**Additional Experiment 1 Results**

**Robustness check**

Decisions about cutoffs, exclusions, and analyses were made prior to analyzing the data. These decisions were established in order to remove inaccurate trials, outlier trials, as well as participants who may not have been fully engaged in the task (based on their performance on attention check trials). The log-transformations were the only choice that was data-derived, because we plotted our distributions and measured skewness. Based on those analyses, we log-transformed change blindness duration to eliminate skewness and to achieve as close as a normal distribution.

To verify that our findings reported in our manuscript do not depend on these decisions, we performed the following robustness analyses for Experiment 1 in which none of our exclusion criteria were applied and in which we look at raw and log-transformed data. In this analysis, we include participants who missed attention check trials in the change blindness task and the likelihood ratings task, as well as all inaccurate trials (i.e., the participant did not click on the changing object) and trials that were ±3 SDs from the participant’s mean RT (i.e., outliers) in the change blindness task. (Note that trials in the change blindness task with no response or that timed out were still excluded since change blindness duration was not recorded.) These data were entered into our primary linear mixed effects model for predicting change blindness duration with change detection likelihood ratings as a fixed ordinal effect and participant, image pair, and stimulus set as random intercepts.

When all participants and trials were analyzed, change detection likelihood ratings significantly predicted log-transformed change blindness duration, *β* = -0.12, *p* < .001, 95% CI = [-0.18, -0.06] and raw change blindness duration, *β* = -0.13, *p* < .001, 95% CI = [-0.20, -0.06], such that changes rated as likely to be spotted were detected faster than changes rated as unlikely to be spotted. These results are consistent–although slightly weaker–with what we report in our manuscript (i.e., when these exclusions and transformations were applied), *β* = -0.18, *p* < .001, 95% CI = [-0.25, -0.11].

Since there were a lot more attention check failures in Experiment 2, we wanted to conduct similar robustness analyses to determine if our findings stemmed from our exclusion criteria. We re-analyzed the data without excluding any of the new participants who failed attention check trials in 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), size, and eccentricity of the change as fixed effects and participant, image pair, and stimulus set as random intercepts. The results of these analyses, and the result we report in our manuscript, are laid out in the table below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Likelihood ratings** | Size | Eccentricity | Interaction |
| Robustness (log RT) | ***β* = -0.02, *p* = .396, 95% CI = [-0.06, 0.02]** | *β* = -0.08, *p* < .05, 95% CI = [-0.15, -0.02] | *β* = 0.08, *p* < .001, 95% CI = [0.04, 0.13] | *β* = -0.02, *p* = .487, 95% CI [-0.08, 0.04] |
| Reported (log RT) | ***β* = -0.05 *p* < .05, 95% CI = [-0.1, -0.004]** | *β* = -0.10, *p* < .01, 95% CI = [-0.17, -0.03] | *β* = 0.08, *p* < .01, 95% CI = [0.03, 0.13] | *β* = 0.002, *p* = .941, 95% CI = [-0.06, 0.07] |
| Percentage change | **1.5%** | 0.25% | 0% | -1.1% |

Three of these findings (significant effects of size and eccentricity, and no significant interaction between likelihood ratings and group) from this robustness analysis are consistent with the results reported in our manuscript. However, likelihood ratings were not significantly predictive of change blindness duration in the robustness analysis (bolded). These results suggest that the predictive relationship we find between likelihood ratings and change blindness duration depends on attentive participants, hence the importance of attention checks.

Generally, the results of these robustness checks confirm that the pattern of results we report in our manuscript do not depend on our predetermined exclusion criteria, cutoffs, or transformations.

**Variability in change blindness duration and likelihood ratings**

**High variability**

Here, we wanted to determine whether our results were driven by a handful of participants and/or images with sufficient variability in response latencies on the change detection task or with sufficient variability in likelihood ratings. First, we performed a median-split analysis of **participants** based on their variability in raw **change blindness duration** on the change detection task. We calculated the standard deviation of change detection duration for each participant, computed the median standard deviation (*Mdn* = 2.88 seconds), and selected participants who had standard deviations greater than the median (that is, people with the most variability in change blindness duration; *n* = 107 participants). We then re-ran our primary analysis from Experiment 1 (a linear mixed effects model predicting change blindness duration with likelihood ratings as an ordinal fixed effect and participant, image, and stimulus set as random intercepts). Likelihood ratings significantly predicted change blindness duration, *β* = -0.26, *p* < .001, 95% CI = [-0.36, -0.16], such that changes rated as likely to be spotted were detected faster than changes rated as unlikely to be spotted.

Next, we analyzed **participants** based on their variability in **likelihood ratings**. We calculated the standard deviation of likelihood ratings for each participant, computed the median standard deviation (*Mdn* = 1.15), and selected participants who had a standard deviation greater than the median (that is, people with the most variability in likelihood ratings; *n* = 107 participants). Fifty-eight of these participants were also included in our analysis of participants and variability in raw change blindness duration. We then re-ran our primary analysis from Experiment 1 (a linear mixed effects model predicting change blindness duration with likelihood ratings as an ordinal fixed effect and participant, image, and stimulus set as random intercepts). Likelihood ratings significantly predicted change blindness duration, *β* = -0.18, *p* < .001, 95% CI = [-0.27, -0.09], such that changes rated as likely to be spotted were detected faster than changes rated as unlikely to be spotted.

We then analyzed **images** based on their variability on raw **change blindness duration** on the change detection task. We calculated the standard deviation of change detection duration for each image, computed the median standard deviation (*Mdn* = 3.78 seconds), and selected images who had standard deviations greater than the median (that is, images with the most variability in change blindness duration; *n* = 239 images). We then re-ran our primary analysis from Experiment 1 (a linear mixed effects model predicting change blindness duration with likelihood ratings as an ordinal fixed effect and participant, image, and stimulus set as random intercepts). Likelihood ratings significantly predicted change blindness duration, *β* = -0.19, *p* < .001, 95% CI = [-0.29, -0.10], such that changes rated as likely to be spotted were detected faster than changes rated as unlikely to be spotted.

Finally, we analyzed **images** based on their variability on **likelihood ratings**. We calculated the standard deviation of likelihood ratings for each image, computed the median standard deviation (*Mdn* = 0.99), and selected images who had a standard deviation greater than the median (that is, images with the most variability in likelihood ratings; *n* = 238 images). One hundred twenty-two of these images were also included in our analysis of images and variability in raw change blindness duration. We then re-ran our primary analysis from Experiment 1 (a linear mixed effects model predicting change blindness duration with likelihood ratings as an ordinal fixed effect and participant, image, and stimulus set as random intercepts). Likelihood ratings significantly predicted change blindness duration, *β* = -0.18, *p* < .001, 95% CI = [-0.27, -0.08], such that changes rated as likely to be spotted were detected faster than changes rated as unlikely to be spotted.

Based on the combined results, it appears that the relationship between change blindness duration and likelihood ratings occurs even when we limit our analysis to both participants and images that exhibit sufficient variability in each measure.

**Low variability**

Here, we wanted to determine whether our results were driven by a handful of participants and/or images with insufficient variability in response latencies on the change detection task or with insufficient variability in likelihood ratings. First, we analyzed **participants** based on their variability in raw **change blindness duration** on the change detection task. We calculated the standard deviation of change detection duration for each participant, computed the median standard deviation (*Mdn* = 2.88 seconds), and selected participants who had standard deviations less than the median (that is, people with the least variability in change blindness duration; *n* = 107 participants). We then re-ran our primary analysis from Experiment 1 (a linear mixed effects model predicting change blindness duration with likelihood ratings as an ordinal fixed effect and participant, image, and stimulus set as random intercepts). Likelihood ratings significantly predicted change blindness duration, *β* = -0.16, *p* < .001, 95% CI = [-0.25, -0.07], such that changes rated as likely to be spotted were detected faster than changes rated as unlikely to be spotted.

Next, we analyzed **participants** based on their variability in **likelihood ratings**. We calculated the standard deviation of likelihood ratings for each participant, computed the median standard deviation (*Mdn* = 1.15), and selected participants who had a standard deviation less than the median (that is, people with the least variability in likelihood ratings; *n* = 107 participants). Fifty-eight of these participants were also included in our analysis of participants and variability in raw change blindness duration. We then re-ran our primary analysis from Experiment 1 (a linear mixed effects model predicting change blindness duration with likelihood ratings as an ordinal fixed effect and participant, image, and stimulus set as random intercepts). Likelihood ratings significantly predicted change blindness duration, *β* = -0.32, *p* < .001, 95% CI = [-0.44, -0.20], such that changes rated as likely to be spotted were detected faster than changes rated as unlikely to be spotted.

We then analyzed **images** based on their variability on raw **change blindness duration** on the change detection task. We calculated the standard deviation of change detection duration for each image, computed the median standard deviation (*Mdn* = 3.78 seconds), and selected images who had standard deviations less than the median (that is, images with the least variability in change blindness duration; *n* = 239 images). We then re-ran our primary analysis from Experiment 1 (a linear mixed effects model predicting change blindness duration with likelihood ratings as an ordinal fixed effect and participant, image, and stimulus set as random intercepts). Likelihood ratings significantly predicted change blindness duration, *β* = -0.12, *p* < .05, 95% CI = [-0.22, -0.03], such that changes rated as likely to be spotted were detected faster than changes rated as unlikely to be spotted.

Finally, we analyzed **images** based on their variability on **likelihood ratings**. We calculated the standard deviation of likelihood ratings for each image, computed the median standard deviation (*Mdn* = 0.99), and selected images who had a standard deviation less than the median (that is, images with the least variability in likelihood ratings; *n* = 237 images). One hundred twenty-one of these images were also included in our analysis of images and variability in raw change blindness duration. We then re-ran our primary analysis from Experiment 1 (a linear mixed effects model predicting change blindness duration with likelihood ratings as an ordinal fixed effect and participant, image, and stimulus set as random intercepts). Likelihood ratings significantly predicted change blindness duration, *β* = -0.27, *p* < .001, 95% CI = [-0.38, -0.16], such that changes rated as likely to be spotted were detected faster than changes rated as unlikely to be spotted.

Based on the combined results, it appears that the relationship between change blindness duration and likelihood ratings occurs even when we limit our analysis to both participants and images that exhibit insufficient variability in each measure.

**Variability in predictions**

To analyze the variability in the relationship between change blindness duration and change detection duration, we calculated the correlation between each measure for every participant and averaged those correlations. When we do so for Experiment 1, we get an average correlation of -0.12 (*SD* = 0.23). The corresponding distribution is plotted below (the dotted line corresponds to the average correlation). As depicted, there is a small negative association such that changes rated as likely to be spotted were detected faster than changes rated as unlikely to be spotted.



To determine how much of the effect is driven by a small subset of participants/items (those with sufficient variability), we approached this in two ways. First, we removed participants and images with the top 10% of variability in change blindness duration and likelihood ratings and analyzed the data in our primary model from Experiment 1 (a linear mixed effects model predicting change blindness duration with likelihood ratings as an ordinal fixed effect and participant, image, and stimulus set as random intercepts) and plotted individual correlations between change blindness duration and likelihood rating against the amount of variability that participants exhibited in change blindness duration and likelihood ratings.

For the first approach, we removed **participants** with the top 10% of variability in **change blindness duration** and analyzed the remaining 90% of participants. Likelihood ratings significantly predicted change blindness duration, *β* = -0.17, *p* < .001, 95% CI = [-0.24, -0.09]. Next, we removed **participants** with the top 10% of variability in **likelihood ratings** and analyzed the remaining 90% of participants. Likelihood ratings significantly predicted change blindness duration, *β* = -0.25, *p* < .001, 95% CI = [-0.33, -0.17]. We then removed **images** with the top 10% of variability in raw **change blindness duration** and analyzed the remaining 90% of images. Likelihood ratings significantly predicted change blindness duration, *β* = -0.15, *p* < .001, 95% CI = [-0.23, -0.08]. Finally, we removed **images** with the top 10% of variability in **likelihood ratings** and analyzed the remaining 90% of images. Likelihood ratings significantly predicted change blindness duration, *β* = -0.23, *p* < .001, 95% CI = [-0.31, -0.16]. For all four results, changes rated as likely to be spotted were detected faster than changes rated as unlikely to be spotted, suggesting that the effects do not appear to be driven by a small subset of participants/images with sufficient variability.

For the second approach, we plotted individual correlations between change blindness duration and likelihood rating against the amount of variability that participants exhibited separately in change blindness duration and likelihood ratings. For change blindness duration, there was a significant negative correlation. *r*(209) = -.16, *p* = .018, meaning that participants with a greater negative correlation between change blindness duration and likelihood rating (indicating that changes rated as likely to be spotted were detected faster than changes rating as unlikely to be spotted) exhibited more variability in change blindness duration.



However, there was no significant correlation for the likelihood ratings (*p* = .971), suggesting that variability in likelihood ratings is not associated with the pattern of our main result – that changes rated as likely to be spotted are detected faster than changes rated as unlikely to be spotted.



**Variability in stimulus sets**

To determine the extent that the effects were driven by certain stimulus sets (particularly those where there was more variability in change detection times), we ran the fixed effects model from Experiment 1 (predicting changing blindness duration with change detection likelihood ratings as a fixed effect and participant, image pair, and stimulus set as random intercepts) on the data three times–without the Wolfe1 stimulus set, without the Wolfe2 stimulus set, and without both Wolfe1 and Wolfe2 stimulus sets. Change detection likelihood ratings continued to significantly predict change blindness duration when the Wolfe1 stimulus set was removed, *β* = -0.17, *p* < .001, 95% CI = [-0.24, -0.09]. (Note: each dot represents a single image).



Change detection likelihood ratings also continued to significantly predict change blindness duration when the Wolfe2 stimulus set was removed, *β* = -0.14, *p* = .002, 95% CI = [-0.22, -0.05]. (Note: each dot represents a single image).



However, when both Wolfe1 and Wolfe2 stimulus sets were removed, change detection likelihood ratings did not significantly predict change blindness duration, *β* = -0.06, *p* = .301, 95% CI = [-0.17, 0.05]. (Note: each dot represents a single image).



Based on these results, the Wolfe1 and Wolfe2 stimulus sets appear to be important for the relationship between likelihood ratings and change blindness duration. We acknowledge this could be due to the variance within these sets, but we also want to point out that when we remove Wolfe1 AND Wolfe2 from the analysis, we have reduced our data by more than 75% (482 images to 117 images), which could also lead to this non-significant relationship.

**Trials with unusually long detection times**

To explore how trials with unusually long detection times contributed to our results, we excluded trials exceeding 15 seconds. This resulted in 593 trials excluded, or approximately 12% of total trials. Change detection likelihood ratings continued to significantly predict change blindness duration when trials with unusually long detection times (greater than 15 seconds) were excluded, *β* = -0.15, *p* < .001, 95% CI = [-0.21, -0.08].



We also extended this analysis by having a more conservative definition of “unusually long”: when the cut-off was set to 8 seconds (excluding 2692 trials, or approximately 54% of total trials), change detection likelihood ratings continued to significantly predict change blindness duration, *β* = -0.07, *p* = .025, 95% CI = [-0.13, -0.01].

**Split-half reliability**

We performed a split-half reliability analysis to assess how measurement error sets an upper limit on the predictiveness of metacognition. We randomly split each participant’s data in half (by first shuffling trial order and then dividing the trials into halves). Since this split was random, we did it 10 times to eliminate any ordering effects within participants. We then calculated the correlation between change blindness duration and likelihood ratings and computed an average correlation for each half for each participant. The correlation between each half was significant, *r*(203) = 0.85, *p* < .001, suggesting that all images contribute equally to the predictiveness of likelihood ratings on change blindness duration.

**Remembering easy and hard changes**

To determine whether the easy/hard follow-up analysis in Experiment 1 (Section 2.2) was dependent on selecting the middle 12 images, we ran the same easy/hard analysis on the middle 10, 8, 6, and 4 trials, as well as the corresponding extreme 14, 16, 18, and 20 trials, respectively. The results of each model are below.

1. Middle **10** trials: *β* = -0.09, *p* < .05, 95% CI = [-0.15, -0.02];

Extreme 14 trials: *β* = -0.27, *p* < .001, 95% CI = [-0.36, -0.17]

1. Middle **8** trials: *β* = -0.07, *p* < .05, 95% CI = [-0.14, -0.008];

Extreme 16 trials:   *β* = -0.25, *p* < .001, 95% CI = [-0.34, -0.15]

1. Middle **6** trials: *β* = -0.08, *p* < .05, 95% CI = [-0.15, -0.01];

Extreme 18 trials:  *β* = -0.25, *p* < .001, 95% CI = [-0.34, -0.16]

1. Middle **4** trials: *β* = -0.05, *p* = .170, 95% CI = [-0.11, 0.02];

Extreme 20 trials: *β* = -0.24, *p* < .001, 95% CI = [-0.33, -0.16]

In general, these results suggest that the analysis is not dependent on the number of middle trials included in the model. However, when a small number of middle trials are analyzed (in this case, 4), the result becomes non-significant.