. The descriptions on the attention-check trial were exactly the same and should have been rated as such. There was no time constraint and participants were told to take enough time to read through the descriptions before making a response. The task immediately advanced to the next trial once the participant made a response (numbers 1–6 on the participant's keyboard). Each pair of descriptions was rated by an average of 4 participants (SD = 1 participant, range = 2 to 10 participants). Thus, in total, each image pair was rated by an average of 21 participants (SD = 8 participants, range = 10 to 56 participants).

4.2. Results

For the description similarity ratings task, the average rating across all trials was 3.93 (SD = 0.45). The percentage that each value on the rating scale was used across all trials is as follows: 1 = 5.9%, 2 = 9.9%, 3 = 18.1%, 4 = 25.3%, 5 = 34.2%, 6 = 6.7%.

Similarity ratings were analyzed with the change blindness durations and metacognitive judgements collected in Experiment 1. We first examined whether semantic similarity ratings were predictive of metacognitive judgements of change detection and constructed a linear mixed-effects model with semantic similarity ratings as a single ordinal fixed effect and participant, image pair, and stimulus set as random intercepts. The semantic similarity between the unmodified and modified images was significantly predictive of metacognitive judgements of change detection ($\beta = -0.29$, $p < .01$, 95% CI = [−0.49, −0.10]), such that changes rated as having more similar descriptions were judged as being more difficult to detect than changes rated as having less similar descriptions (Fig. 5A).

We next examined whether change detection likelihood ratings continued to predict change blindness duration when accounting for the semantic similarity between the images in a pair. We constructed a linear mixed-effects model predicting change blindness duration with change detection likelihood ratings, semantic similarity ratings, size and eccentricity of the change as fixed effects and participant, image pair, and stimulus set as random intercepts. Replicating Experiments 1 and 2, change detection likelihood ratings ($\beta = -0.19$, $p < .001$, 95% CI = [−0.25, −0.12]), size ($\beta = -0.06$, $p < .01$, 95% CI = [−0.10, −0.02]), and eccentricity ($\beta = 0.06$, $p < .001$, 95% CI = [0.03, 0.09]) of the change significantly predicted change blindness duration. However, not only did semantic similarity not modulate the relationship between change detection likelihood ratings and change blindness, but semantic similarity itself was not a significant predictor of change blindness duration ($\beta = 0.003$, $p = .849$, 95% CI = [−0.03, 0.03]; Fig. 5B). These results suggest that participants' change detection metacognition is a robust predictor of change blindness for scenes and is not accounted for by the semantic similarity of the unmodified and modified images.

4.3. Discussion

When analyzed as a single predictor, ratings of semantic similarity (based on linguistic descriptions of the unmodified and modified scenes) significantly predicted metacognitive judgements of change detection difficulty. However, when we included semantic similarity and

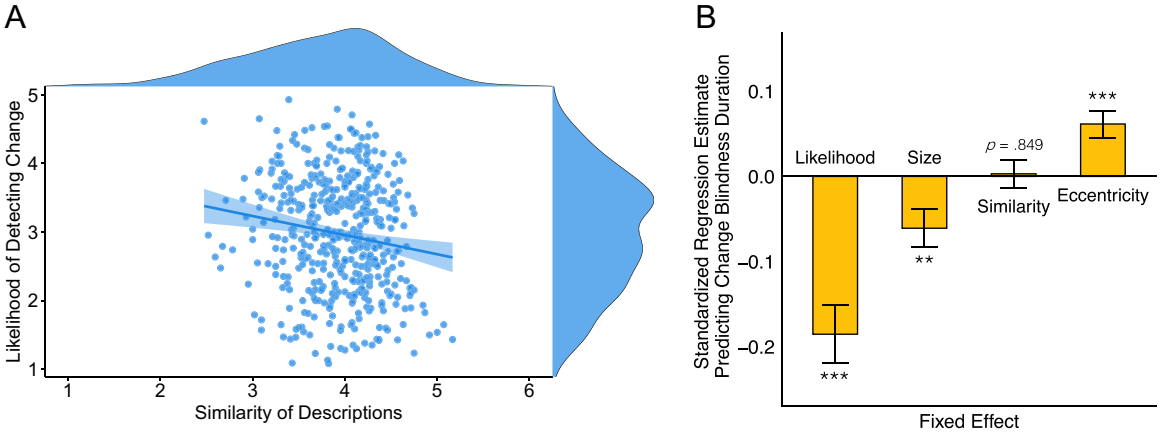

Fig. 5. A. Semantic similarity ratings (based on linguistic descriptions) predict metacognitive judgements of change detection likelihood (Experiment 3). The average similarity rating and likelihood rating is plotted for each image pair and each dot represents an image pair. Density distributions are provided for semantic similarity ratings and metacognitive judgements of change detection likelihood. B. Standardized regression estimates from the linear mixed-effects model predicting change blindness duration (Experiment 3), indicating that likelihood ratings, size, eccentricity, but not semantic similarity predict change blindness duration. *Note:* $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

metacognitive judgements in our model of change blindness duration, semantic similarity did *not* predict change blindness duration, and metacognitive judgements continued to predict change blindness duration. Together, these results suggest that people rely upon the semantic similarity of scenes when judging change difficulty (indexed by metacognitive judgements), but not when detecting change (indexed by change blindness duration). This finding differs from previous research that showed that semantically inconsistent changes (e.g., a toothbrush appearing on a desk) are detected faster compared to semantically consistent changes (e.g., a pencil appearing on a desk; Hollingworth & Henderson, 2000; Stirk & Underwood, 2007) and that highly meaningful changes and changes of central interest to the scene were detected faster (O'Regan et al., 1999; Pringle et al., 2001; Rensink et al., 1997). One difference between the current study and previous research is that we did not deliberately incorporate inconsistent changes into the scenes and most of the changes could be considered semantically consistent. We also highlighted the change to participants and very likely forced attention to the changed feature, which may have not been included in the descriptions otherwise. As a consequence of explicitly drawing attention to the change, participants may have provided descriptions of the unmodified and modified scenes that were more dissimilar than if the change was not highlighted. Additionally, ratings of semantic similarity in our study were directly obtained from participants on a 6-point scale whereas experimenters in other studies binarily categorized the scene as either semantically inconsistent or semantically consistent, or as a central interest change or a marginal interest change.

One limitation to this study is that our measure of semantic similarity relies on linguistic descriptions of the scenes, yet participants certainly know more about these scenes than what they write. Thus, there is likely additional semantic similarity between the scenes that is not captured by these written responses. However, our goal with including this measure was to account for another factor, beyond low-level differences, that may have potentially predicted change detection performance.

Overall, these results suggest that while people may use semantic similarity in making their metacognitive judgements, semantic similarity based on linguistic descriptions of changes in scenes does not underlie the predictive relationship between metacognitive judgements and change blindness duration.

5. General discussion

We found that participants' judgements about their ability to detect clearly visible changes to visual scenes significantly predicted the duration of their change blindness when the changes occurred in a deliberate change blindness paradigm: changes rated as likely to be spotted were detected faster than changes rated as unlikely to be spotted. These judgements of change detection likelihood (made by comparing the two versions of the scene side-by-side) predicted change detection performance in a different task (using the "mud splash" flicker paradigm), even when low-level and high-level indicators of change magnitude did not. Although people are generally overconfident that they will notice a change in situations designed to evoke change blindness (Levin et al., 2000), our results show that people are aware that some changes are easier to detect than others. Thus, we have found evidence for both the generality of change-detection judgements and that change detection metacognition—how people *think* they will perform in a change blindness paradigm—tracks the relative difficulty of change blindness scenarios, rather than indiscriminately overestimating performance.

Our results have implications for understanding "change blindness blindness," in which people are unaware of their own failures of awareness, as well as for predicting future failures of awareness.

5.1. Why is change blindness surprising if it is predictable?

One of the most distinctive aspects of change blindness is that the changes in scenes that viewers fail to detect—often for several seconds or even minutes—are typically *detectable* changes (e.g., a building disappearing, a car replaced by a truck, or a person changing identity). People believe these changes are obvious, both to themselves and to others (Levin et al., 2000; Ortega et al., 2018), yet in demonstrations, they are surprised at their failure to readily detect the changes. Based on the surprising nature of change blindness, it seems reasonable to conclude that people have poor metacognition about change blindness, or at the very least, that they may be unable to predict which changes they would fail to notice. However, our results challenge this conclusion by showing that participants' judgements about the likelihood of detecting changes significantly predicted which changes to visual scenes are difficult to notice.

If people are aware of the relative difficulty of the changes and their change detection metacognition significantly predicts the magnitude of their change blindness, why then is change blindness a compelling demonstration of a failure of awareness? First, change detection predictions may become accurate only after participants have seen several example stimuli and can learn the distribution of change difficulty, and thus, if people only view one or two scenes (such as in a simple demonstration of change blindness), they would be unlikely to predict what they will fail to see. In Experiment 1, we found no significant difference when participants predicted change blindness in the first half of trials and in the last half of trials, indicating there was no metacognitive advantage with more exposure to different kinds of changes. Second, change detection predictions may reflect metacognition only about the relative likelihood of detecting the change and not about the absolute duration of time needed to detect the change, and thus people may be most surprised by how long it takes them to notice a change. Although the changes rated hardest to detect took approximately 3.5 s more than the changes rated easiest to detect in Experiment 1, the easiest changes still took participants about 8 s to detect. In the context of a classroom demonstration or TV ad, which have an average duration of 15 s (Ciccarelli, 2021), the surprise about one's own change blindness most likely stems from the fact that it requires several seconds at all to detect an obvious change right in front of one's eyes, and not that some changes may take a little longer to notice. Future studies may test this possibility more explicitly by asking participants to provide estimates of change blindness duration, although human duration judgements are influenced by many different factors, including environmental, physical, cognitive, and emotional factors (Zakay & Block, 1996) that may undermine the utility of gathering duration estimates. Using anchors is known to resolve some of the inaccuracies and biases when giving estimates of task duration (König, 2005) and it will be important to consider which anchors are appropriate, given the variance in change blindness duration for different images and different people.

5.2. The generality of metacognitive judgements of change blindness

Our results demonstrate that change detection judgements are general in two ways. First, participants' judgements about how likely they were to detect a change when comparing the unmodified and modified versions of the scenes side-by-side predicted their change blindness duration in the "mud splash" flicker paradigm—an entirely different task. Second, changedetection likelihood judgements from one group of participants predicted change blindness duration in an entirely different group of participants. In both cases, these judgements remained predictive of change blindness duration even when controlling for low-level and high-level indicators of change magnitude. Our results contrast with an earlier case in which judgements did not generalize from one task to another nor one group of participants to another (Levin et al., 2000). The generality of change detection likelihood judgements could be due to the properties of our stimuli and due to robust metacognition about how people *think* they will perform in a change blindness paradigm.

Although returning participants demonstrated knowledge about their own change detection abilities, when a new group of participants

(who had not previously participated in the change blindness experiment) provided change detection likelihood ratings, their ratings were just as effective at predicting change blindness in the returning participants. This result suggests that change detection metacognition depends more on properties intrinsic to images, rather than memory for past performance or high-level cognitive processes that may be subject to individual experience. Typically, when metacognition is assessed in a visual awareness paradigm, participants view relatively homogenous stimuli on each trial (i.e., a Gabor grating presented at threshold) and provide confidence ratings about their trial-wise performance (e.g., Jachs et al., 2015). This approach keeps low-level visual signals constant, while allowing high-level processes to vary (e.g., across time and across people). In these cases, it would be unlikely that one group of participants would be able to predict the trial-by-trial performance of a separate group of participants.

In contrast, in the current study, there was a tremendous amount of variability among the images: the visual nature of the change to be detected was variable, as was the semantic content of the images. Although high-level processes could still vary across time and across people in our study, the image-level variability may be the dominant input to participants' change detection likelihood ratings. However, low-level image properties such as the size or eccentricity of the change did not predict change detection likelihood ratings directly and change detection likelihood ratings remained predictive of change blindness duration even when these properties were controlled for. Moreover, when the similarity between written descriptions of the unmodified and modified versions of the scene was included, this measure of semantic similarity did not modulate the predictive relationship between change detection likelihood ratings and change blindness duration. In this case, semantic similarity did predict likelihood ratings directly, but again, the ratings remained predictive of change blindness duration when semantic similarity was controlled for.

Because the relationship between metacognitive judgements and change blindness was not explained by low-level visual properties or semantic similarity in the current study, this suggests that the input to change detection metacognition may rely on synthesizing different sources of visual information that are not captured by image-level properties. However, this does not necessarily mean that low-level visual properties or semantic similarity were never used. It is likely that these predictors were used to some degree, but not used consistently enough to show an effect in the current study. Furthermore, we relied on similarity ratings of written descriptions of unmodified and modified scenes as a measure of semantic similarity between scenes, which only captures some aspects of semantic similarity. It is also possible that many different factors, including some image-level properties that were not included in this study, can explain the relationship between metacognitive judgements and change blindness on an image-by-image basis. In fact, it seems very likely that the dominant factor people use to predict change blindness duration varies across images and potentially across people. Here, we were primarily interested in uncovering the general and consistent relationships between our predictors and change blindness duration.

5.3. Conclusion

We have demonstrated that assessing metacognition through simple change detection likelihood judgements is both general (insofar as ratings from one task predict change blindness in another task, and ratings from one group of people predict change blindness in another group) and effective (insofar as ratings explain unique variance in change blindness above and beyond image properties). Metacognitive judgements of this sort may thus be extremely useful in predicting future failures of awareness.

CRediT authorship contribution statement

Adam J. Barnas: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Emily J. Ward:** Conceptualization, Validation, Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

None.

Acknowledgements

This work was supported by the Wisconsin Alumni Research Foundation (WARF) and the University of Wisconsin-MadisonOffice of the Vice Chancellor for Research and Graduate Education.

Appendix A. Supplementary data

The data and materials for all experiments are publicly available at the following repository: https://github.com/adamjbarnas/MetaCognition_Archive. Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2022.105208.

References

Bateman, D., Eaton, J. W., Wehbring, R., & Hauberg, S. (2015). *The GNU octave 4.0 reference manual 1/2: Free your numbers*. London: Samurai Media Limited.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed models using lme4. *Journal of Statistical Software, 67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Beck, M. R., Angelone, B. L., & Levin, D. T. (2004). Knowledge about the probability of change affects change detection performance. *Journal of Experimental Psychology: Human Perception and Performance, 30*, 778–791. https://doi.org/10.1037/0096-1523.30.4.778

Beck, M. R., Levin, D. T., & Angelone, B. (2007). Change blindness blindness: Beliefs about the roles of intention and scene complexity in change detection. *Consciousness and Cognition, 16*, 31–51. https://doi.org/10.1016/j.concog.2006.01.003

Block, R. A., & Zakay, D. (1997). Prospective and retrospective duration judgments: A meta-analytic review. *Psychonomic Bulletin & Review, 4*, 184–197. https://doi.org/10.3758/BF03209393

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*, 433–436. https://doi.org/10.1163/156856897X00357

Buhrmester, M. D., Talaifar, S., & Gosling, S. D. (2018). An evaluation of Amazon's Mechnical Turk, its rapid rise, and its effect use. *Perspectives on Psychological Science, 13*, 149–154. https://doi.org/10.1177/1745691617706516

Ciccarelli, D. (2021, May 19). What is the most effective length for a TV commercial? Voices. https://www.voices.com/blog/effective_length_for_tv_commercials/.

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS One, 8*, Article e57410. https://doi.org/10.1371/journal.pone.0057410

Davenport, J. L., & Potter, M. C. (2004). Scene consistency in object and background perception. *Psychological Science, 15*, 559–564, 10.111/j.0956-7976.2004.00719.x.

Ehinger, K. A., Allen, K., & Wolfe, J. M. (2016). Change blindness for cast shadows in natural scenes: Even informative shadow changes are missed. *Attention, Perception, & Psychophysics, 78*, 978–987. https://doi.org/10.3758/s13414-015-1054-7

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience, 8*, 1–9. https://doi.org/10.3389/fnhum.2014.00443

Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., & Rees, G. (2010). Relating introspective accuracy to individual differences in brain structure. *Science, 329*, 1541–1543. https://doi.org/10.1126/science.1191883

Galvin, S. J., Podd, J. V., Drga, V., & Whitmore, J. (2003). Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions. *Psychonomic Bulletin & Review, 10*, 843–876. https://doi.org/10.3758/bf03196546

Gaspar, J. G., Neider, M. B., Simons, D. J., McCarley, J. S., & Kramer, A. F. (2013). Change detection: Training and transfer. *PLoS One, 8*, Article e67781. https://doi.org/10.1371/journal.pone.0067781

Halkjelsvik, T., & Jørgensen, M. (2012). From origami to software development: A review of studies of judgment-based predictions of performance time. *Psychological Bulletin, 138*, 238–271. https://doi.org/10.1037/a0025996

Hollingworth, A. (2003). Failures of retrieval and comparison constrain change detection in natural scenes. *Journal of Experimental Psychology: Human Perception and Performance, 29*, 388–403. https://doi.org/10.1037/0096-1523.29.2.388

Hollingworth, A., & Henderson, J. M. (2000). Semantic informativeness mediates the detection of changes in natural scenes. *Visual Cognition, 7*, 213–235. https://doi.org/10.1080/135062800394775

Jachs, B., Blanco, M. J., Grantham-Hill, S., & Soto, D. (2015). On the independence of visual awareness and metacognition: A signal detection theoretic analysis. *Journal of*

Experimental Psychology. Human Perception and Performance, 41, 269–276. https://doi.org/10.1037/xhp0000026
Johansson, P., Hall, L., Sikström, S., & Olsson, A. (2005). Failture to detect mismatches between intention and outcome in a simple detection task. *Science, 310*, 116–119. https://doi.org/10.1126/science.1111709
Komsta, L., & Novomestky, F. (2015). *Moment, cumulants, skewness, kurtosis, and related tests. R package version 0.14*.
König, C. J. (2005). Anchors distort estimates of expected duration. *Psychological Reports, 96*, 253–256. https://doi.org/10.2466/pr0.96.2.253-256
Kunimoto, C., Miller, J., & Pashler, H. (2001). Confidence and accuracy of near-threshold discrimination responses. *Consciousness and Cognition, 10*, 294–340. https://doi.org/10.1006/ccog.2000.0494
LaPointe, M. R. P., Lupiáñez, J., & Milliken, B. (2013). Context congruency effects in change detection: Opposing effects on change detection and identification. *Visual Cognition, 21*, 99–122. https://doi.org/10.1080/13506285.2013.787133
Lau, H. C., & Passingham, R. E. (2006). Relative blindsight in normal observers and the neural correlate of visual consciousness. *Proceedings of the National Academy of Sciences, 103*, 18763–18768. https://doi.org/10.1073/pnas.0607716103
de Leeuw, J. R. (2015). jsPsych: A JaveScript library for creating behavioral experiments in a web browser. *Behavior Research Methods, 47*, 1–12. https://doi.org/10.3758/s13428-014-0458-y
Levin, D. T. (2002). Change blindness blindness as visual metacognition. *Journal of Consciousness Studies, 9*(5–6), 111–130.
Levin, D. T., Drivdahl, S. B., Momen, N., & Beck, M. R. (2002). False predictions about the detectability of visual changes: The role of beliefs about attention, memory, and the continuity of attend objects in causing change blindness blindness. *Consciousness and Cognition, 11*, 507–527. https://doi.org/10.1016/s1053-8100(02)00020-x
Levin, D. T., Momen, N., Drivdahl, S. B., & Simons, D. J. (2000). Change blindness blindness: The metacognitive error of overestimating change-detection ability. *Visual Cognition, 7*, 397–412. https://doi.org/10.1080/135062800394865
Levin, D. T., & Simons, D. J. (1997). Failure to detect changes to attended objects in motion pictures. *Psychonomic Bulletin & Review, 4*, 501–506. https://doi.org/10.3758/BF03214339
Levin, D. T., & Simons, D. J. (2000). Perceiving stability in a changing world: Combining shots and integrating views in motion pictures and the real world. *Mead Psychology, 2*, 357–380. https://doi.org/10.1207/S1532785XMEP0204_03
Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature, 390*, 279–281. https://doi.org/10.1038/36846
Ma, L.-Q., Xu, K., Wong, T.-T., Jiang, B.-Y., & Hu, S.-M. (2013). Change blindness images. *IEEE Transactions on Visualization and Computer Graphics, 19*, 1808–1819. https://doi.org/10.1109/TVCG.2013.99
Mandler, J. M., & Ritchey, G. H. (1977). Long-term memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory, 3*, 386–396. https://doi.org/10.1037/0278-7393.3.4.386
Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition, 21*, 422–430, 10.1016.j.concog.2011.09.021.
McAnally, K. I., Morris, A. P., & Best, C. (2017). Metacognitive monitoring and control in visual change detection: Implications for situation awareness and cognitive control. *PLoS One, 12*, Article e0176032. https://doi.org/10.1371/journal.pone.0176032
Metcalfe, J., & Shimamura, A. P. (1994). *Metacognition: Knowing about knowing.* Cambridge, MA: MIT Press.
Mitroff, S. R., Simons, D. J., & Levin, D. T. (2004). Nothing compares 2 views: Change blindness can occur despite preserved access to the changed information. *Perception & Psychophysics, 66*, 1268–1281. https://doi.org/10.3758/bf03194997
O'Regan, J. K., Rensink, R. A., & Clark, J. J. (1999). Change-blindness as a result of "mudsplashes". *Nature, 398*, 34. https://doi.org/10.1038/17953
Ortega, J., Montañes, P., Barnhart, A., & Kuhn, G. (2018). Exploiting failures in metacognition through magic: Visual awareness as a source of visual metacognition bias. *Consciousness and Cognition, 65*, 152–168. https://doi.org/10.1016/j.concog.2018.08.008
Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., … Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods, 51*, 195–203. https://doi.org/10.3758/s13428-018-01193-y
Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*, 437–442. https://doi.org/10.1163/156856897X00366
Phillips, W. A. (1974). On the distinction between sensory storage and short-term visual memory. *Perception & Psychophysics, 16*, 283–290. https://doi.org/10.3758/BF03203943
Pringle, H. L., Irwin, D. E., Kramer, A. F., & Atchley, P. (2001). The role of attentional breadth in perceptual change detection. *Psychonomic Bulletin & Review, 8*, 89–95. https://doi.org/10.3758/BF03196143
Rensink, R. A., O'Regan, J. K., & Clark, J. J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science, 8*, 368–373. https://doi.org/10.1111/j.1467-9280.1997.tb00427.x
Rensink, R. A., O'Regan, J. K., & Clark, J. J. (2000). On the failure to detect changes in scenes across brief disruptions. *Visual Cognition, 7*, 127–145. https://doi.org/10.1080/135062800394720
Roy, M. M., & Christenfeld, N. J. S. (2008). Effect of task length on remembered and predicted duration. *Psychonomic Bulletin & Review, 15*, 202–207. https://doi.org/10.3758/PBR.15.1.202
Roy, M. M., Christenfeld, N. J. S., & McKenzie, C. R. M. (2005). Underestimating the duration of future events: Memory incorrectly used or memory bias? *Psychological Bulletin, 131*, 738–756. https://doi.org/10.1037/0033-2909.131.5.738
Sareen, P., Ehinger, K. A., & Wolfe, J. M. (2015). CB database: A change blindness database for objects in natural indoor scenes. *Behavior Research Methods, 48*, 1343–1348. https://doi.org/10.3758/s13428-015-0640-x
Scholl, B. J., Simons, D. J., & Levin, D. T. (2004). 'Change blindness' blindness: An implicit measure of a metacognitive error. In D. T. Levin (Ed.), *Thinking and seeing: Visual metacognition in adults and children* (pp. 145–164). Cambridge, MA: MIT Press.
Scott, R. B., Dienes, Z., Barrett, A. B., Bor, D., & Seth, A. K. (2014). Blind insight: Metacognitive discrimination despite chance task performance. *Psychological Science, 25*, 2199–2208. https://doi.org/10.1177/0956797614553944
Scott-Brown, K. C., Baker, M. R., & Orbach, H. S. (2000). Comparison blindness. *Visual Cognition, 7*, 253–267. https://doi.org/10.1080/135062800394793
Simons, D. J., & Levin, D. T. (1998). Failure to detect changes to people during a real-world interaction. *Psychonomic Bulletin & Review, 5*, 644–649. https://doi.org/10.3758/BF03208840
Simons, D. J., & Rensink, R. A. (2005). Change blindness: Past, present, and future. *Trends in Cognitive Science, 9*, 16–20. https://doi.org/10.1016/j.tics.2004.11.006
Snodgrass, M., Bernat, E., & Shevrin, H. (2004). Unconscious perception: A model-based approach to method and evidence. *Perception & Psychophysics, 66*, 846–867. https://doi.org/10.3758/BF03194978
Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs: General and Applied, 74*, 1–29. https://doi.org/10.1037/h0093759
Standing, L. (1973). Learning 10000 pictures. *Quarterly Journal of Experimental Psychology, 25*, 207–222. https://doi.org/10.1080/14640747308400340
Stirk, J. A., & Underwood, G. (2007). Low-level visual saliency does not predict change detection in natural scenes. *Journal of Vision, 7*, 1–10. https://doi.org/10.1167/7.10.3
Türkan, B. N., İyilikci, O., & Amado, S. (2021). Ways of processing semantic information during different change detection tasks. *Visual Cognition, 29*, 366–378. https://doi.org/10.1080/13506285.2021.1927276
Vandenbroucke, A. R. E., Sligte, I. G., Barrett, A. B., Seth, A. K., Fahrenfort, J. J., & Lamme, V. F. (2014). Accurate metacognition for visual sensory memory representations. *Psychological Science, 25*, 861–873. https://doi.org/10.1177/0956797613516146
Varakin, D. A., Levin, D. T., & Collins, K. M. (2007). Comparison and representation failures both cause real-world change blindness. *Perception, 36*, 737–749. https://doi.org/10.1068/p5572
Vogel, E. K., Woodman, G. F., & Luck, S. J. (2001). Storage of features, conjunctions, and objects in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance, 27*, 92–114. https://doi.org/10.1037/0096-1523.27.1.92
Zakay, D., & Block, R. A. (1996). The role of attention in time estimation processes. *Advances in Psychology, 115*, 143–164. https://doi.org/10.1016/S0166-4115(96)80057-4