

Running Head: Measurement error in unconscious cognition

**IN PRESS: Psychonomic Bulletin & Review**

**24/03//2021**

**Raising Awareness about Measurement Error in Research  
on Unconscious Mental Processes**

Miguel A. Vadillo<sup>1</sup>, Simone Malejka<sup>2</sup>,  
Daryl Y. H. Lee<sup>2</sup>, Zoltan Dienes<sup>3</sup>, & David R. Shanks<sup>2</sup>

<sup>1</sup>Departamento de Psicología Básica, Universidad Autónoma de Madrid, Spain

<sup>2</sup>Department of Experimental Psychology, University College London, UK

<sup>3</sup>School of Psychology, University of Sussex, UK

Mailing address:

Miguel A. Vadillo  
Departamento de Psicología Básica  
Facultad de Psicología  
Universidad Autónoma de Madrid  
28049 Madrid, Spain

e-mail: miguel.vadillo@uam.es

Autor note: MAV was supported by grants 2016-T1/SOC-1395 and 2020-5A/SOC-19723 (Comunidad de Madrid, Programa de Atracción de Talento Investigador) and PSI2017-85159-P (Agencia Estatal de Investigación y Fondo Europeo de Desarrollo Regional). All the authors were supported by grant ES/P009522/1 from the UK Economic and Social Research Council. We are indebted to Ben Colagiuri and Evan Livesey for sharing their data with us. Correspondence concerning this article should be addressed to Miguel A. Vadillo, Departamento de Psicología Básica, Facultad de Psicología, Universidad Autónoma de Madrid, 28049 Madrid, Spain. E-mail: miguel.vadillo@uam.es

**Abstract**

Experimental psychologists often neglect the poor psychometric properties of the dependent measures collected in their studies. In particular, a low reliability of measures can have dramatic consequences for the interpretation of key findings in some of the most popular experimental paradigms, especially when strong inferences are drawn from the absence of statistically significant correlations. In research on unconscious cognition, for instance, it is commonly argued that the lack of a correlation between task performance and measures of awareness or explicit recollection of the target stimuli provides strong support for the conclusion that the cognitive processes underlying performance must be unconscious. Using contextual cuing of visual search as a case study, we show that given the low reliability of the dependent measures collected in these studies, it is usually impossible to draw any firm conclusion about the unconscious character of this effect from correlational analyses. Furthermore, both a psychometric meta-analysis of the available evidence and a cognitive-modeling approach suggest that, in fact, we should expect to see very low correlations between performance and awareness at the empirical level, even if both constructs are perfectly related at the latent level. Convincing evidence for the unconscious character of contextual cuing and other effects will most likely demand richer and larger datasets, coupled with more powerful analytic approaches.

**Keywords:** Contextual cuing; Meta-analysis; Reliability; Unconscious learning

Experimental psychologists typically care very little about the psychometric properties of their measures. This is no coincidence, as in fact some of the most popular experimental tasks used in psychological research yield strong effects despite being unreliable (Hedge, Powell, & Sumner, 2018). If we ask a sample of participants to complete the Stroop task, for instance, most of them will show the basic effect (Haaf & Rouder, 2019). However, if we now ask the same participants to complete the task for a second time, their performance across sessions will only be weakly correlated, at best (Enkavi et al., 2019; Hedge et al., 2018). The Stroop and other popular experimental paradigms produce very strong effects at the group level, as a result of aggregating data from many trials and relying on within-participants control conditions. This effectively removes many sources of noise in the data, including not only random (or error) variance, but also a significant proportion of variance due to individual differences across participants.

In practice, researchers working in the ‘experimental tradition’ (Cronbach, 1957) have strong incentives to reduce the impact of individual differences as much as possible. Reducing every possible source of variation increases effect sizes and, consequently, improves statistical power without recourse to larger samples (Brysbaert, 2019; De Schryver, Hughes, Rosseel, & De Houwer, 2016; Parsons, Kruijt, & Fox, 2019; Zimmerman & Zumbo, 2015). Whether or not this approach also reduces the reliability of the dependent measures is most often inconsequential. But occasionally experimental researchers do submit their data to types of analyses that hinge critically on their psychometric properties, sometimes unknowingly so. For instance, if researchers are interested in whether performance in the Stroop, flanker, and other tasks make use of a common skill or resource to inhibit representations or responses, then reliability does become an issue, because the dependent measures provided by these tasks cannot be

expected to correlate with each other when they fail to capture systematic differences across individuals (Rey-Mermet, Gade, & Oberauer, 2018; Rouder, Kumar, & Haaf, 2019).

As we will show in the present article, neglecting the reliability of dependent measures has been particularly consequential and misleading in areas of research where strong inferences are made on the basis of null correlations, particularly in the study of unconscious mental processes. A common strategy for demonstrating that a mental process is unconscious is to show that there is no significant correlation between participants' task performance and their level of awareness of the stimuli or of the contingencies that drive performance in that task. As an illustrative example of this approach and its shortcomings, consider a recent study by Berkovitch and Dehaene (2019). Participants were asked to report whether a series of target words were nouns or verbs. Immediately before each target word, a masked and putatively subliminal prime (also a noun or a verb) was presented for 100 ms. The authors found a standard compatibility effect between primes and targets: Participants responded faster to target words that were preceded by primes of the same grammatical category. Most interestingly, the size of this effect was not positively correlated with participants' ability to discriminate the primes as nouns or verbs in a visibility task conducted at the end of the experiment. *Prima facie*, this result, replicated in five experiments, appears to provide compelling evidence that the priming effect observed in responses to target words must have been unconscious. If the effect depended on conscious perception of the primes, then participants who were better able to perceive these primes in the visibility test should also have shown stronger priming effects.

A moment's thought, however, reveals that this logic relies on the (implausible) assumption that the measures of visibility and priming taken in the experiment are free from measurement error or, in other words, are perfectly reliable. Let us imagine that, contrary to this assumption, the reliability of these measures was rather low and that most

of the variation in visibility and priming across participants was due to measurement error. In that case, we would have little reason to expect that participants who obtained a high score in the visibility test were really more able to perceive the primes consciously than participants with lower scores, because most inter-individual differences would be driven by noise only. Because of this, there would also be little reason to expect a positive correlation between these scores and the size of priming effects across participants, even if they are related at the latent level. Unfortunately, Berkovitch and Dehaene (2019) did not report any information about the reliability of their dependent variables, rendering it impossible to assess whether the lack of correlation between visibility scores and priming was really due to the unconscious character of priming, as argued by the authors, or simply to the low reliability of both dependent measures.

Of course, Berkovitch and Dehaene (2019) are not alone. Countless studies in the domain of implicit cognition rely on the same logic to assess whether a cognitive process is unconscious (e.g., Colagiuri & Livesey, 2016; Hedger, Garner, & Adams, 2019; Jiang, Sha, & Sisk, 2018; Salvador et al., 2018), almost always without reporting the reliability of the dependent variables or even considering its critical role in the interpretation of statistical analyses. For pragmatic reasons, the present article will focus on a specific implicit learning effect where this analytic approach has been used extensively, namely, contextual cuing of visual search (Chun & Jiang, 1998; Sisk, Remington, & Jiang, 2019), but the arguments are general.

In a typical contextual cuing experiment, participants are instructed to search for a rotated T-shaped target among a number of L-shaped distractors and report its left/right orientation. Some search displays are repeated several times during the experiment (repeated condition), while others are presented just once (new condition), although participants are not instructed about this manipulation. Across blocks, search times become

substantially faster for repeated than for new patterns, showing that repeated exposure to the search displays results in long-lasting learning (a substantial majority of participants individually show such an effect). This learning effect is typically assumed to be unconscious because at the end of the experiment participants seem to be unable to discriminate the repeated from the new patterns or to recall explicitly the location on the screen where the target was located in repeated displays (Chun & Jiang, 2003; Goujon, Didierjean, & Thorpe, 2015; Sisk et al., 2019).

As we have noted elsewhere (e.g., Dienes, 2015; Smyth & Shanks, 2008; Vadillo, Konstantinidis, & Shanks, 2016), this line of argumentation is potentially misleading, as participants' inability to perform significantly above chance in these awareness tests might simply be the result of the low statistical power of these experiments. Even firm advocates of the hypothesis that contextual cuing is driven by unconscious processes recognize that performance in awareness tests often tends to be significantly above chance, if a sufficient number of participants is tested (Colagiuri & Livesey, 2016; Sisk et al., 2019). In addition, this approach ignores individual differences across participants: Even if at the group level participants perform no better than chance in the awareness test, this does not mean that all participants were completely unaware; and, vice versa, if the sample, taken as a whole is significantly above chance, this does not mean that all participants had explicit knowledge.

There is a second argument for claiming that contextual cuing is unconscious that is based precisely on individual differences in learning and awareness. Even when participants show a significant level of awareness at the group level, the size of contextual cuing, as measured by visual search advantage for repeated patterns, is usually uncorrelated with the degree of awareness shown by each participant (e.g., Conci & von Mühlenen, 2011; Dixon, Zelazo, & De Rosa, 2010; Geringswald, Baumgartner, & Pollmann, 2012; Geyer, Shi, & Müller, 2010; Manginelli, Langer, Klose, & Pollmann,

2013; Olson, Jiang, & Moore, 2005; Peterson & Kramer, 2001; Preston & Gabrieli, 2008; Schankin, & Schubö, 2009; Zang, Geyer, Assumpcao, Müller, & Shi, 2016). As the reader will immediately notice, this is essentially the line of argumentation used by Berkovitch and Dehaene (2019) in our previous example and, as such, rests on the assumption that measurement error in standard measures of contextual cuing and awareness is small or negligible. Although numerous contextual cuing experiments have tested the correlation between visual search and different measures of awareness, none of them has reported the reliabilities of these measures, which renders these analyses uninformative as to the question of whether contextual cuing is driven by unconscious processes.

There are several reasons to suspect that, in general, these reliabilities must be lower than dictated by psychometric standards for correlational research. First, as mentioned in the opening paragraph, the measures provided by some of the most popular experimental tasks are notoriously unreliable. Given that the dependent variables collected in many of these tasks (e.g., difference of response times in two conditions) are relatively similar to the measures collected in contextual cuing experiments, it is reasonable to be skeptical about the reliability of the latter. Secondly, the few implicit learning studies that have reported reliabilities have usually revealed disappointingly low values (e.g., Kalra, Gabrieli, & Finn, 2019; Kaufman et al., 2010; Siegelman & Frost, 2015; Smyth & Shanks, 2008; Vadillo, Linssen, Orgaz, Parsons, & Shanks, 2020; West, Vadillo, Shanks, & Hulme, 2018).

The present article explores the impact of measurement error on the results of studies that rely on correlational approaches to defend the unconscious character of learning. In the following sections, we estimate the reliabilities of the different measures of learning and awareness employed in this literature based on data collected in our own laboratory and on a large data set collected by other researchers. With this information at hand, we then move

on to reassess the results of previous contextual cuing studies that relied on correlational tests of awareness. Finally, we show that, to a considerable extent, the poor correlations between contextual cuing and awareness observed in the empirical data are consistent with a model of contextual cuing that accounts for participants' performance under the assumption that visual search and awareness (i.e., recognition) are driven by a common latent mental representation.

### **Reliability of Visual Search and Awareness Measures in Contextual Cuing**

To assess the reliability of the dependent measures used in contextual cuing studies, we conducted four high-powered experiments with large sample sizes covering three different types of awareness/recognition tests that have been used extensively in this literature. Although we performed and selected studies with different procedures, materials, and designs, the experimental method across all studies is prototypical for the wider contextual cuing literature. The only noticeable divergences with respect to the typical procedures were that (a) our awareness tests comprised more trials than usual and (b) Experiments 1 and 2 did not include control trials (i.e., new patterns) during the first half of the experiment. The reader can find a detailed description of the methods employed in each experiment in Appendix A. The general results (i.e., those traditionally reported for these experiments), available in Appendix A, show that all four experiments produced strong contextual cuing effects. Performance in the awareness tests was also significantly above chance in all cases. The main question that we address here is whether contextual cuing and awareness measures were correlated with each other across participants and, in case of observing a non-significant correlation, to what extent measurement error provides a reasonable explanation.



The main difference across experiments was the type of test used to measure awareness at the end of each experiment. In Experiment 1, participants' recognition of the repeated patterns was tested with a two-alternative forced-choice (2AFC) task (e.g., Chaumon, Schwartz, & Tallon-Baundry, 2008; Colagiuri, Livesey, & Harris, 2011). On each trial, participants were shown two patterns consecutively. One of them was a repeated pattern from the learning stage and the other one a new pattern that had never been shown before. Participants were asked to indicate which of them was a familiar pattern from the first stage of the experiment. Their responses were collected using a 6-point rating scale with values 1 (*definitely the first pattern*), 2 (*probably the first pattern*), 3 (*guess the first pattern*), 4 (*guess the second pattern*), 5 (*probably the second pattern*), and 6 (*definitely the second pattern*). This procedure (inspired by Colagiuri & Livesey, 2016) departs slightly from many contextual cuing experiments where participants are asked to provide a binary response (e.g., *the first pattern*) instead of using a confidence scale. Because of this, we analyzed these responses using two different procedures. On the one hand, we computed a binary awareness score by simply treating any response on the correct side of the scale (regardless of confidence) as correct and any other response as incorrect. Any score above 50% correct responses would be counted as above-chance performance on this measure. On the other hand, we computed a rating score that retained the information conveyed by confidence responses. For this measure, we recoded each participant's responses onto a 1-6 scale where 1 referred to the incorrect response with maximal confidence and 6 referred to the correct response with maximal confidence. Scores higher than 3.5 indicate above chance performance on this dependent variable.

The awareness test for Experiments 2 and 3 was an old/new recognition task, which is by far the most common type of awareness test among published contextual cuing studies (e.g., Chun & Jiang, 1998; Manns & Squire, 2001; Rosero et al., 2019). On each

trial of the testing stage, participants were presented with just one pattern, which could be either a repeated pattern from the learning stage or a completely new display, and they were asked to say whether they thought that this pattern had previously been presented during the first part of the experiment. As in Experiment 1, responses were collected with a 6-point rating scale with values 1 (*definitely it was not*), 2 (*probably it was not*), 3 (*guess it was not*), 4 (*guess it was*), 5 (*probably it was*), and 6 (*definitely it was*). Again, we used these responses to compute two different measures of awareness. For the binary score, we considered all correct detections of repeated patterns as ‘hits’ (regardless of confidence) and all incorrect responses to new patterns as ‘false alarms’ (again regardless of confidence). The proportion of hits and false alarms for each participant was converted to a  $d'$  score from Signal Detection Theory (Macmillan & Creelman, 2005). We also computed a rating score by simply subtracting the mean confidence response given to new patterns from the mean confidence response given to repeated patterns. For both variables, a value of zero indicates chance-level performance.

Finally, in Experiment 4, participants’ awareness was assessed with a generation task (Chun & Jiang, 2003; Smyth & Shanks, 2008). Participants were told that on each trial they would see a pattern that had been presented in the first stage of the experiment, except that the target was now replaced by an additional distractor of the same color. Their task was to click on any distractor on the screen where they thought the target was hidden. For consistency with Experiments 1-3, immediately after clicking on a distractor, they were also asked to report their confidence in their choice with a 3-point scale with values 1 (*I am just guessing*), 2 (*I think I am probably right*), and 3 (*I am sure I am right*). Only repeated patterns were presented during the generation task. In this case, the binary measure of awareness was the proportion of trials on which participants clicked in the correct quadrant of the response screen, where chance-level performance is indicated by a proportion of

25% correct responses. The rating score assigned each trial a score between 1 (highest confidence in an incorrect quadrant) to 6 (highest confidence in the correct quadrant). Note that, unlike the other measures considered in Experiments 1-4, there is no obvious chance-level threshold for this variable, although higher values correspond to higher levels of awareness. A tentative threshold for chance-level performance may be estimated by randomly permuting each participants' quadrant choices and ratings across trials. In our sample, the mean rating score observed across 1,000 such permutations was 3.14. This is the reference value that we will use in all subsequent analyses with this dependent variable.

In all four experiments, the magnitude of contextual cuing for each participant was computed by subtracting mean search times in repeated patterns from mean search times in new patterns. In Experiments 1 and 2, only repeated patterns were presented during the first half of the learning stage. Consequently, contextual cuing scores were computed using only data from the second half of the learning stage. In Experiments 3 and 4, participants were exposed to both types of patterns from the beginning of the visual search task. However, data from the first block were ignored in the analyses, because no evidence of learning can be expected before participant have seen repeated patterns several times. Appendix A provides detailed information about data selection criteria.

As can be seen in Figure 1, the correlation between contextual cuing and awareness was numerically positive in Experiments 1, 2, and 4, and negative in Experiment 3, regardless of whether awareness was measured with binary or rating scores. All correlations were numerically weak, ranging from  $-.115$  to  $.117$ , and indeed none of them reached statistical significance in two-tailed tests. Given the numerous failures to detect a significant correlation between learning and awareness in the research reviewed above, these null results are perhaps unsurprising. They do, however, come from studies (smallest  $N = 104$ , after data exclusions) with higher power than most previous studies, which have a

median sample size of  $N = 16$  (Vadillo et al., 2016), and the confidence intervals on the correlations (reported below) are accordingly much narrower.

The most interesting question that we want to address here is whether those null correlations could be due to the low reliability of the standard measures of awareness and learning in this task. In Table 1, we show the average split-half reliabilities of these dependent variables. To reduce the impact of chance in our assessment of reliability, we repeated the analyses for each dependent variable and experiment over 1,000 random equal-sized splits, with the only constraint that both halves of each split should contain the same number of trials with each repeated configuration<sup>1</sup>. Table 1 shows the mean uncorrected split-half correlation across iterations, together with the meta-analytic average correlations across experiments. Note that these split-half correlations underestimate the true reliability of the dependent variables, because they are based on just one half of the trials. The reliability of the measure obtained with the full sequence of trials is naturally larger and can be estimated with the Spearman-Brown correction, which is also shown in Table 1.

As can be seen, the reliability of contextual cuing scores seems to be relatively stable across experiments, despite the minor procedural differences among them, with an average reliability of .454. The reliabilities of awareness measures, in contrast, change dramatically from one experiment to another, ranging from .339 to .642 for binary scores, and from .352 to .753 for rating scores. The meta-analytic averages suggest that reliabilities tend to be slightly higher for rating scores overall. But, even for those, only Experiment 4 reaches

---

<sup>1</sup> To illustrate, in Experiment 1 the split-half reliability is based on RTs across 12 blocks of trials, each including 8 repeated and 8 new displays, for a total of 196 trials per participant. These were split randomly into 2 sets of 98 trials with the constraint that the 12 presentations of each particular display were equally divided (i.e., 6 RTs) between the two halves of each split. This was done to avoid artificial suppression of the reliability estimate as a result of chance allocation of displays to the halves. For instance, if a particular display was associated with unusually fast (or slow) RTs, then a random split in which say 10 of these RTs were included in one half and only 2 in the other would lead to a reduction in the resulting correlation that would not be indicative of true unreliability. A similar constraint was applied in computing the reliability of the awareness measures. We thank Evan Livesey for suggesting this procedure.

acceptable levels of reliability. In general, less than half of the variance seen in contextual cuing and awareness scores reflect true differences across participants.

The low reliabilities found in the experiments conducted in our laboratories might not be representative of other studies using similar procedures. To test the generalizability of our results, we reanalyzed the data from three experiments conducted by Colagiuri and Livesey (2016). For the sake of clarity, we will refer to these three data sets as Experiments CL1, CL2, and CL3 henceforth. To the best of our knowledge, these three studies represent the most ambitious attempt to quantify the correlation between learning and awareness in contextual cuing published so far, with sample sizes of 63, 84 and 766. Table 1 shows the reliabilities for the awareness measures and contextual cuing scores using the same procedure employed in the analysis of our own experiments. In general, the reliability of contextual cuing is quite consistent with the estimates from our own experiments, while the reliabilities of awareness measures are systematically lower, especially for Experiments CL1 and CL2. These results give us some confidence in the conclusion that the dependent variables used in this type of study are generally unreliable, beyond the data obtained in our own experiments.

Before continuing, it might be convenient to pause for a moment and consider the practical implications of the overly low reliabilities reported in Table 1. Consider, for instance, the seemingly low correlation between contextual cuing and the binary measure of awareness in Experiment 1. The reliabilities of these two measures impose an upper limit on the maximum observable correlation that can be expected to occur between them. Imagine that the constructs measured by variables  $x$  and  $y$  are correlated with strength  $r_{\text{lat}}$  at the latent level and that these two measures have reliabilities  $r_{\text{xx}}$  and  $r_{\text{yy}}$ . Unless the reliabilities of  $x$  and  $y$  are perfect, any non-zero observed correlation between the variables,  $r_{\text{obs}}$ , will be lower than  $r_{\text{lat}}$ , because measurement error will attenuate the observed

correlation. On average,  $r_{obs}$  will be equal to  $r_{lat} \cdot \sqrt{r_{xx} \cdot r_{yy}}$  (Schmidt & Hunter, 2014; Spearman, 1904). Given an observed correlation and the reliabilities of the dependent variables, it is possible to use the reverse logic to estimate the underlying correlation at the latent level with the expression  $r_{obs}/\sqrt{r_{xx} \cdot r_{yy}}$ . For the particular case of Experiment 1, this means that, although the observed correlation between contextual cuing and awareness is only .061, the disattenuated correlation at the latent level is  $.061/\sqrt{.483 \cdot .339} = .151$ , a small but far from trivial correlation. To put this correlation coefficient in context, this effect size is slightly larger than the relieving effect of ibuprofen on self-reported pain,  $r = .14$ , or the effect of antihistamine use on self-reported severity of runny nose and sneezing symptoms,  $r = .11$  (Meyer et al., 2001, Table 1).

Of course, the boundaries of a 95% confidence interval (CI) should also be corrected for attenuation, such that even if the point estimate of the correlation changes from .061 to .151, the corresponding  $p$ -value of the correlation coefficient remains unchanged (Schmidt & Hunter, 2014). In the particular case of Experiment 1, the boundaries of the 95% CI of the observed correlation were [-.133, .251]. If Spearman's disattenuation formula is applied to these boundaries for approximate purposes, the 95% CI becomes [-.329, .620]. In other words, Experiment 1 actually provides very little information about the latent correlation between contextual cuing and awareness, as only moderately negative and strongly positive values are excluded by the 95% CI. This conclusion is somewhat striking, bearing in mind that with a sample size of 104 participants, the statistical power of Experiment 1 is considerably larger than the usual contextual cuing experiment. Furthermore, as we will discuss in later sections, the reliability of the awareness test used in Experiment 1 is likely to be higher than in most experiments, given that it comprises an unusually large number of trials. In sum, what the information reported in Table 1 suggests

is that it is extremely difficult to estimate the true correlation between cuing and awareness with any acceptable level of accuracy.

### **A Psychometric Meta-analysis**

As explained above, if one knows the reliabilities of two dependent measures, it is possible to correct the observed correlations between them with Spearman's (1904) disattenuation formula, but doing so sometimes results in extremely wide confidence intervals. Hence, individual experiments, even those with extremely large sample sizes, usually provide very little information about the plausible value of the correlation at the latent level once reliability is taken into account. As an alternative approach, it might still be possible to use meta-analytic methods to collate data from several experiments to obtain a more precise estimate. Interestingly, there is whole approach to evidence synthesis, known as psychometric meta-analysis, built on the idea that the goal of a meta-analysis is not simply to average out the effect sizes that have been empirically observed in a set of studies, but instead to estimate what those effect sizes would have looked like in an ideal experiment unconstrained by limitations of sample size, measurement error, and other artifacts (Schmidt, 2010; Schmidt & Hunter, 2014; Wiernik & Dahlke, 2020). Unlike traditional meta-analytic methods (e.g., Borenstein, Hedges, Higgins, & Rothstein, 2009), in psychometric meta-analysis, correlation coefficients are collated after correcting them for measurement error. Furthermore, the weight of each study in the meta-analytic average depends not only on sample size, but also on measurement error.

Following up on this idea, we conducted two meta-analyses collating data from our four experiments and the three experiments conducted by Colagiuri and Livesey (2016), one of them using standard meta-analytic methods and the second using psychometric meta-analysis. Collectively, these seven experiments comprise data from 1,383

participants. The standard meta-analysis was conducted with the ‘metafor’ package for R (Viechtbauer, 2010), first converting all the observed  $r$  coefficients to Fisher’s  $z$  and then converting the resulting meta-analytic average back to an  $r$  coefficient. The psychometric meta-analysis was conducted with the ‘psychmeta’ package for R (Wiernik & Dahlke, 2020). The standard meta-analysis yielded a mean correlation of .039, [-.014, .092], suggesting that the average observed correlation must be close to zero. In contrast, the psychometric meta-analysis returned a much wider CI for the latent correlation, [-.016, .225], and a substantially larger mean correlation of  $r = .104$ . This disattenuated correlation is still non-significant at  $p < .05$  and, consequently, we cannot reject the hypothesis that the correlation is as small as zero. But with these data at hand we also cannot reject the hypothesis that it is as large as .225 either.

### **Learning and Awareness in Other Studies**

A critical reader might argue that the seven experiments analyzed in the previous sections represent just a small and perhaps unrepresentative subset of the numerous studies that have explored the correlation between learning and awareness in contextual cuing. However, there are good reasons to believe that the rest of the studies are, if anything, more heavily affected by these methodological problems. The precision of the estimates obtained in these studies depends on both sample size and reliability. Regarding the former, as we have noted above, contextual cuing experiments are usually conducted with a median sample size of just 16 participants (Vadillo et al., 2016). This figure pales in comparison with the sample size of any of the seven experiments included in the previous meta-analysis.

Perhaps more importantly, the reliability of the dependent measures used in most studies must be considerably lower than those observed in this subset of experiments we



have analyzed. One of the main determinants of reliability is the number of data points collected in a test. In the present set of experiments, the awareness test comprised 32 trials for the 2AFC test in Experiments 1, CL1, and CL3; 64 for the recognition test in Experiments 2, 3, and CL2; and 32 for the generation task in Experiment 4. In contrast, among the studies reviewed by Vadillo et al. (2016), the median numbers of trials were 12, 24, and 12, for the same three types of test, respectively. In Figure 2, we extrapolate what the expected reliabilities would be for experiments with such numbers of trials using the Spearman-Brown prediction formula. As we have three different estimates for 2AFC and recognition tasks, the predicted reliabilities are based on their average (weighted by sample size). The expected reliabilities for the awareness tests using the typical number of trials are .120, .161, and .402 for 2AFC, recognition, and generation tests, respectively. Of course, these predictions are subject to sampling error and assume that the number of trials would be the only relevant difference between Experiments 1-4 and CL1-CL3 and previous studies. Needless to say, they may underestimate, perhaps grossly, the true reliabilities of the measures collected in other experiments. But, in any case, they suggest that reliabilities are likely to be problematic in most or all of the contextual cuing experiments conducted with this general procedure.

To better appreciate the scale of the problem, Table 2 provides a list of all the published experiments that have analyzed the correlation between contextual cuing and awareness relying on an awareness test for which we can estimate reliability with the data from Experiments 1-4 and CL1-CL3. Detailed information about the literature search strategy is available in Appendix B. Note that the 18 experiments included in Table 2 are not the only ones that have relied on a correlational analysis to conclude that contextual cuing was unconscious. Table B1 lists additional studies that we could not enter into the following analyses because they failed to meet the inclusion criteria.

All the studies listed in Table 2 assessed awareness with old/new recognition tests, although the scoring algorithm changed from one study to another. While most studies measured awareness with a  $d'$  score, others correlated the size of contextual cuing with accuracy or with the proportion of hits in the recognition test. We computed reliability estimates for all three types of awareness measures using the data from Experiments 2, 3, and CL2 and then extrapolated the reliability that would correspond to awareness tests with different numbers of trials using the Spearman-Brown prediction formula, as done in Figure 2. As expected, the estimated reliabilities shown in Table 2 are too low to allow any valid inference from these measures in correlational analyses. The only exception to this conclusion is perhaps the set of experiments that use hit rate as the measure of awareness, which reach reliabilities slightly above .45. Note, however, that this gain in reliability comes at a cost in terms of interpretability: Without additional information, a high hit rate could reflect either good explicit recognition or simply a bias towards identifying all patterns (including new ones) as “old” (Macmillan & Creelman, 2005). In summary, given the likely (un)reliabilities of the dependent measures and the relatively small sample sizes used in these studies, their results offer very little information about the latent correlation between learning and awareness.

### **Precision of Participant-Level Estimates**

In the previous sections, we have shown that standard measures of learning and awareness used in contextual cuing experiments are alarmingly unreliable and that this constitutes an obstacle for any attempt to ascertain whether this type of learning is unconscious or not on the basis of correlational analyses. However, so far, we have not addressed why these dependent variables are unreliable or whether we could expect them to be any different. The fact that measures of learning and awareness are unreliable entails

that most of the variation seen across participants is not due to genuine differences between individuals, but simply to random error and chance. Under these circumstances, the specific score obtained by participants provides very little information about their true level of performance (Stallings & Gillmore, 1971).

To illustrate this point, Figure 3 depicts the means (dots) and 95% CIs (vertical lines) of the contextual cuing score and measured awareness for each individual participant in Experiments 1-4. For the sake of simplicity, only the binary awareness scores are included in the figure. The grey line and band in each panel denote the mean and 95% CI of performance across participants. Note that the bands are quite narrow and systematically above chance, consistent with the conclusion that our procedures elicited a powerful contextual cuing effect and a noticeable level of awareness. At the individual level, in contrast, CIs are consistent with a wide range of values and provide only vague information about the true level of performance. For most individual participants, performance was not significantly above chance (lower end of the CI above the dashed line) and, in general, the CIs of the vast majority of participants overlap extensively with each other. Except for the most extreme cases, it is difficult to know whether the scores obtained by any two participants are genuinely different from each other or, alternatively, whether the differences between them are entirely due to chance.

This low level of precision is not an accidental feature of contextual cuing and awareness scores. To the contrary, it is a logical consequence of the distributional properties of visual search times and awareness scores, which any plausible model of visual search and recognition will predict. As an illustrative example, Figure 4A shows the two ex-Gaussian distributions that provide the best fit to the distribution of search times for the first participant tested in Experiment 3. Previous research shows that the ex-Gaussian distribution provides an excellent model of response times (RTs) in visual search tasks like

the ones used in contextual cuing (Palmer, Horowitz, Torralba, & Wolfe, 2011). Although contextual cuing is a powerful effect at the group level, it is somewhat surprising to see that the distributions of RTs predicted by the model for the two experimental conditions (repeated and new displays) are highly overlapping, with just a small advantage for repeated search displays. It is possible to simulate multiple experiments for this particular participant by sampling a certain number of trials from the two model distributions and then obtaining a contextual cuing score by subtracting the mean RT extracted from the repeated distribution from the mean RT from the new distribution. Figure 4B shows the results of simulating 10,000 such experiments, each one with the same number of trials that were used to fit the model to this particular participant (in this case, 169 repeated trials and 173 new trials, after excluding the first epoch, trials with errors and outliers). As can be seen, a clear contextual cuing effect emerges, with an average visual search advantage of about 100 ms for repeated displays. But it is also clear that the result is subject to alarming levels of sampling error. In a significant proportion of cases, the contextual cuing effect is even reversed (i.e., a negative RT difference). It is obvious from these simulations that, if multiple experiments were conducted with this particular participant, the average observed effect would change dramatically from one replication to another.

The same point can be made for performance in the awareness test. A Signal Detection Theory (SDT) analysis of data from this particular participant (assuming equal variances for “repeated” and “new” distributions) yielded a discriminability score of  $d' = 0.24$  and a response criterion of  $c = 0.27$ . Figure 4C recreates the SDT model that would correspond to these parameter values. Again, using the model, it is possible to simulate what this participant would do in multiple independent experiments. Each hit rate would be given by the area under the distribution for repeated patterns that is higher than  $c$ , and each false alarm rate would be given by the area under the distribution for new patterns also

above  $c$ . Figure 4D shows the results of 10,000 such simulated experiments, each one with 32 repeated trials and 32 new trials, as in the actual experiment. As in the case of contextual cuing scores, the results of the simulation show that the observed  $d'$  scores can change dramatically from one simulation to another. Although the mean of the  $d'$  distribution is clearly positive and virtually identical to the observed  $d'$  of this specific participant, many simulations give rise to negative values well below zero.

Given the results of both simulations, it is perhaps unsurprising that the respective CIs in Figure 3 are so wide. The poor precision in the estimation of participant-level parameters is exactly what we would expect given the distributional properties of participants' responses in these kinds of task.

### **So, What Are the Largest Correlations we Can Expect to Find between Contextual Cuing and Awareness? A Modeling Approach**

The preceding discussion demonstrates that an ex-Gaussian model of visual search times predicts low precision of individual-level estimates of contextual cuing, at least for standard tasks comprising around 100-200 trials in each condition. Similarly, SDT predicts imprecise estimates of  $d'$  with awareness tests based on just 64 trials. It is possible to follow up this modeling approach to grasp what are the largest correlations we can expect to find between cueing and awareness—even when assuming a perfect relationship between the two constructs at the latent level. Let us imagine that the ex-Gaussian model of reaction times and the SDT model of awareness outlined in the previous sections are combined into a single model that accounts for cuing and recognition by assuming that a single latent variable drives performance in both tasks (for a similar approach in the context of implicit memory, see Berry, Shanks, Speekenbrink, & Henson, 2012). Intuitively, one would expect such a model to predict a strong correlation between

contextual cuing and recognition. Contrary to this, the following simulations show that the predicted correlations can be astonishingly low, as long as the model retains the amount of trial-by-trial noise observed in data. For the sake of simplicity, we only attempted to model the results of the two experiments that used a conventional old/new recognition test (Experiments 2 and 3).

In the following model, we assume that the performance of participant  $i$  in a recognition test is driven by two latent variables,  $\delta_i$  and  $\gamma_i$ , that might not be exactly equal to the corresponding  $d'$  and  $c$  scores observed empirically due to sampling error. The probability that the participant will recognize a pattern as repeated in the recognition test,  $p_{ki}$ , is given by equation

$$p_{ki} \sim \Phi(\beta_i | x_k \cdot \delta_i, 1) \quad (\text{Equation 1})$$

where  $\Phi$  is the cumulative normal distribution function with mean  $x_k \cdot \delta_i$  and standard deviation 1 at point  $\beta_i$ ; indicator variable  $x_k = \{0, 1\}$  classifies the data as belonging to condition “new” or “repeated”; data point  $\beta_i$  is equal to  $\delta_i/2 + \gamma_i$ ; that is, both  $\beta_i$  and  $\gamma_i$  refer to the response criterion in SDT, but  $\gamma_i$  is defined with respect to the mean between the signal and no-signal distributions, centered at  $\delta_i$  and 0, respectively, while  $\beta_i$  is defined in the same units as  $\delta_i$ .

Similarly, visual search times for participant  $i$  in condition  $k$ ,  $rt_{ki}$ , are given by

$$rt_{ki} \sim \text{ExGauss}(a_i - x_k \cdot b \cdot \delta_i, \sigma_i, \tau_i) \quad (\text{Equation 2})$$

where  $\text{ExGauss}$  is the probability density function of the ex-Gaussian distribution;  $a_i$  is the mean of the normal component in condition new; as in Equation 1,  $x_k = \{0, 1\}$  codes for conditions new and repeated, respectively;  $b \cdot \delta_i$  is the magnitude of the contextual cuing effect, which depends on  $\delta_i$  (from the SDT model) and a fixed scaling factor,  $b$ , which is constant for all participants;  $\sigma_i$  and  $\tau_i$  are the standard deviations of the normal and exponential components, respectively.

In total, the model includes five free parameters:  $\delta_i$ ,  $\gamma_i$ ,  $a_i$ ,  $\sigma_i$ , and  $\tau_i$ . The most important part of the model is the expression  $b \cdot \delta_i$ , which ensures that the visual search advantage for repeated patterns is always proportional to the parameter that drives performance in the awareness test,  $\delta_i$ . That is to say, the size of contextual cuing is given by the recognition parameter  $\delta_i$  scaled by a fixed factor  $b$ . This scaling factor was the same for all participants within an experiment and was selected in such a way that it maximized the fit to the observed relationship between contextual cuing and awareness.

In Experiment 2, for instance, the mean level of awareness is a  $d'$  score of 0.298 and the mean magnitude of contextual cuing is 310.88 ms. These two estimates might be subject to sampling error, but they are not biased by measurement error. Therefore, any regression line relating awareness and cuing at the latent level must cross the coordinates (0.298, 310.88). In addition, if the model assumes that contextual cuing and awareness are driven by a single latent variable, then at the latent level there should be no contextual cuing for an ideal participant with zero awareness, that is, the regression line should also cross the coordinates (0, 0). If the linear regression relating awareness and cuing at the latent level must cross these coordinates, then the slope of the regression must be equal to  $310.88 / 0.298 = 1,041.53$  ms/ $d'$  unit (for a similar approach, see the ‘ratio-of-means’ heuristic in Dienes, 2019). Consequently, this was the fixed value of  $b$  in Experiment 2. Following the same logic,  $b$  was set to  $87.33 / 0.175 = 499.23$  ms/ $d'$  unit in Experiment 3.

We fitted this model individually to the data from each participant in Experiments 2 and 3 using maximum likelihood estimation. The procedure is described in Appendix C. Table 3 shows the mean best fitting parameters across participants as well as the correlation between predicted and observed performance across participants. As can be seen, model predictions are strongly correlated with the respective observed visual search

times in both conditions and with the response criterion,  $c$ , in the awareness test. The correlation between predicted and observed  $d'$  scores is weaker, but still strong.

Figure 5A shows the results of a single simulation of the model with the best-fitting parameters from each participant in Experiment 2. Perhaps the most striking feature of the results is the discrepancy between the red line, denoting the slope of the relationship between awareness and cuing at the latent level, and the black line, denoting the observed slope for this particular simulation. In the same vein, although cuing is entirely dependent on awareness at the latent level (i.e., given a true score of zero in the awareness test the predicted true cuing score is also zero), the correlation between the variables is quite weak at the empirical level. For the particular simulation depicted in Figure 5A, the correlation between cuing and awareness is .388. Of course, this value is subject to sampling error. To obtain a clearer view of the average results predicted by the model, Figure 5B shows the distribution of correlation coefficients between cuing and awareness across 1,000 simulations of Experiment 2. On average, the predicted correlation is .296. Figures 5C and 5D show the same information for Experiment 3. The predicted correlation between cuing and awareness is substantially lower in this case. To test the generalizability of these results, we also fitted the model to Experiment CL2 from Colagiuri and Livesey (2016) and simulated data from the model with the best-fitting parameters for each participant. As shown in Figures 5E and 5F, the average correlation predicted by the model is even lower for this experiment.

It is also worth noting that, although the intercept of the model is zero, the simulations produce positive intercepts at the observed level. That is, for a participant with an observed awareness score of zero, the predicted observed cuing effect is systematically higher than zero. Many studies in different areas of research on unconscious mental processes have relied on this *above-zero regression intercept* (Greenwald et al., 1995)



pattern of results to claim that a cognitive process is unconscious, on the grounds that if the predicted effect for a participant with no awareness is higher than zero, then the effect must be influenced by factors that are unrelated to awareness (e.g., Berkovitch & Dehaene, 2019; Greenwald & De Houwer, 2017; Reuss, Kiesel, & Kunde, 2015; Salvador et al., 2018). Our simulations show that, however intuitive, this logic is deeply flawed: A positive intercept at the observed level does not necessarily indicate that the intercept at the latent level is also positive. A behavioral effect might be significant among participants who show no positive evidence of awareness and yet the effect might be entirely driven by awareness (see also Klauer, Draine, & Greenwald, 1998; Miller, 2000; Shanks, 2017).

To be clear, we are not putting forward the model outlined above as an adequate model of the contextual cuing task. Even researchers who disagree with the conclusion that contextual cuing is unconscious would probably feel uncomfortable with the assumption that contextual cueing and awareness are perfectly correlated at the latent level. Note also that we are making no attempt to compare the performance of this model against an alternative model assuming zero correlation at the latent model. Our model ignores crucial aspects of the contextual cuing task, including for instance the fact that reaction times change over the course of the task and that this change is steeper for repeated patterns than for new patterns and that contextual cuing might not only affect the mean of the response time distribution but also its variance (i.e.,  $\sigma$  and  $\tau$  in the model).

Our goal is not to claim that our implementation is an appropriate model of contextual cuing, but to show that even a model that assumes a perfect correlation at the latent level will predict rather low empirical correlations between learning and awareness, provided it preserves the amount of trial-by-trial noise typically found in these dependent variables. To better appreciate the implications of this, let us take the largest average correlation in Figure 5,  $r = .296$ , as a liberal estimate of the empirical correlation that one

can expect to find between learning and awareness. Let us also assume that we try to detect this correlation with the usual sample size of the studies reviewed in Table 2, that is, a median sample size of 16 participants. The power of such a study to detect the correlation between learning and awareness would be just 20%. It is clear that if we take as a reference any of the other predicted correlations in Figure 5 or if we relax the assumption that learning and awareness are perfectly related at the latent level, the chances that these experiments will return a significant result becomes negligible, even when they are strongly correlated at the latent level.

### **General Discussion**

In this article, we have marshaled a range of evidence challenging the view that contextual cuing is driven by unconscious mental processes. Our particular focus has been the claim, advanced in many previous studies, that low and non-significant correlations between performance and awareness must be diagnostic of an unconscious process driving contextual cuing. When due heed is paid to the low reliabilities of the cuing and awareness measures, we find that performance and awareness would be observationally uncorrelated, or only slightly correlated, even when they were perfectly associated at the latent level. Stated differently, without better data and methods, the correlational approach is inadequate to discriminate between alternative accounts of the relationship between conscious and unconscious mental processes.

We suspect that the problems of reliability highlighted in the present article apply to many experimental paradigms in unconscious cognition research. In general, the minimal information that we have about the reliabilities of the measures collected in other paradigms tends to confirm these concerns (e.g., Kalra et al., 2019; Kaufman et al., 2010; Siegelman & Frost, 2015; Vadillo et al., 2020; West et al., 2018). Across the previous

sections, we have used contextual cuing as a convenient example because the lack of correlation between visual search and awareness has become a popular means to support the claim that this form of learning is unconscious. Furthermore, as far as we know, contextual cuing is the only implicit learning paradigm where the correlation between learning and performance has been tested in at least one large-scale study involving several hundred participants, conclusively eliminating lack of statistical power as a candidate explanation for near-zero, non-significant correlations.

In fact, the research described in the present article started out as an attempt to follow up on the impressive set of studies reported by Colagiuri and Livesey (2016), extensively discussed in previous sections. Their three experiments, in turn, were originally motivated by a previous article published by some of us (Vadillo et al., 2016), where we described a simple model of contextual cuing that did indeed predict that measures of learning and awareness should be weakly but positively correlated. In retrospect, we suspect that this particular model, which ignored the potential role of reliability, may have encouraged researchers (including ourselves) to adopt an overoptimistic view of the extent to which contextual cuing and awareness measures can be expected to correlate.

Based on the reliability information summarized in Table 1, we can estimate that the largest correlation between contextual cuing scores and awareness measures that one can expect to find in typical contextual cuing tasks will rarely be larger than .30 or .40, even when the correlation between them is perfect at the latent level. As can be seen in Figure 5, our model, which also assumes a perfect correlation at the latent level, makes even more pessimistic predictions, with expected empirical correlations in the range of .15 to .30. If the (rather implausible) idea of a perfect latent correlation is abandoned, it becomes clear that the correlation coefficients can easily become tiny at the observed level, to such an

extent that even samples comprising hundreds of participants may fail to detect a true correlation at the latent level.

Contextual cuing studies have relied on recognition being a test of conscious awareness in order to argue that the contextual cuing effect is unconscious. We have treated recognition as an awareness measure, in order to respond to the claims in this literature. However, strictly speaking, in showing that a single-process model can account for low correlations between contextual cuing and recognition, we have only established that a single source of knowledge can underlie both cuing and recognition. This knowledge may be conscious, unconscious, or both. Indeed, according to two major theories of consciousness, higher order theory and global workspace theory (Dienes & Seth, 2018), both contextual cuing and recognition can be based on unconscious knowledge, because both are tested with respect to the nature of the world (cf. Berry & Dienes, 1993). However, what we have shown is that a non-significant correlation between the two measures is not an argument for any one of them being based on unconscious knowledge. Moreover Experiment 4, which employed a generation test, obtained results very similar to the Experiments 1-3 employing recognition tests. In the face of question marks over whether recognition tests are acceptable measures of awareness, advocates of unconscious influences in contextual cuing have defended generation tests as an alternative (Chun & Jiang, 2003).

#### *Can we improve the reliability of awareness tests?*

It is often easier to diagnose a problem than to fix it. Unfortunately, there are no easy solutions to the challenge posed by the low reliability of the dependent measures collected in contextual cuing experiments in particular and studies on unconscious mental processes in general. A typical recommendation to improve the reliability of psychological measures

is to increase the number of observations taken from each participant. As can be seen in Figure 2, it follows from the Spearman-Brown prediction formula that the reliability of any dependent measure will approach 1.0 with a sufficiently large number of observations. This prediction, however, hinges critically on the assumption that all the observations tap onto the same latent construct. Unfortunately, this is not always the case. Enkavi et al. (2019) found that several laboratory-based measures of self-regulation never reached acceptable levels of test-retest reliability—regardless of task length. Our own results provide converging evidence that increasing the number of trials might not always improve reliability. For instance, the reliability estimates of contextual cuing scores were similar in Experiments 1-4, even though they were estimated on the basis of data from Epochs 6-12 in Experiments 1 and 2 (because no new patterns were included during the first half of the experiment) and Epochs 2-12 in Experiments 3 and 4 (where all epochs included new patterns and only Epoch 1 was excluded from the analysis, as no evidence of learning was expected during the first trials).

Furthermore, in the case of contextual cuing, there are good reasons to expect that increasing the length of the task beyond the usual number of trials might change the nature of the cognitive processes involved. It is possible that performance in a contextual cuing task is initially guided by implicit or unconscious processes and that, as participants gain more experience, they eventually become aware of the repetitions (e.g., Goujon, Didierjean, & Poulet, 2014; for a similar claim in a different paradigm, see Bechara, Damasio, Tranel, & Damasio, 1997). Similarly, if the awareness test includes too many blocks, then the new patterns used as lures to test recognition might eventually become so familiar that participants are simply unable to tell them from repeated patterns.

Another reason why contextual cuing scores and measures of awareness tend to be unreliable is that they are usually computed on the basis of difference scores. The

magnitude of contextual cuing is defined as the difference between search times for new and repeated patterns. In the same vein,  $d'$  scores are computed as the difference between the hit rate and false-alarm rate (although converted using the inverse cumulative standard normal distribution). Difference scores are often unreliable because they remove many sources of inter-individual differences (e.g., cognitive ability or response speed), and at the same time include twice as much random variance from the two individual components being subtracted (Enkavi et al., 2019; Hedge, Powell, & Sumner, 2018). In the case of contextual cuing, although the relative search advantage for repeated patterns is unreliable, mean search times for repeated patterns are highly reliable. For instance, in Experiments 1-4, the (Spearman-Brown corrected) split-half reliabilities of this dependent measure are .947, .911, .981, and .980, respectively. Given this high reliability, it is tempting to use it as a measure of contextual cuing, ignoring the baseline search times for new patterns. And, in fact, search times for repeated patterns do inversely correlate (if only weakly) with measures of awareness. For Experiments 1-4, these correlations are  $r = -.200, p = .042$ ;  $r = -.075, p = .437$ ;  $r = -.071, p = .440$ ; and  $r = -.131, p = .130$ , respectively; with a meta-analytic average of  $r = -.118, p = .011$ . However, what these analyses gain in terms of reliability is also lost in terms of interpretability: Search times for repeated patterns are probably influenced by many factors that are of no interest to researchers in this domain (such as individual differences in perceptual-motor speed). Unlike the difference between conditions, raw search times have little value as a measure of learning and, consequently, whether or not they correlate with measures of awareness is largely irrelevant.

### *Limitations*

Of course, the present study is not without limitations. Throughout this article, we have recurrently used methods for the estimation of reliability and for the disattenuation of

correlations that rely on classical test theory. All psychometric methods designed to deal with measurement error rest on assumptions that might not always hold and classical test theory is no exception. Specifically, classical test theory assumes that participants' true scores and errors are uncorrelated, that errors in different tasks are uncorrelated and that, on average, errors are zero (Hambleton & Jones, 1993). Psychometric meta-analysis not only inherits these assumptions, but comes with problems of its own. For instance, it is more sensitive to publication bias than traditional meta-analytic methods.

In addition, and most importantly, the unit of analysis in classical test theory is usually the score obtained by each participant in each task, and in the particular case of contextual cuing, this approach might be misleading. The evidence collected so far suggests that participants may only learn about a small set of the patterns that are presented repeatedly over the experiment (Colagiuri & Livesey, 2016; Smyth & Shanks, 2008). If so, averaging data across all repeated patterns to obtain a single contextual cuing score for each participant might remove the most important source of variation in these datasets. When participants' scores are computed across all trials, they might look too homogeneous to detect any meaningful differences among them; but below this homogeneity, there might be substantial variation in learning at the pattern-level that could, perhaps, correlate with awareness scores. Following up on this idea, Colagiuri and Livesey (2016) analyzed the correlation between contextual cuing and awareness at the pattern level and found no evidence of a positive correlation. In fact, they even detected a small negative correlation.

It is important to note, however, that the problem of measurement error applies to pattern-level analyses as well. If participants only learn about a few patterns, one would expect to find more true variation in learning and awareness scores at the pattern level. But, in contrast, each repeated pattern is only presented a small number of times throughout the task, which means that these measures will also suffer from more sampling

error. Our data suggest that these two factors (i.e., more true variation and more sampling error) cancel each other out and, in the end, pattern-level scores show similar levels of reliability as participant-level scores. Specifically, our reliability estimates of the contextual cuing scores at the pattern level were .372, .367, .527, and .487 for Experiments 1-4, respectively, and our reliability estimates of the binary awareness scores were .216, .372, .372, and .501.<sup>2</sup>

### *Concluding comments*

Future research should not only try to improve the reliabilities of the measures typically collected in unconscious-learning experiments: Ideally, it should also adopt methodological approaches specifically designed to ameliorate and quantify the impact of measurement error. For instance, Rouder and Haaf (2019) implemented a hierarchical model that can estimate the latent correlation between performance in two difference tasks after accounting for trial-by-trial variation. Unlike the methods adopted in the present study, Rouder and Haaf's approach yields *portable* estimates of latent correlations and test-retest reliabilities, that is, estimates of the ideal values that would be obtained in a hypothetical scenario with an infinite number of trials in each task. This method also allows researchers to compute Bayes factors quantifying the weight of the evidence in favor of the hypothesis that the latent correlation is zero against the alternative hypothesis that it is different from zero. Given that this area of research relies extensively on the interpretation of null correlations, which are often uninterpretable in frequentist statistics,

---

<sup>2</sup> For each participant, we calculated the split-half reliability of contextual cuing scores for each of the 8 repeated displays across 1,000 random equal-sized splits and then averaged these reliabilities across iterations. The same constraint described in Footnote 1 about equal numbers of occurrences of each pattern in each half of a split was applied in this analysis. On each iteration, the magnitude of contextual cuing for each pattern was computed by subtracting the RTs for the repeated pattern from the RTs to a new pattern that contained the target in the exact same location and color. We computed the split-half correlations using a linear-mixed model, with random intercepts for participants, to account for the dependency between different data points from the same participant.



we think that adopting this Bayesian stance is a must for future research. More recently, Malejka, Vadillo, Dienes, and Shanks (in press) have developed additional models for the analysis of unconscious cognitive processes that also take into account the reliabilities of dependent variables and the imprecision of participant-level estimates. As in the case of Rouder and Haaf, these models can also be used to compute Bayes factors in favor of the null hypothesis of no correlation.

An alternative strategy to circumvent the problems posed by low reliabilities is to replace correlational analyses by a purely experimental approach. Instead of obtaining learning and awareness measures and then testing for the correlation between them, it seems more effective to manipulate the level of awareness and test for its impact on learning, or vice versa. To our knowledge, only a couple of studies (Chun & Jiang, 2003, Experiment 2; Kroell, Schlagbauer, Zinchenko, Müller, & Geyer, 2019) have adopted this approach, with contradictory findings so far. Unlike most contextual cuing experiments, Chun and Jiang (2003, Experiment 2) provided participants with explicit instructions alerting them that some patterns would be presented many times during the experiment, and that it would be in their advantage to remember them to boost their visual search performance. This instruction made no significant difference to the magnitude of contextual cuing, which was, if anything, numerically smaller in the explicit condition. The conclusion extracted by the authors was that contextual cuing must be unconscious, because instructions aimed to promote explicit learning made no difference in the results. However, the findings of the study must be interpreted with caution. Despite the instructions, participants were not significantly above chance in the awareness test, suggesting that the manipulation was probably ineffective.

In contrast, in a recent study by Kroell et al. (2019), participants were asked to memorize a set of “explicit” displays at the beginning of the experiment, which were then

intermixed with “implicit” repeated patterns during a standard contextual cuing task. Contextual cuing was substantially larger for the “explicit” displays, suggesting that conscious learning did contribute to this task. Although it is still too early to draw any firm conclusion from these studies, we suspect that this experimental strategy is more likely to yield insights about the conscious or unconscious nature of contextual cuing than the comparatively large body of research addressing this question with purely correlational methods.

In any case, we hope to have shown that one of the most common criteria to demonstrate that a particular mental process is unconscious is deeply flawed: The absence of correlation between performance in a task and measures of awareness provides no convincing evidence for the claim that the cognitive processes underlying performance are unconscious, unless both constructs are quantified with sufficiently reliable measures. Future research should improve the reliability of these measures and adopt methods that are better suited to deal with measurement error.

### **Open Practices Statement**

All data, materials, and analysis scripts related to the present study are publicly available at <https://osf.io/fuzvn/>. All data exclusions, manipulations, and measurements conducted in Experiments 1-4 are reported in Appendix A.

## References

- Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (1997). Deciding advantageously before knowing the advantageous strategy. *Science*, **275**, 1293-1295.
- Berkovitch, L., & Dehaene, S. (2019). Subliminal syntactic priming. *Cognitive Psychology*, **109**, 26-46.
- Berry, C. J., Shanks, D. R., Speekenbrink, M., & Henson, R. N. A. (2012). Models of recognition, repetition priming, and fluency: Exploring a new framework. *Psychological Review*, **119**, 40-79.
- Berry, D. C., & Dienes, Z. (1993). *Implicit learning: Theoretical and empirical issues*. Hove: Lawrence Erlbaum.
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: John Wiley & Sons.
- Brysbart, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition*, **2**:16.
- Chaumon, M., Schwartz, D., & Tallon-Baudry, C. (2008). Unconscious learning versus visual perception: Dissociable roles for gamma oscillations revealed in MEG. *Journal of Cognitive Neuroscience*, **21**, 2287-2299.
- Chun, M. M., & Jiang, Y. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology*, **36**, 28-71.
- Chun, M. M., & Jiang, Y. (2003). Implicit, long-term spatial contextual memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **29**, 224-234.
- Colagiuri, B., & Livesey, E. J. (2016). Contextual cuing as a form of nonconscious learning: Theoretical and empirical analysis in large and very large samples. *Psychonomic Bulletin & Review*, **23**, 1996-2009.

- Colagiuri, B., Livesey, E. J., & Harris, J. A. (2011). Can expectancies produce placebo effects for implicit learning? *Psychonomic Bulletin & Review*, **18**, 399-405.
- Conci, M., & von Mühlelen, A. (2011). Limitations of perceptual segmentation on contextual cueing in visual search. *Visual Cognition*, **19**, 203-233.
- Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, **12**, 671-584.
- De Schryver, M., Hughes, S., Rosseel, Y., & De Houwer, J. (2016). Unreliable yet still replicable: A comment on LeBel and Paunonen (2011). *Frontiers in Psychology*, **6**:2039.
- Dienes, Z. (2015). How Bayesian statistics are needed to determine whether mental states are unconscious. In M. Overgaard (Ed.), *Behavioural methods in consciousness research* (pp. 199-220). Oxford, UK: Oxford University Press.
- Dienes, Z. (2019). How do I know what my theory predicts? *Advances in Methods and Practices in Psychological Science*, **2**, 364-377.
- Dienes, Z., & Seth, A. K. (2018). Conscious versus unconscious processes. In G. C. L. Davey (Ed.), *Psychology* (BPS Textbooks in Psychology; pp. 262-323). Wiley: Chichester.
- Dixon, M. L., Zelazo, P. D., & De Rosa, E. (2010). Evidence for intact memory-guided attention in school-aged children. *Developmental Science*, **13**, 161-169.
- Enkavi, A. Z., Eisenberg, I. W., Bissett, P. G., Mazza, G. L., MacKinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2019). Large-scale analysis of test-retest reliabilities of self-regulation measures. *Proceedings of the National Academy of Sciences*, **116**, 5472-5477.

- Geringswald, F., Baumgartner, F., & Pollmann, S. (2012). Simulated loss of foveal vision eliminates visual search advantage in repeated displays. *Frontiers in Human Neuroscience*, **6**:134.
- Geringswald, F., Herbig, A., Hoffmann, M. B., & Pollmann, S. (2013). Contextual cueing impairment in patients with age-related macular degeneration. *Journal of Vision*, **13**:28.
- Geringswald, F., & Pollmann, S. (2015). Central and peripheral vision loss differentially affects contextual cueing in visual search. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **41**, 1485-1496.
- Geyer, T., Mueller, H. J., Assumpcao, L., & Gais, S. (2013). Sleep-effects on implicit and explicit memory in repeated visual search. *PLoS ONE*, **8**, e69953.
- Geyer, T., Shi, Z., & Müller, H. J. (2010). Contextual cueing in multiconjunction visual search is dependent on color- and configuration-based intertrial contingencies. *Journal of Experimental Psychology: Human Perception and Performance*, **36**, 515-532.
- Goujon, A., Didierjean, A., & Poulet, S. (2014). The emergence of explicit knowledge from implicit learning. *Memory & Cognition*, **42**, 225-236.
- Goujon, A., Didierjean, A., & Thorpe, S. (2015). Investigating implicit statistical learning mechanisms through contextual cueing. *Trends in Cognitive Sciences*, **19**, 524-533.
- Greenwald, A. G., & De Houwer, J. (2017). Unconscious conditioning: Demonstration of existence and difference from conscious conditioning. *Journal of Experimental Psychology: General*, **146**, 1705-1721.
- Haaf, J. M., & Rouder, J. N. (2019). Some do and some don't? Accounting for variability of individual difference structures. *Psychonomic Bulletin & Review*, **26**, 772-789.

- Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, **50**, 1166-1186.
- Hedger, N., Garner, M., & Adams, W. J. (2019). Do emotional faces capture attention, and does this depend on awareness? Evidence from the visual probe paradigm. *Journal of Experimental Psychology: Human Perception and Performance*, **45**, 790-802.
- Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, **12**, 38-47.
- Jiang, Y. V., Sha, L. Z., & Sisk, C. A. (2018). Experience-guided attention: Uniform and implicit. *Attention, Perception, & Psychophysics*, **80**, 1647-1653.
- Kalra, P. B., Gabrieli, J. D. E., & Finn, A. S. (2019). Evidence of stable individual differences in implicit learning. *Cognition*, **190**, 199-211.
- Kaufman, S. B., Deyoung, C. G., Gray, J. R., Jiménez, L., Brown, J., & Mackintosh, N. (2010). Implicit learning as an ability. *Cognition*, **116**, 321-340.
- Klauer, K. C., Draine, S. C., & Greenwald, A. G. (1998). An unbiased errors-in-variables approach to detecting unconscious cognition. *British Journal of Mathematical and Statistical Psychology*, **51**, 253-267.
- Kroell, L. M., Schlagbauer, B., Zinchenko, A., Müller, H. J., & Geyer, T. (2019). Behavioural evidence for a single memory system in contextual cuing. *Visual Cognition*, **27**, 551-562.
- Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide*. Mahwah, NJ: Erlbaum.
- Makovski, T. (2018). Meaning in learning: Contextual cueing relies on objects' visual features and not on objects' meaning. *Memory & Cognition*, **46**, 58-67.

- Malejka, S., Vadillo, M. A., Dienes, Z., & Shanks, D. R. (in press). Correlation analysis to investigate unconscious mental processes: A critical appraisal and mini-tutorial. *Cognition*.
- Manginelli, A. A., Langer, N., Klose, D., & Pollmann, S. (2013). Contextual cueing under working memory load: Selective interference of visuospatial load with expression of learning. *Attention, Perception, & Psychophysics*, **75**, 1103-1117.
- Manns, J. R., & Squire, L. R. (2001). Perceptual learning, awareness, and the hippocampus. *Hippocampus*, **11**, 776-782.
- Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., Eisman, E. J., Kubiszyn, T. W., & Reed, G. M. (2001). Psychological testing and psychological assessment: A review of the evidence. *American Psychologist*, **56**, 128-165.
- Miller, J. (2000). Measurement error in subliminal perception experiments: Simulation analyses of two regression methods. *Journal of Experimental Psychology: Human Perception and Performance*, **26**, 1461-1477.
- Olson, I. R., Jiang, Y., & Moore, K. S. (2005). Associative learning improves visual working memory performance. *Journal of Experimental Psychology: Human Perception and Performance*, **31**, 889-900.
- Palmer, E. M., Horowitz, T. S., Torralba, A., & Wolfe, J. M. (2011). What are the shapes of response time distributions in visual search? *Journal of Experimental Psychology: Human Perception and Performance*, **37**, 58-71.
- Parsons, S., Kruijt, A.-W., & Fox, E. (2019). Psychological science needs a standard practice of reporting the reliability of cognitive-behavioral measures. *Advances in Methods and Practices in Psychological Science*, **2**, 378-395.
- Peterson, M. S., & Kramer, A. F. (2001). Attentional guidance of the eyes by contextual information and abrupt onsets. *Perception & Psychophysics*, **63**, 1239-1249.

- Preston, A. R., & Gabrieli, J. D. E. (2008). Dissociation between explicit memory and configural memory in the human medial temporal lobe. *Cerebral Cortex*, **18**, 2192-2207.
- Reuss, H., Kiesel, A., & Kunde, W. (2015). Adjustments of response speed and accuracy to unconscious cues. *Cognition*, **134**, 57-62.
- Rey-Mermet, A., Gade, M., & Oberauer, K. (2018). Should we stop thinking about inhibition? Searching for individual and age differences in inhibition ability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **44**, 501-526.
- Rosenbaum, G. M., & Jiang, Y. V. (2013). Interaction between scene-based and array-based contextual cueing. *Attention, Perception, & Psychophysics*, **75**, 888-899.
- Rosero, M. A., Winkelmann, T., Pohlack, S., Cavalli, J., Nees, F., & Flor, H. (2019). Memory-guided attention: Bilateral hippocampal volume positively predicts implicit contextual learning. *Brain Structure and Function*, **224**, 1999-2008.
- Rouder, J. N., & Haaf, J. M. (2019). A psychometrics of individual differences in experimental tasks. *Psychonomic Bulletin & Review*, **26**, 452-467.
- Rouder, J., Kumar, A., & Haaf, J. M. (2019). *Why most studies of individual differences with inhibition tasks are bound to fail*. Unpublished manuscript. Available at <https://doi.org/10.31234/osf.io/3cjr5>
- Salvador, A., Berkovitch, L., Vinchier, F., Cohen, L., Naccache, L., Dehaene, S., & Gaillard, R. (2018). Unconscious memory suppression. *Cognition*, **180**, 191-199.
- Schankin, A., & Schubö, A. (2009). Cognitive processes facilitated by contextual cueing: Evidence from event-related brain potentials. *Psychophysiology*, **46**, 668-679.
- Schankin, A., Stursberg, O. & Schubö, A. (2008). The role of implicit context information in guiding visual-spatial attention. In B. Caputo & M. Vincze (Eds.), *Cognitive vision* (pp. 93-106) Berlin: Springer.



- Schmidt, F. L. (2010). Detecting and correcting the lies that data tell. *Perspectives on Psychological Science*, **5**, 233-242.
- Schmidt, F. L., & Hunter, J. E. (2014). *Methods of meta-analysis: Correcting error and bias in research findings (3rd Edition)*. Thousand Oaks, CA: Sage.
- Shanks, D. R. (2017). Regressive research: The pitfalls of post hoc data selection in the study of unconscious mental processes. *Psychonomic Bulletin & Review*, **24**, 752-775.
- Siegelman, N., & Frost, R. (2015). Statistical learning as an individual ability: Theoretical perspectives and empirical evidence. *Journal of Memory and Language*, **81**, 105-120.
- Sisk, C. A., Remington, R. W., & Jiang, Y. V. (2019). Mechanisms of contextual cuing: A tutorial review. *Attention, Perception, & Psychophysics*, **81**, 2571-2589.
- Smyth, A. C., & Shanks, D. R. (2008). Awareness in contextual cuing with extended and concurrent explicit tests. *Memory & Cognition*, **36**, 403-415.
- Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, **15**, 72-101.
- Stallings, W. M., & Gillmore, G. M. (1971). A note on "accuracy" and "precision". *Journal of Educational Measurement*, **8**, 127-129.
- Vadillo, M. A., Konstantinidis, E., & Shanks, D. R. (2016). Underpowered samples, false negatives, and unconscious learning. *Psychonomic Bulletin & Review*, **23**, 87-102.
- Vadillo, M. A., Linssen, D., Orgaz, C., Parsons, S., & Shanks, D. R. (2020). Unconscious or underpowered? Probabilistic cuing of visual attention. *Journal of Experimental Psychology: General*, **149**, 160-181.
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, **36**(3), 1-48.

- West, G., Vadillo, M. A., Shanks, D. R., & Hulme, C. (2018). The procedural learning deficit hypothesis of language learning disorders: We see some problems. *Developmental Science*, **21**:e12552.
- Wiernik, B. M., & Dahlke, J. A. (2020). Obtaining unbiased results in meta-analysis: The importance of correcting for statistical artifacts. *Advances in Methods and Practices in Psychological Science*, **3**, 94-123.
- Zang, X., Geyer, T., Assumpcao, L., Müller, H. J., & Shi, Z. (2016). From foreground to background: How task-neutral context influences contextual cueing of visual search. *Frontiers in Psychology*, **7**:852.
- Zang, X., Jia, L., Müller, H. J., & Shi, Z. (2015). Invariant spatial context is learned but not retrieved in gaze-contingent tunnel-view search. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **41**, 807-819.
- Zellin, M., Conci, M., von Mühlennen, A., & Müller, H. J. (2011). Two (or three) is one too many: Testing the flexibility of contextual cueing with multiple target locations. *Attention, Perception, & Psychophysics*, **73**, 2065-2076.
- Zellin, M., Conci, M., von Mühlennen, A., & Müller, H. J. (2013). Here today, gone tomorrow -Adaptation to change in memory-guided visual search. *PLoS ONE*, **8**:e59466.
- Zimmerman, D. W., & Zumbo, B. D. (2015). Resolving the issue of how reliability is related to statistical power: Adhering to mathematical definitions. *Journal of Modern Applied Statistical Methods*, **14**, 9-26.

### Figure Captions

*Figure 1.* Correlation between the size of contextual cuing and two different measures of awareness (binary and rating scores) across participants in Experiments 1-4. The binary measures of awareness are the proportion of correct responses in the 2AFC test in Experiment 1,  $d'$  in Experiments 2-3, and the proportion of correct responses in the quadrant generation task in Experiment 4. Rating scores took into account participants' confidence on each response of the awareness test (see the main text). In all the experiments, contextual cuing was defined as the difference between reaction times for new patterns and reaction times for repeated patterns. Dotted lines denote chance-level performance.

*Figure 2.* Predicted reliabilities of three binary measures of awareness (2AFC, recognition, and generation accuracy). Predictions are extrapolated from the empirical reliabilities observed in Experiments 1-4 and CL1-CL3 (circles) using the Spearman-Brown prediction formula. The black diamond denotes the predicted reliability for an experiment with the typical number of test trials.

*Figure 3.* Precision of participant-level estimates of contextual cuing scores and binary measures of awareness. Participants are ordered from smallest to largest by mean score (dots) and color bars depict 95% CIs. The grey line and band in each panel denote the mean and 95% CI of performance across participants.

*Figure 4.* Panel A: Best-fitting ex-Gaussian distributions to reaction times in the repeated and new conditions for one participant in Experiment 3. Panel B: Distribution of

predicted contextual cuing scores for the same participant based on the model depicted in Panel A. Panel C: Best-fitting SDT model to recognition data from the same participant. Panel D: Distribution of predicted  $d'$  scores based on the model depicted in Panel C, with sampling as described in the main text.

*Figure 5.* Panels A, C, E: Results of simulations of the single process model with the best-fitting parameters for each participant in Experiments 2-3 and CL2, respectively. Each datapoint represents one simulated participant. The red line denotes the underlying relationship between contextual cuing and recognition performance at the latent level. The black line denotes the best-fitting regression at the observed level. Panels B, D, F: Distribution of correlation coefficients observed across 1,000 simulations.

Table 1. Split-half Correlations and Reliabilities Across Experiments

|                                      | Contextual cuing       |                | Awareness test: Binary |                | Awareness test: Rating |                |
|--------------------------------------|------------------------|----------------|------------------------|----------------|------------------------|----------------|
|                                      | Split-half correlation | Spearman-Brown | Split-half correlation | Spearman-Brown | Split-half correlation | Spearman-Brown |
| Experiment 1                         | .319                   | .483           | .204                   | .339           | .309                   | .472           |
| Experiment 2                         | .209                   | .346           | .264                   | .417           | .477                   | .646           |
| Experiment 3                         | .367                   | .537           | .211                   | .348           | .214                   | .352           |
| Experiment 4                         | .272                   | .428           | .473                   | .642           | .603                   | .753           |
| Meta-analysis of Experiments 1-4     | .293                   | .454           | .296                   | .457           | .414                   | .585           |
| Experiment CL-1                      | .310                   | .474           | .067                   | .126           | .216                   | .355           |
| Experiment CL-2                      | .268                   | .423           | .113                   | .203           | .104                   | .189           |
| Experiment CL-3                      | .308                   | .470           | .155                   | .268           | .197                   | .329           |
| Meta-analysis of Experiments CL1-CL3 | .304                   | .467           | .146                   | .254           | .190                   | .320           |

*Note:* Meta-analytic averages of split-half correlations were computed converting all correlation coefficients to Fisher's  $z$  scores and submitting them to a Random Effects meta-analysis using the metafor package for R. The resulting averages were then back-converted to  $r$  coefficients. All the

Spearman-Brown corrected estimates were computed by applying the Spearman-Brown prediction formula to the split-half correlations reported on their left.

Table 2. Studies Quantifying the Correlation Between Contextual Cuing and Awareness

| Study           | Sample size | Type of test           | No. trials | Imputed reliability | Observed correlation [and 95% CI] |
|-----------------|-------------|------------------------|------------|---------------------|-----------------------------------|
| PG.08           | 23          | Recognition (hit rate) | 12         | .466                | .100 [-.326, .492]                |
| SSS.08          | 16          | Recognition ( $d'$ )   | 32         | .203                | .290 [-.240, .687]                |
| SS.09           | 14          | Recognition ( $d'$ )   | 32         | .203                | .095 [-.459, .596]                |
| DZD.10          | 12          | Recognition (accuracy) | 8          | .066                | -.160 [-.672, .456]               |
| GSM.10.2        | 10          | Recognition ( $d'$ )   | 32         | .203                | .280 [-.424, .773]                |
| ZCVM.11.3       | 22          | Recognition ( $d'$ )   | 24         | .161                | -.030 [-.446, .397]               |
| GBP.12          | 25          | Recognition (accuracy) | 24         | .173                | -.203 [-.554, .209]               |
| GHHP.13.amd.bin | 16          | Recognition (accuracy) | 24         | .173                | -.141 [-.595, .381]               |
| GHHP.13.con     | 16          | Recognition (accuracy) | 24         | .173                | -.368 [-.730, .156]               |
| MLKP.13.1       | 36          | Recognition (hit rate) | 12         | .466                | -.248 [-.533, .088]               |
| MLKP.13.3       | 35          | Recognition (hit rate) | 12         | .466                | .134 [-.209, .447]                |
| MLKP.13.4       | 40          | Recognition (hit rate) | 12         | .466                | .113 [-.206, .410]                |
| ZCVM.13.1a      | 12          | Recognition ( $d'$ )   | 24         | .161                | -.170 [-.678, .448]               |
| ZCVM.13.4       | 14          | Recognition ( $d'$ )   | 24         | .161                | .420 [-.142, .777]                |

|            |    |                        |    |      |                     |
|------------|----|------------------------|----|------|---------------------|
| ZJMS.15.1  | 13 | Recognition ( $d'$ )   | 24 | .161 | -.130 [-.635, .453] |
| ZGAMS.16.1 | 10 | Recognition ( $d'$ )   | 48 | .275 | .050 [-.598, .659]  |
| ZGAMS.16.2 | 10 | Recognition ( $d'$ )   | 48 | .275 | .040 [-.605, .653]  |
| RWPCNF.19  | 40 | Recognition (hit rate) | 12 | .466 | .040 [-.275, .347]  |

*Note:* List of published studies reporting a correlation between contextual cuing and awareness across participants. The acronyms in the Study column stand for the initials of the authors' family names followed by the year of publication (enabling them to be uniquely identified in the References), and experiment number and/or condition. Sample size, type of test, number of trials, and observed correlations were taken directly from the main text of the articles. The imputed reliability was calculated from the reliability estimates of our Experiments 1-4 and Experiments CL1-CL3, using the Spearman-Brown prediction formula, as explained in the main text. Other studies that estimated the correlation between cuing and awareness, but failed to meet our selection criteria, are listed in Table B1.



Table 3. Best-Fitting Parameters and Model Performance

| <b>A. Best fitting parameters</b>                                |                |                 |                      |                    |                      |
|--|----------------|-----------------|----------------------|--------------------|----------------------|
|  | $\delta$       | $\gamma$        | $a$                  | $\sigma$           | $\tau$               |
| <b>Experiment 2</b>  | 0.17<br>(0.20) | -0.01<br>(0.39) | 1,392.90<br>(513.63) | 452.87<br>(370.67) | 1,804.79<br>(484.27) |
| <b>Experiment 3</b>  | 0.10<br>(0.11) | 0.04<br>(0.29)  | 765.35<br>(171.20)   | 146.73<br>(96.11)  | 618.12<br>(168.96)   |
| <b>B. Correlation between observed and predicted performance</b> |                |                 |                      |                    |                      |
|  | RT repeated    | RT new          | $d'$                 | $c$                |                      |
| <b>Experiment 2</b>  | .923           | .925            | .713                 | .996               |                      |
| <b>Experiment 3</b>  | .990           | .990            | .660                 | .999               |                      |

*Note:* Table 3A summarizes the mean (and standard deviation) of best fitting parameters of the model for Experiments 2 and 3 across participants. Parameters  $\delta$  and  $\gamma$  denote the latent  $d'$  and  $c$  parameters from SDT in the awareness test, with the constraint that the size of contextual cueing in the visual search task is also dependent on  $\delta$  (see the main text). Reaction times are sampled from an ex-Gaussian distribution where  $a$  is the mean of the normal component for the new patterns in milliseconds, and  $\sigma$  and  $\tau$  are the standard deviations of the normal and exponential components, respectively, also in milliseconds. Table 3B reports the correlation

between observed and predicted measures of performance in the visual search task (mean RT in conditions repeated and new) and in the recognition test ( $d'$  and  $c$  scores from SDT-analysis).

## Appendix A

### Method and Results of Experiments 1-4

The four experiments analyzed in the main article were conceived as conceptual replications of the three experiments reported by Colagiuri and Livesey (2016). However, there were minor differences among them regarding (a) the specific type of awareness test, (b) the experimental design and (c) the shape of the stimuli used as distractors and targets.

Experiment 1, inspired by Experiments CL1 and CL3, was conducted in April 2016 as part of a student laboratory class at UCL. To deal with the time constraints imposed by the class, we reduced its length by presenting only repeated patterns in the first half of the experiment. This reduced the number of contextual cuing trials by 25%, but it also meant that the magnitude of contextual cuing could only be measured using trials from the second half of the experiment. Experiment 2, combining samples run at both UCL and UAM during June and December 2018, tried to replicate Experiment CL2, but again, presenting only repeated patterns during the first part of the experiment. The task was programmed with different software and slightly different stimuli than in Experiment 1, which rendered the visual search task slightly more difficult for participants. To discount the possibility that this led to any meaningful difference in the results, Experiment 3, conducted at UAM from December 2018 to March 2019, implemented a small change in the shape of the distractors aimed at facilitating visual search. It also included random trials from the beginning of the experiment, as in the original experiments conducted by Colagiuri and Livesey. Finally, Experiment 4, conducted simultaneously at UAM and UCL between March and May 2019, used the exact same materials, procedure, and design as Experiment 3 during the visual search task, but in combination with a different type of awareness test

at the end of the experiment. In the following sections, we provided detailed information about the methods employed in each experiment.

### *Method*

#### *Participants*

Our minimum planned sample size for Experiments 1-4 was 100 participants, although, in practice, we tested more participants in all cases for a range of reasons including student class size and the desire to improve the precision of our reliability estimates. A total of 112 undergraduate students at UCL took part in Experiment 1; 60 participants from the UCL participant pool and 63 from the UAM pool took part in Experiment 2; 126 participants from the UAM pool took part in Experiment 3; 64 participants from the UAM pool and 80 participants from the UCL pool took part in Experiment 4. Participants were tested either in isolated, individual cubicles or at computers in a large and quiet laboratory class sufficiently separated to avoid interference. All participants at UAM received course credit for their participation, whereas participants at UCL received course credit (Experiment 1) or could choose between £5 or course credit (Experiments 2 and 4). All participants provided informed consent before the experiment started. The studies were approved by the ethics committees of UAM (Ref. CEI-80-1473) and UCL (Ref. CPB/2010/004).

#### *Stimuli*

Stimulus presentation and response collection were controlled with Cogent 2000 and Cogent Graphics toolboxes for MATLAB, in Experiment 1, and with PsychToolbox for MATLAB in Experiments 2-4. In all cases, each search display consisted of the presentation of 15 L-shaped distractors and one T-shaped target on a grey background.

These stimuli appeared in random locations of an invisible  $12 \times 12$  grid, with the constraint that each quadrant should contain four distractors, or three distractors and the target.

Distractors and targets subtended approximately 0.8 degrees from a viewing distance of 60 cm. In Experiment 2, the vertical line of the distractors was slightly offset, which rendered them more similar to the target and, consequently, made the visual search task more difficult. Stimuli could be red, green, blue, or yellow. The distractors could be rotated either  $0^\circ$ ,  $90^\circ$ ,  $180^\circ$ , or  $270^\circ$ , but the target could only be rotated  $90^\circ$  or  $270^\circ$  as compared to the letter T. For any given pattern in the repeated condition, the orientation and colour of distractors was kept constant in subsequent presentations. The colour of the target was also kept constant in all the repeated presentations of a consistent pattern, but its orientation was randomly determined on a trial-by-trial basis. At the beginning of each experiment, 16 locations of the grid (four per quadrant), roughly equidistant from the center of the screen, were preselected to contain the targets. Distractors could never appear in these locations. Half of these locations (two per quadrant) were reserved for search displays in the repeated condition, while the other half were used for displays in the new condition.

### *Design and Procedure*

During the visual search task, participants were exposed to 24 blocks of trials. In Experiments 3-4, each block contained the same eight repeated search displays and eight new search displays. In Experiments 1-2, the first 12 blocks contained only the eight repeated search displays, while all subsequent blocks included both the eight repeated and the eight new search displays. Each visual search trial began with a fixation cross presented for 1 sec followed immediately by the search display. Participants had to respond to the orientation of the target using the keys <Z> for targets rotated  $90^\circ$  clockwise and <M> for targets rotated  $270^\circ$  clockwise. If the response was correct, the search display disappeared and a grey screen was presented during a 1-sec inter-trial interval. If the

response was incorrect the search display was replaced by the message “WRONG!” (or its Spanish equivalent) written in red capital letters. The error message remained at the centre of the screen for 2 sec before proceeding to the inter-trial interval. Participants had a 20 sec rest break after every 100 trials.

After completing the visual search task, all participants completed an awareness test. In Experiment 1, the awareness test used a two-alternative forced-choice task. The test consisted of four blocks of eight trials. On each trial, participants were presented with two search displays in sequence, one of them being a repeated display from the visual search task and the other being a completely new display using one of the target locations reserved for new patterns during the visual search task. Whether the repeated or the new pattern appeared first was determined at random. To control for any learning effects during the awareness test, the same eight new displays were presented on each of the four blocks. On each trial, the first search display was presented for 3 seconds, followed by a 750 msec blank screen and then by the second display for an additional 3 seconds. At the end of each trial, participants were asked which of the previous patterns had been presented repeatedly during the first stage of the experiment. They entered their responses on a 6-point Likert scale with response options 1 (*definitely the first pattern*), 2 (*probably the first pattern*), 3 (*guess the first pattern*), 4 (*guess the second pattern*), 5 (*probably the second pattern*), and 6 (*definitely the second pattern*).

Experiments 2 and 3 used an old/new recognition task for the awareness test. The test comprised four blocks of 16 trials each. On each trial, a search display was presented for 3 sec and immediately afterwards participants were asked to report whether they thought that pattern had been presented during the first part of the experiment. Their responses were collected with a 6-point Likert scale with values 1 (*definitely it was not*), 2 (*probably it was not*), 3 (*guess it was not*), 4 (*guess it was*), 5 (*probably it was*), and 6 (*definitely it was*).

Half of the trials in each block presented search displays from the repeated condition in the visual search task, while the other half used new patterns using the target locations reserved for new displays during the visual search task. As in Experiment 1, the new displays generated for the awareness test repeated across blocks to control for learning effect during the awareness test.

The awareness test in Experiment 4 relied on a generation task. The test comprised four blocks of eight trials. On each trial, a repeated search display from the visual search task was presented, but with the target replaced by an L-shaped distractor of the same color. Participants were told that the patterns they would see had already been presented during the visual search task, except that the target had been replaced by a distractor. Their task was to click on the distractor that replaced the target. Immediately after clicking on a distractor with the mouse, they were asked to rate their confidence in their response with a 3-point scale with values 1 (*I am just guessing*), 2 (*I think I am probably right*), and 3 (*I am sure I am right*).

### *Results*

To ensure that participants were paying attention to the task, we removed all data from any participant who failed to achieve 95% accuracy in the visual search task. A total of 8, 13, 4, and 10 participants from Experiments 1-4, respectively, failed to meet this criterion and were therefore excluded from all further analyses. In addition, for the analysis of visual search times, we ignored trials with incorrect responses, trials with RTs longer than 10 sec, and trials immediately after a rest break. We then computed the mean and standard error of the remaining RTs for each participant and we further removed any trial with an RT more than 3 standard deviations away from the mean. To reduce noise in the data, RTs from adjacent blocks were binned into two-block epochs. In Experiments 1-2,

RTs from the first half of the experiment were ignored in all subsequent analyses, because only the second half of the experiment contained both repeated and new search displays.

Table A1 shows the results of four condition (repeated vs. new)  $\times$  epoch (6-12 in Experiments 1-2 and 1-12 in Experiments 2-4) repeated-measures analyses of variance (ANOVAs) on RTs for Experiments 1-4, respectively. As can be seen, the main effect of condition was significant in all cases, confirming that our methods elicited robust contextual cuing effects. The main effect of epoch was also significant in all cases, showing a general decrease in RTs with task practice. Finally, the condition  $\times$  epoch interaction was significant in Experiments 3-4, showing the decline in RTs over the visual search task was steeper for repeated than for new search displays. However, this only holds for the two experiments where both types of displays were presented from the beginning of the task.

As explained in the main text, the measures collected in the awareness test of the experiments allowed us to compute two different indices of awareness for each participant: a “binary” score that only takes into consideration whether responses were correct or not (e.g., a  $d'$  score for Experiments 2 and 3) and an “ordinal” score that also considers the level of confidence in each response. In Table A2 we present the descriptive statistics of both types of awareness scores in Experiments 1-4, together with the results of  $t$ -tests comparing mean performance against chance (except for the ordinal measure in Experiment 4, for which chance-level performance is undefined). Performance was significantly above chance in all cases, confirming previous evidence that participants show positive signs of awareness in well-powered contextual cuing experiments (Vadillo et al., 2016).



*Table A1. Analysis of Reaction Times During the Learning Stage of Experiments 1-4*

| <b>Effect / Experiment</b>                 | <b>df</b> | <b>F</b> | <b>p</b> | <b><math>\eta_p^2</math></b> |
|--|-----------|----------|----------|------------------------------|
| <b>Condition (repeated vs. new)</b>        |           |          |          |                              |
| Experiment 1                               | 1, 103    | 146.81   | < .001   | .59                          |
| Experiment 2                               | 1, 109    | 112.52   | < .001   | .50                          |
| Experiment 3                               | 1, 121    | 120.43   | < .001   | .50                          |
| Experiment 4                               | 1, 133    | 109.14   | < .001   | .45                          |
| <b>Epoch</b>                               |           |          |          |                              |
| Experiment 1                               | 5, 515    | 23.12    | < .001   | .18                          |
| Experiment 2                               | 5, 545    | 19.57    | < .001   | .15                          |
| Experiment 3                               | 11, 1331  | 96.19    | < .001   | .44                          |
| Experiment 4                               | 11, 1463  | 103.68   | < .001   | .43                          |
| <b>Condition <math>\times</math> Epoch</b> |           |          |          |                              |
| Experiment 1                               | 5, 515    | 0.43     | .831     | < .01                        |
| Experiment 2                               | 5, 545    | 1.17     | .324     | .01                          |
| Experiment 3                               | 11, 1331  | 7.02     | < .001   | .05                          |
| Experiment 4                               | 11, 1463  | 4.30     | < .001   | .03                          |

*Table A2. Analysis of Performance in the Awareness Test*

| <b>Measure / Experiment</b> | <b>Mean</b> | <b>95% CI</b> | <b><i>t</i></b> | <b>df</b> | <b><i>p</i></b> | <b><i>d<sub>z</sub></i></b> |
|-----------------------------|-------------|---------------|-----------------|-----------|-----------------|-----------------------------|
| <b>Binary measures</b>      |             |               |                 |           |                 |                             |
| Experiment 1                | 0.55        | [0.53, 0.57]  | 5.25            | 103       | <.001           | 0.52                        |
| Experiment 2                | 0.30        | [0.22, 0.37]  | 7.94            | 109       | <.001           | 0.76                        |
| Experiment 3                | 0.17        | [0.11, 0.24]  | 5.43            | 121       | <.001           | 0.49                        |
| Experiment 4                | 0.30        | [0.28, 0.32]  | 4.65            | 133       | <.001           | 0.40                        |
| <b>Rating measures</b>      |             |               |                 |           |                 |                             |
| Experiment 1                | 3.70        | [3.64, 3.76]  | 6.32            | 103       | <.001           | 0.62                        |
| Experiment 2                | 0.33        | [0.25, 0.41]  | 8.10            | 109       | <.001           | 0.77                        |
| Experiment 3                | 0.18        | [0.12, 0.23]  | 6.07            | 121       | <.001           | 0.55                        |
| Experiment 4                | 3.24        | [3.20, 3.29]  | 4.54            | 133       | <.001           | 0.38                        |

## **Appendix B**

### **Literature Search and Inclusion/Exclusion criteria**

Our literature search strategy sought to identify all published studies testing the correlation between the size of contextual cuing and performance in an awareness test across participants. For papers published before November 2013, we considered the 73 empirical studies meta-analysed by Vadillo et al. (2016). In that review, studies were considered for inclusion only if they relied on a standard contextual cuing task where the location of distractors in a static visual search display predicted the location of the target in the same display (Criterion 1). Additionally, studies were excluded if the search display included natural scenes, because it is widely acknowledged that the contextual cuing effect elicited by those stimuli is accompanied by conscious recollection of the search displays (Criterion 2). From the articles reviewed by Vadillo et al. (2016), we selected only studies where the authors reported the correlation between contextual cuing and awareness scores across participants (Criterion 3). Furthermore, we only considered studies computing awareness scores for which we could impute a reliability estimate based on the data collected in our Experiments 1-4 (Criterion 4). Following this strategy, we were able to retrieve 14 independent correlation coefficients (see Table 2).

For papers published during or after November 2013, we obtained a list of all the studies citing the seminal article by Chun and Jiang (1998) in the Web of Science from November 1st 2013 to July 17th 2019 and we screened them following the aforementioned selection criteria. From the 439 articles returned by the Web of Science, we discarded 386 on the basis of information provided by the title or abstract. The full texts of the remaining 53 articles were read by the first author. This led to the identification of four additional correlation coefficients eligible for inclusion (see Table 3).

In total, this literature search allowed us to identify 18 eligible experiments published in 13 articles (Dixon et al., 2010; Geringswald et al., 2012, 2013; Geyer et al., 2010; Manginelli et al., 2013; Preston & Gabrieli, 2008; Rosero et al., 2019; Schankin & Schubö, 2009; Schankin, Stursberg, & Schubö, 2008; Zang et al., 2015, 2016; Zellin, Conci, von Mühlenen, & Müller, 2011, 2013). Table B1 lists additional experiments that explored the correlation between contextual cuing and awareness in one way or another but that failed to meet one of the inclusion criteria.

*Table B1. Characteristics of Excluded Studies and Reasons for Exclusion*

| Study                                      | Reason for exclusion  |
|--|---|
| Peterson & Kramer (2001)                   | According to the main text, correlation analyses were conducted in Experiments 1 and 2 but exact correlation coefficients were not reported (Criterion 3).  |
| Olson, Jiang, & Moore (2005)               | This study was excluded from Vadillo et al. (2016) because the experimental procedure was different from the standard contextual cuing task (Criterion 1).  |
| Smyth & Shanks (2008)                      | Search times and awareness scores were not correlated across participants, but across patterns within each participant (Criterion 3).   |
| Geyer, Shi, & Müller (2010, Experiment 3)  | The relationship between learning and awareness was explored by several means (e.g., correlations across patterns within each participant and <i>t</i> -tests of cuing scores based on median split of awareness), but the authors did not report a simple correlation between cuing and awareness scores across participants (Criterion 3).  |
| Conci & von Mühlennen (2011, Experiment 2) | Correlations between cuing and awareness were computed at the pattern level, not at the participant level (Criterion 3)   |
| Geyer, Mueller, Assumpcao & Gais (2013)    | The authors compared cuing for recognized and nonrecognized patterns, but they did not report correlations across participants (Criterion 3).   |
| Rosenbaum & Jiang (2013)                   | This study was excluded from Vadillo et al. (2016) because it used naturalistic scenes (Criterion 2).   |
| Geringswald & Pollmann (2015)              | Awareness was measured with a generation task with repeated and new patterns, with the difference in proximity between clicks and targets in each condition as the measure of awareness. We cannot compute a reliability estimate for this awareness scores from our own data, because the only experiment where we used a generation task (Experiment 4) did not include new patterns (Criterion 4). |
| Makovski (2018)                            | Awareness was measured with familiarity ratings, for which we lack a reliability estimate (Criterion 4).  |

## Appendix C

### Fitting the Single-Process Model to Participants' Data

Reaction times from the visual search task were filtered following the same trial selection criteria described in Appendix A. For participants in Experiments 3, we further removed data from Epoch 1 before fitting the model, as participants are not expected to show a contextual cuing effect during the initial trials. For Experiment 2, only trials from the second half of the visual search task were used to fit the model. The model was fitted individually to data from all valid visual search trials and the awareness test trials using the Neelder-Mead in R's `optim` function to find the parameter values that minimized the negative log-likelihood of the model.

Given that both experiments included more contextual cuing than awareness test trials, we scaled the log-likelihood of the awareness test to ensure that both tasks had the same weight in constraining the model optimization. To find suitable starting parameters for the optimization, we first fitted an ex-Gaussian distribution to all valid search times from each participant, ignoring their experimental condition (repeated or new). These parameters were obtained with the 'retimes' package for R. The best fitting  $\mu$ ,  $\sigma$ , and  $\tau$ , were used as starting parameters for  $a$ ,  $\sigma$ , and  $\tau$ , respectively. For the awareness test,  $d'$  and  $c$  scores were computed from the hit and false alarm rates, and the resulting estimates were then used as starting parameters for  $\delta$  and  $\gamma$ , respectively.

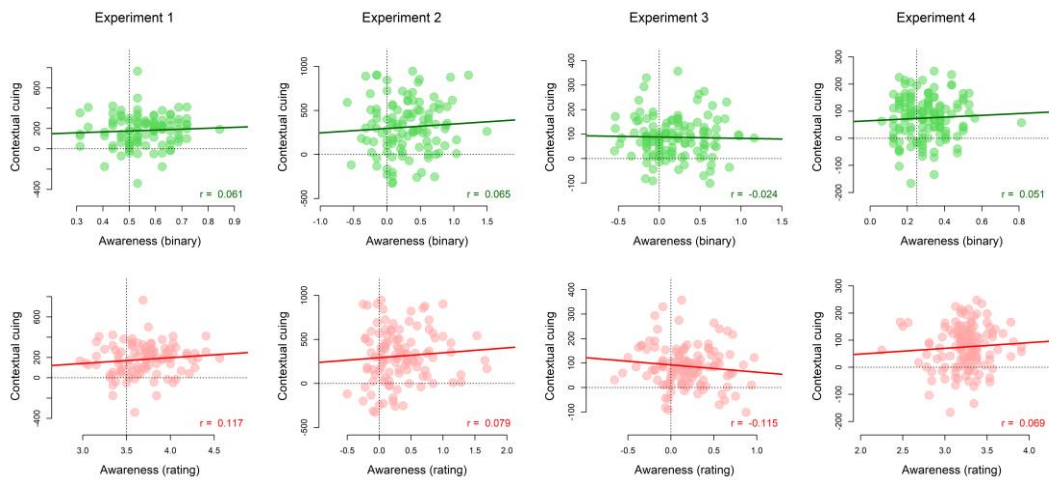


Figure #1

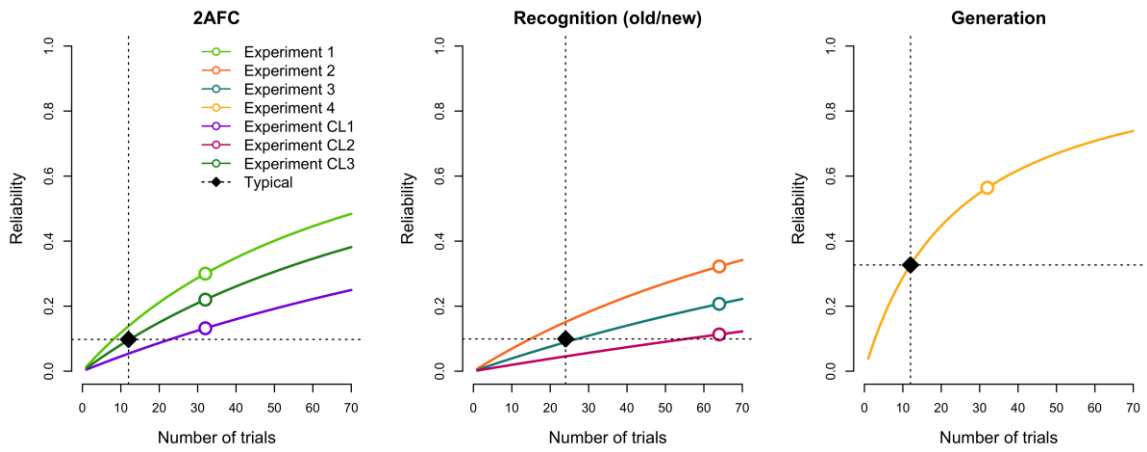


Figure #2



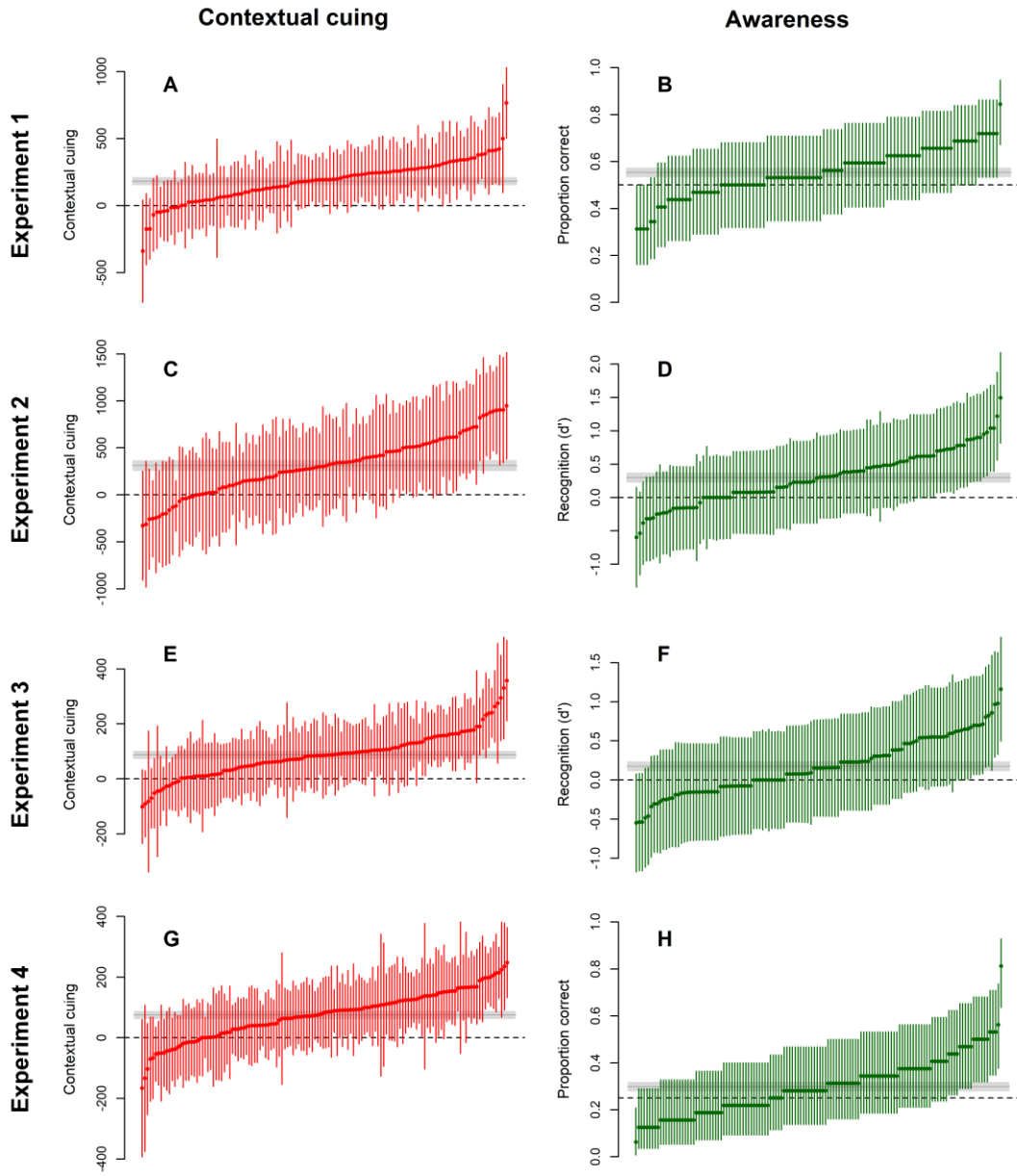


Figure #3

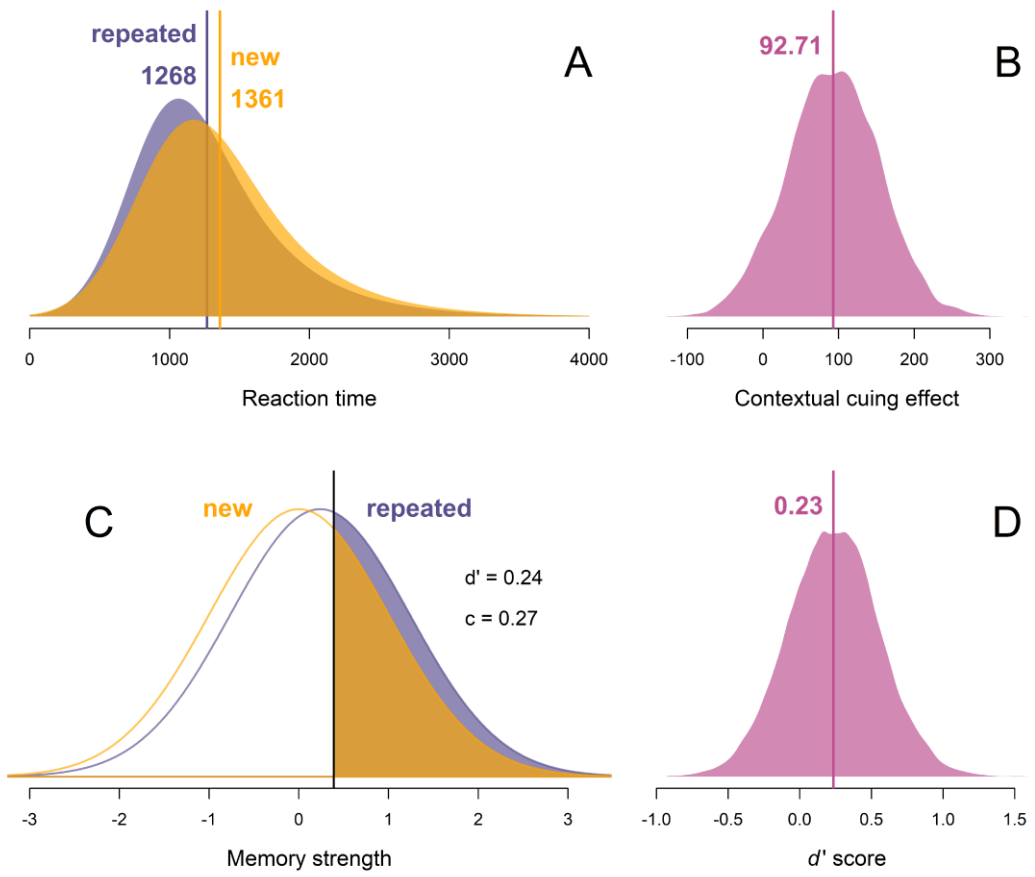


Figure #4

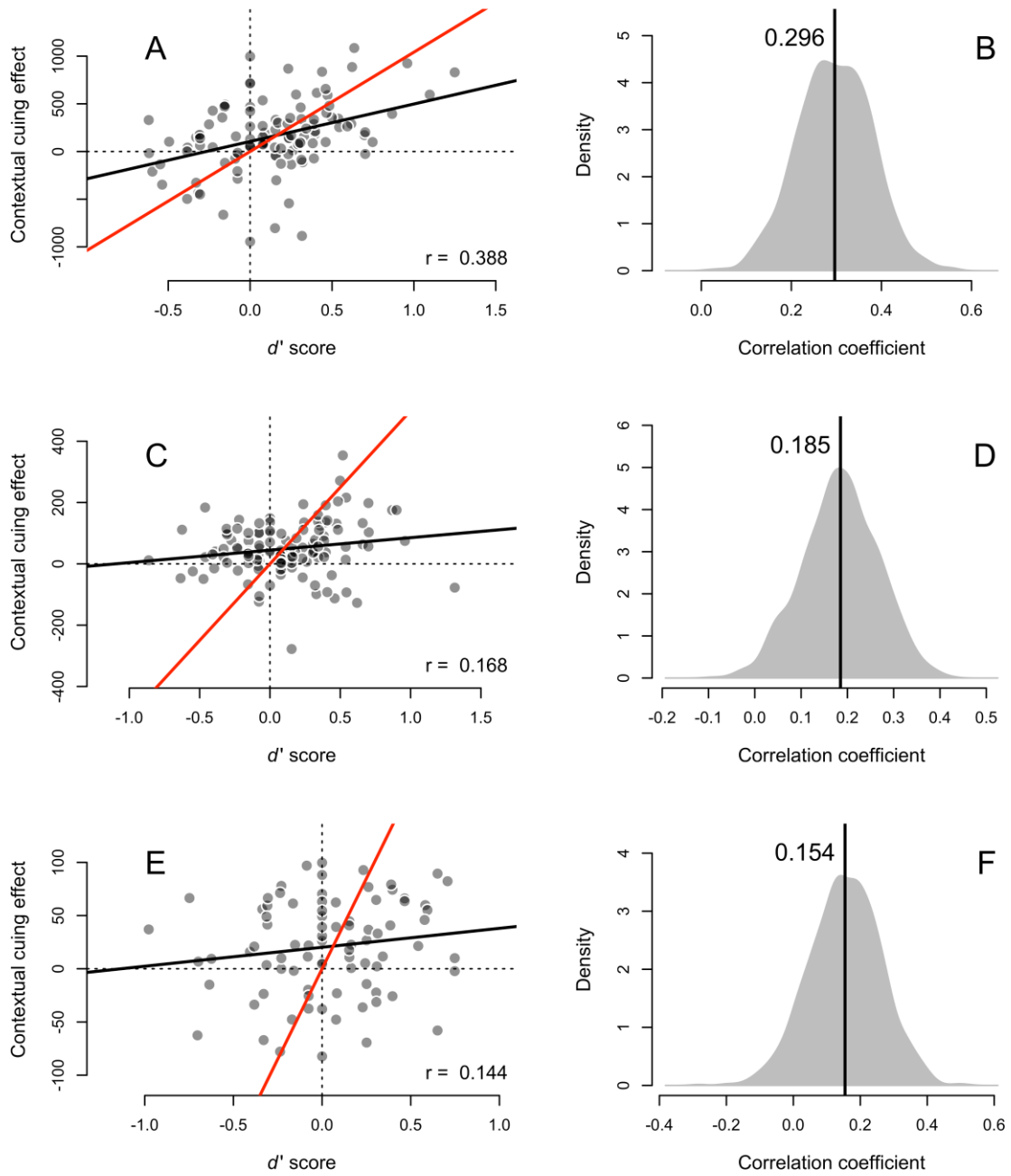


Figure #5