Psychon Bull Rev (2016) 23:1996–2009
DOI 10.3758/s13423-016-1063-0

Contextual cuing as a form of nonconscious learning: Theoretical and empirical analysis in large and very large samples

Ben Colagiuri[1,2] · E. J. Livesey[1]

Published online: 24 May 2016

Abstract Numerous studies have demonstrated that associative learning can affect visual cognition. In one such effect, search times for a target hidden among similar distractors are faster for repeated search configurations compared with novel configurations. This contextual cuing effect is particularly interesting, because researchers routinely have failed to find evidence of recognition of the repeated configurations, concluding that the effect is a form of nonconscious learning. Vadillo, Konstantinidis, and Shanks (2016) recently criticized this conclusion on a number of methodological and conceptual grounds that suggest the area suffers from a high probability of false-negative results on awareness tests and misinterpretation of weak or absent relationships between cuing and awareness measures. We developed further predictions from theoretical models assuming that single or independent memory sources drive learning and awareness and discuss how these predictions fare in three new contextual cuing experiments involving large (n > 60) and very large samples (n > 600). The data support the absence of a positive relationship between recognition and the cuing effect both at the participant and configuration level, the probability of which being a false negative is very low in a model assuming a single memory source drives learning and awareness. This was the case using both conventional and Bayesian analyses. The combination of this theoretical and empirical analysis suggests that contextual cuing is not dependent on cue recognition and provides evidence that it reflects a genuine form of nonconscious learning.

Keywords Associative learning · Awareness · Implicit learning · Implicit memory

Introduction

Few topics in experimental psychology have generated as much research and so little consensus as nonconscious learning. Whereas there have been many results claiming to demonstrate learning without awareness, critical reviews of the literature routinely conclude that most of the evidence for nonconscious learning is unconvincing (Lovibond & Shanks, 2002; Shanks & St John, 1994). Consequently, some theorists propose that all forms of human learning can be explained in terms of a single learning system, of which explicit knowledge is the unifying currency (Mitchell, De Houwer, and Lovibond, 2009). Nonetheless, a recent learning phenomenon, labeled contextual cuing (Chun & Jiang, 1998) has attracted attention based on initial evidence, suggesting that it may occur in the absence of conscious knowledge.

Contextual cuing involves an incidental learning procedure disguised as a visual search task. In Chun and Jiang's (1998) original demonstration, participants search for a target, the letter T, rotated 90 degrees clockwise or counter clockwise. This target is hidden among a field of distractors, such as numerous letter Ls, rotated 0, 90, 180, or 270 degrees. The participant's task is to locate the T and respond by reporting whether it is rotated clockwise or counter clockwise. Some of the configurations are repeated, meaning that the target-distractor arrangement can serve as a cue for the target's

Electronic supplementary material The online version of this article (doi:10.3758/s13423-016-1063-0) contains supplementary material, which is available to authorized users.

✉ Ben Colagiuri
ben.colagiuri@sydney.edu.au

[1] School of Psychology, University of Sydney, Sydney, NSW 2006, Australia

[2] School of Psychology, University of New South Wales, Sydney, Australia

location. Contextual cuing occurs when reaction times are faster on these cued trials (repeated configurations) than on uncued trials (novel configurations).

The typical method of assessing whether participants develop explicit knowledge of the relationship between distractors and target position involves a post-experiment test in which participants are asked either to make "old" versus "new" judgments about both repeated and novel configurations (recognition test: Chun & Jiang, 1998) or to predict the location of the target given a configuration of distractors, which could be one that was repeated or a novel configuration (generation test: Chun & Jiang, 2003). These two awareness tests have produced similar results; many studies either failed to find evidence of better recognition for repeated configurations (Chun & Jiang, 1998; Chun & Phelps, 1999; Colagiuri, Livesey, & Harris, 2011; Howard, Howard, Dennis, Yankovich, & Vaidya, 2004; Manns & Squire, 2001; Rausei, Makovski, & Jiang, 2007) or found generation performance equivalent to chance (Chun & Jiang, 2003; Jimenez & Vazquez, 2011), despite clear evidence of contextual cuing effects. Even in studies that have detected some degree of knowledge of the repeated configurations, that is performance above chance, contextual cuing appears unaffected by the participant's level of awareness, with cuing effects reliably observed in sub-groups of participants who show no explicit awareness (Preston & Gabrieli, 2008; Smyth & Shanks, 2008 Experiment 2, but not Experiment 1; Vaidya, Huger, Howard, & Howard, 2007). On this basis, a recent review by Goujon, Didierjean, & Thorpe (2015) came to the conclusion that the contextual cuing effect does not require conscious knowledge. Certainly on face value, such evidence for learning in the absence of awareness appears to contradict single-system accounts of learning that necessarily entail conscious deliberation (Mitchell et al., 2009).

Recently, however, Vadillo, Konstantinidis, & Shanks (2016) reviewed the literature on contextual cuing and noted several methodological and conceptual shortcomings that cast considerable doubt on much of the evidence used to argue that contextual cuing reflects nonconscious learning. One of their major criticisms was that most contextual cuing experiments were simply too underpowered to confidently assess the presence of awareness and that the prevalence of false negatives therefore might be quite high. For instance, they noted that the statistical power to detect significant awarness was only .21 when using the median sample size (N=16) and estimated effect size from their meta-analysis, which was comparable to the actual number of significant results on awareness tests from the experiments they reviewed. Furthermore, when Vadillo et al. meta-analysed the results on available awareness tests they found a significant positive effect with Cohen's d_z = 0.31 (95% confidence interval [CI] 0.24-0.37). Vadillo et al. made several further persuasive arguments about why the evidence for nonconscious learning had been overstated in the contextual cuing literature. Here, we present a theoretical and empirical analysis taking those criticisms into consideration and provide evidence supporting the status of contextual cuing as reflecting nonconscious learning. The empirical data come from three new contextual cuing experiments. The full methodological details of the three experiments are presented in the *Supplementary Materials*. The reader may prefer to review those before going further. Briefly, the three experiments involved standard contextual cuing tasks with large (Experiments 1, n = 63, and 2, n = 84) and very large samples (Experiment 3, n = 766). They used a combination of two-alternative forced choice (Experiments 1 and 3) and old-new recognition tests (Experiment 2). All three experiments showed significant contextual cuing effects (Cohen's d range 0.84-0.99, all $p < 0.001$; Table S1) as well as recognition that was significantly higher than chance (Cohen's d range 0.31-0.47, all $p < 0.001$; Table S2). The critical theoretical and empirical analysis pertains to the relationship *between* cuing and recognition, which we describe in detail next.

Pitfalls of claiming evidence of nonconscious learning

There is good reason to doubt whether a weak correlation between cuing and recognition really constitutes strong evidence for their psychological independence and therefore nonconscious learning. Even a model in which priming of reaction times (i.e., cuing) and recognition are served by exactly the same memory system can account for a weakly positive but nonsignificant correlation between the two measures (Berry, Shanks, & Henson, 2008a, b). This is because performance on a recognition test may be affected by sources of error that are independent from those that affect performance on the primary task, in this case contextual cuing.

A common practice in studying the role of awareness in learning is to divide participants in a post hoc fashion into groups that show evidence of awareness and those that do not. In some cases, authors have claimed evidence for nonconscious learning if participants who do not show awareness still reveal evidence of learning on another measure (e.g., cuing) and if aware and unaware participants do not appear to differ substantially on that measure. In such cases, awareness appears to have little relationship with learning. While some critics have explicitly advocated the use of this form of analysis (Lovibond and Shanks, 2002), the practice has recently come under stronger criticism (Shanks & Berry, 2012). Interpreting the presence of a significant learning effect amongst those deemed to be unaware appears to be particularly problematic. For example in contextual cuing tasks, simply through sampling variation, participants classified as having chance-level recognition of the repeated configurations could nonetheless exhibit a cuing effect. As such, some of the apparent evidence for significant contextual cuing in the absence of recognition may simply be a statistical artifact. Vadillo et al. (2016) illustrate this criticism using a simple statistical model that assumes a single memory source *s*, supplying shared variance on the two performance measures, and independent variance *e*. They used the following equations to calculate an index analogous to contextual cuing (RT; Eq. 1) and an index of awareness (d; Eq. 2), where mean = 1 and sd = 1 for s and mean = 0 and sd = 1 for e.

$$\mathrm{RT} = 100\mathrm{s} + 30\mathrm{e} \qquad (1)$$

$$d = 0.30s + \mathrm{e} \qquad (2)$$

Using this model, Vadillo et al. demonstrated that participants with $d < 0$ still produce a substantial RT effect, well above 0. They also note that the correlation between the two measures is modest at best despite the fact that the source of knowledge was the same for both measures. The authors illustrated this effectively using a sample of 1,000 simulated participants.

Relationship between cuing and recognition as a function of sample size

Theoretical models and simulated data

The above and other arguments made by Vadillo et al. (2016) are, in our view, quite compelling. However, the thrust of the current analysis is based on some further predictions made by this single-system model as function of sample size. These predictions warrant close inspection given that one of the major criticisms of the contextual cuing literature has been the use of small samples.

First, the expected correlation between cuing and recognition (approximately 0.275 according to Eqs. 1 and 2) does not change as a function of sample size, but the reliability of this correlation certainly does increase as the sample increases. In fact, according to the single-system model presented by Vadillo et al (2016), with a very large sample, correlations close to zero are extremely unlikely. Thus, nonsignificant correlations around zero that are calculated from large samples *do* meaningfully question the plausibility of a simple single-system model in which the same knowledge contributes to variance on both measures. Second, although Vadillo et al. found a significant cuing effect in the subsample of their simulated participants with $d < 0$ (suggesting that dividing participants post hoc in this fashion is fairly meaningless), the cuing (RT) difference between participants with $d <0$ and participants with $d > 0$ with large samples would be strongly significant. That is, using this single-system model with a large enough sample, the difference in cuing between aware ($d >$ 0) and unaware ($d < 0$) should almost never be nonsignificant, even if there is a significant cuing effect in the unaware participants.

To illustrate these two observations, we took the basic statistical model reported by Vadillo et al. (2016) and updated the parameters based on our empirical data. Specifically, as per Vadillo et al., we first created a sample of scores for s from a normal distribution with mean and SD of 1 and then multiplied the distribution by the mean cuing effect observed across our three experiments (mean = 55 ms), producing a distribution with mean and SD of 55. Assuming that means and variances of independent random variables sum to produce the observed distribution, we then used the observed sample variance from our data (SD = 60 ms estimated across the 3 experiments, i.e. variance=3600) and subtracted the s variance, producing an estimate of variance from e of SD = 24. We used the same procedure to estimate s and e contributions to the awareness measure d, but following Vadillo et al., we express these in standardized units (from the data mean = 0.4, SD = 1.0) to reflect that types of recognitions tests vary. This provided indices of cuing (RT) and awareness (*d*) shown in Eqs. 3 and 4, respectively.[1]

$$\mathrm{RT} = 55\mathrm{s} + 24\mathrm{e} \qquad (3)$$

$$d = 0.40\mathrm{s} + 0.9\mathrm{e} \qquad (4)$$

Using this model, we ran 10,000 simulated experiments at each of a range of sample sizes varying from N = 12 to N = 1, 000. For each simulated experiment, we ran one-sample *t* tests comparing *d* to a score of zero (i.e., no knowledge expressed on the awareness test) and also calculated the correlation between RT and *d*. We then split the sample in each experiment according to whether $d \leq 0$ (at or below chance) or $d > 0$ (above chance). We ran further one-sample *t* tests comparing the RT effect in each of the subsamples against 0, and an independent samples *t* test comparing the two subsamples. For each of these four *t* tests, we then recorded whether the null was rejected at a critical alpha of $p < 0.05$. Note that because the model possesses a single memory source s that is known to be greater than zero, we can consider a failure to reject the null for the one-sample *t* test on *d* to be a false-negative. Likewise, because the s affects both RT and *d*, we can consider a failure to reject the null on the independent samples *t* test to be a false-negative (participants with higher *d* should display higher RT effects on average). Figure 1A illustrates the proportion of null results for these four tests as a function of sample size. By way of comparison, Fig. 1B illustrates the same functions produced by a model that differs only in that s is sampled independently for RT and for *d*, as would be expected if one were to assume that two completely independent sources of memory supported contextual cuing and explicit knowledge about repeated configurations (i.e. a 'full' independence model).

[1] Note that when using Vadillo et al.'s (2016) parameters, i.e. those in Eqs. 1 and 2, the simulations produce a highly similar pattern of results.

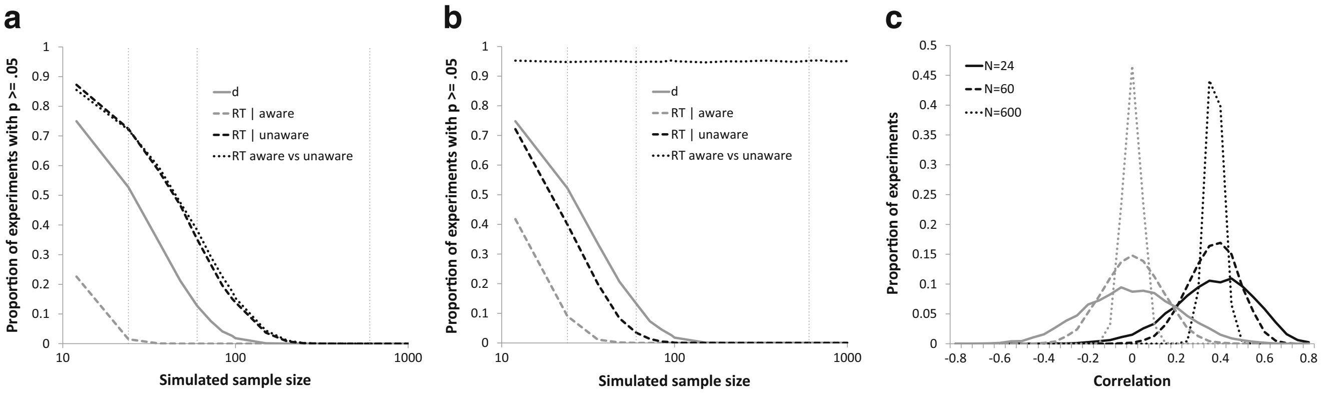

Fig. 1 Results of simulated cuing experiments that vary the sample size between 12 and 1000 and examine two hypothetical measures, RT and d. Panels A & B show proportion of simulated experiments that result in a failure to reject the null ($p \geq 0.05$) as a function of sample size for four critical *t* tests: above chance performance on awareness test (*d*); cuing RT effect amongst aware participants (RT | aware); cuing RT effect amongst unaware participants (RT | unaware); difference in size of cuing effect for aware versus unaware participants (RT aware vs unaware). Simulations in Panel A used a single memory source s for both measures. Simulations in Panel B used independent s for each measure. Note that sample size is displayed on a log scale, dotted vertical lines cross at N = 24, N = 60, and N = 600. Panel C displays the proportion of simulated experiments with correlations between RT and d ranging from −0.8 to 0.8 (in bins of 0.05), for three sample sizes, N = 24, N = 60, and N = 600. Black lines indicate results from the single s simulations from Panel A. Grey lines indicate results from the independent s simulations from Panel B

Examining Fig. 1A, it is clear that even in a single-system model the proportion of false negatives for all tests is very high for small samples but that it decreases markedly by the time N = 100. Notably, the proportion of statistically nonsignificant RT effects in the subsample with $d \leq 0$ essentially falls to zero as sample size increases, further confirming Vadillo et al.'s (2016) suggestion that such an analysis is not informative about the nature of the underlying memory systems involved. However, this is also true of the *t* test comparing the aware and unaware subsamples, suggesting that a difference in cuing effect should generally be evident between the two so long as a relatively large sample is used. In comparison, the model with two independent memory sources shown in Fig. 1B predicts essentially the same pattern for the three one-samples *t* tests (the model still assumes some explicit knowledge and predicts that aware and unaware participants will show an RT effect) but not the independent samples *t* test comparing the two subsamples of participants who scored $d \leq 0$ and $d > 0$. This is because the sampling of s that contributes to *d* scores is completely independent of the sampling of s that contributes to RT. Thus, the null hypothesis is not rejected approximately 95% of the time, as one would expect from a true null effect.

The third panel of Fig. 1 illustrates the relative frequency of the correlation between RT and *d* using samples of N = 24, 60, and 600. It is noteworthy that in the single-system model, although the correlations have the same modal value (i.e., the expected correlation is about 0.37 regardless of N), the correlations for the large samples are far more narrowly distributed around this value. In this model, a zero correlation with a very large sample would be quite unusual. For example, even with an N of 60, fewer than 1% of experiments simulated using Eqs. 3 and 4 would yield $r \leq 0.05$, and with an N of 600, virtually no simulated experiments yield $r \leq 0.2$. In contrast, the correlations from the independent s model are of course expected to be zero and are distributed around this value accordingly.

Consistent with Vadillo et al.'s (2016) criticisms, these analyses suggest that a swathe of contextual cuing studies with relatively small samples may not be very informative about the potential independence of the cuing effect from conscious knowledge. Critically, however, these analyses also outline the circumstances under which evidence for nonconscious learning could be obtained from contextual cuing experiments with large samples, i.e., if the independent samples *t* test comparing RT between unaware ($d \leq 0$) and aware ($d > 0$) subsamples is not statistically significant and the correlation between the two measures is near-zero.

Empirical tests from our three experiments

In our three experiments, the relationship between cuing and recognition at the participant level (i.e., aggregating across all repeated configurations – the conventional level at which such effects are assessed) was first examined by calculating correlations between the two measures and also by comparing the mean cuing scores produced by participants classified as 'aware' (above chance recognition) versus 'unaware' (at or below chance recognition). The results of both these analyses are shown in Table 1. Notably, despite the presence of above chance awareness in each sample overall, there was no case in which the cuing effect was larger in the aware sample relative to the unaware sample. In fact, the strongest hint of a relationship was a negative one, found in Experiment 1. As demonstrated in the simulated data presented above, assuming the cuing and recognition effects in these experiments came from the same memory source as in the single-system model, the

Table 1 Comparison of cuing effects for "aware" and "unaware" participants

	Correlation		Above chance			At or below chance			Above vs. below		
	r	*p*	N	Mean	SD	N	Mean	SD	t	*p*	d_z
Ex 1	−0.14	0.29	36	39.9	56.9	27	56.9	54.0	−1.20	0.24	0.31
Ex 2	0.11	0.30	53	63.0	55.6	31	53.1	67.3	0.73	0.47	0.17
Ex 3	0.03	0.34	424	57.6	64.2	342	55.9	64.6	0.36	0.72	0.03

Classification of participants as aware or unaware was based on whether participants scored greater than 16/32 in the 2AFC recognition test in Experiments 1 and 3 or whether the scored d' > 0 for the 64 old vs. new trials in Experiment 2.

likelihood of observing no relationship between cuing and recognition in all three experiments (including one with an N exceeding 700) is very low. We will report Bayes factor (BF) tests for these analyses later in a section dedicated to BF analysis. For now, this conventional analysis at the participant level suggests that that there is far from convincing evidence of a positive relationship between cuing and recognition precisely where single-system models predict one should be evident.

However, to be fair to the single-system approach, the Vadillo et al. (2016) calculations simplify the estimates of variance by assuming that s has a standard deviation equal to its mean. Thus the proportion of overall variance attributed to s and e are arbitrarily chosen, and other values could equally apply. The values estimated here attribute around 84% of the variance in RT (i.e., cuing effect) to s and around 84% of the variance in *d* (i.e., the awareness measure) to e. In contrast, the single-system model proposed by Berry, Shanks, Speekenbrink, & Henson (2012), for instance, is free to assume that *any* proportion of the observed variance could be attributable to s. When this proportion is small, the expected correlation is closer to zero. Later, we test whether alternate versions of the single-system model can better fit our observed data. For now, we will simply note that we have simulated the effect of varying the proportion of variance attributed to s and e and have found that zero correlations are highly improbable at most values. For instance, according to the single-system model, with N = 600 the probability of observing a near-zero correlation such that r < 0.05 is vanishingly small unless more than 80% of the sample variance comes from e rather than s on both measures. Even when e accounts for 90% of the variance, we would only expect r < 0.05 approximately 10% of the time. Further, according to the single-system model, with N = 60 the probability of observing a negative correlation such that r < −0.1 (as observed in Experiment 1) is also highly improbable ($p < 0.05$) unless 90% or more of the sample variance comes from e rather than s on both measures. The same applies to analysis comparing participants with $d < 0$ and $d > 0$; at n = 600 the probability of a false negative is remote ($p < 0.05$) unless the proportion of variance coming from e is greater than 80%. As we will discuss later, the plausibility of a model that attributes very little variance to s is questionable. However, as an interim summary, we feel that the single-system model on the whole predicts a positive relationship when large samples are used and their absence is thus noteworthy.

Importance of analysis at the individual configuration level

Theoretical models and simulated data

The participant level analysis above seems to indicate evidence against a single-system model in which a single memory source drives cuing and recognition. This could be taken as evidence for nonconscious learning. However, Vadillo et al. (2016) also raise a further issue with analysis of cuing and awareness, previously examined by Smyth and Shanks (2008). Contextual cuing tasks usually involve between 8-12 unique repeated configurations (our experiments all involved 8), over which visual search RTs and recognition scores are averaged. Thus, one possibility is that cuing and awareness are closely related but that participants only learn a small subset of the repeated configurations. If this is the case, then the averaging process is much more likely to conceal evidence of learning on recognition tests, which often include only two presentations of any single repeated configuration and typically involve only a binary old/new response. Supporting this argument, Smyth and Shanks (2008) used 99% confidence intervals to estimate that, on average, their participants only displayed what they deemed to be convincing evidence of a cuing effect for 1-2 configurations out of 12. If it is only this small subset of configurations that drives the overall cuing effect, then it is quite feasible that averaging at the participant level may obscure a genuine relationship between cuing and recognition.

For our three experiments, we estimated the number of configurations for which learning is 'truly' evident using a different approach (but with similar results). This analysis acknowledges and tries to correct for the high false-positive rate in the number of configurations classified as 'above chance' on the cuing and recognition measures. Positive cuing scores (mean RT for cued trials < mean RT for uncued trials) were considered "above chance," whereas negative scores were considered 'below chance.' This means that for both cuing and recognition, we used a very liberal classification of what constitutes a configuration with above chance performance, with a 0.5 probability of being labeled as above chance when no learning has taken place. For both measures, we assumed that the probability of a configuration producing an above chance score is a combination of the true probability of having learned something about that configuration and the guess rate, as expressed in this equation:

$$p(O) = p(T) \times p(O|T) + (1-p(T)) \times p(O|{}^{\sim}T) \quad (5)$$

Here, p(O) is the observed probability of a configuration being above chance, p(T) is the true probability that useful information has been learned about the configuration (T), p(O|T) is the probability that the configuration will be above chance given T, and p(O|~T) is the probability that the configuration will be above chance given that no useful information has been learned (~T). From this equation, p(T) – and thus also the proportion of the eight repeated configurations for which learning actually occurred – can be estimated. We assumed p(O|~T) equals 0.5, that is the guess rate as described above. For the purposes of this analysis, we assumed p(O|T) = 1. This is a conservative approach, because it assumes that all configurations for which the information relevant to cuing has been learned (those with 'true' cuing) will produce a positive cuing score and similarly that all configurations for which information relevant to recognition has been learned (those with 'true' recognition) will produce recognition scores above chance. As such, this approach is likely to *underestimate* the number or configurations that are truly learned about because it seems rational that true values of p(O|T) are less than 1 (but still higher than the guess rate).

In terms of the cuing effect, the mean number of positive cuing scores out of 8 were on average, 5.13, 5.57, and 5.21 for Experiments 1-3 respectively. Using the above approach, this means that an average of 2.26, 3.14, and 2.42 out of 8 configurations were estimated to demonstrate a 'true' cuing effect. In terms of recognition, an average 4.43, 4.30, and 4.42 out of 8 configurations had recognition scores above chance, which means that an average of 0.86, 0.60, and 0.84 out of 8 configurations were estimated to demonstrate 'true' recognition. These numbers are reasonably consistent with those reported by Smyth and Shanks (2008), although it should be emphasized that they are *minimum* estimates of the number of configurations that were learned. Thus, the real proportions of configurations learned are at least this size but may be larger depending on the sensitivity of the measures employed.

What then should we expect to find from an analysis of the relationship between cuing and awareness at the level of individual configurations? To begin to answer that question, we developed two configuration level models, one that assumes some learning of all configurations ('all cues' model) and one that assumes strong learning on only a subset of configurations ('subset' model) and ran them with single-system assumptions (i.e., a common s) and independent systems assumptions (i.e., independent s). Note that the results of interest from the independence model are essentially the same irrespective of whether an 'all cues' or 'subset' model is used, thus for simplicity, we report only the former. As a starting point, we again took Vadillo et al.'s (2016) simple single-system model of RT and *d* but this time created 8 samples for RT and *d* for each simulated participant. These samples are intended to reflect effects for individual configurations, which naturally have much higher variance than the mean cuing and recognition effects. Thus, we created new estimates based on the observed means and variances at the configuration level across the three experiments, as indicated in Eqs. 6 and 7:

$$\mathrm{RT} = 55\mathrm{s} + 158\mathrm{e} \quad (6)$$

$$d = 0.2\mathrm{s} + 0.98\mathrm{e} \quad (7)$$

For the 'all cues' model, as in the participant level simulations, for the single-system model s was sampled from a normal distribution with mean = 1 and SD = 1 and e from two independent normal distributions with mean = 0 and SD = 1 (in combination with Eqs. 6 and 7). Hence, these estimates are again based on a simple but fairly arbitrary method following Vadillo et al. (2016) of assuming that the mean and SD of s will be the same, with the remainder of the variance attributed to e. Accordingly in this model, Eqs. 6 and 7 attribute approximately 10.8% of the variance in cuing to s and approximately 4% of the variance in the awareness measure to s. These represent values where the proportion of variance coming from s is quite small, and thus the predicted relationship between the two measures is fairly weak.

As noted earlier, the single-system model is free to assume that there is almost no variance in s. In doing so, it is possible for the model to predict near-zero correlations. However, to suggest that the strength of learning (s) is largely invariant across configurations and individuals is, we think, quite implausible. Most obviously, we note that it is entirely incompatible with Smyth and Shanks' (2008) and Vadillo et al.'s suggestion that only a subset of configurations are learned by any given individual. For instance, the analysis above suggests that participants learn about a minimum of 30% of the configurations (although possibly more) to produce the pattern of cuing scores observed in our experiments. If participants only learned approximately 30% of configurations then the observed effect would have to be entirely generated by these configurations, in which case the mean cuing score for 70% of configurations would be zero and the mean cuing score for the remaining 30% would be more than triple the overall mean. Even if this were the *only* source of variance in s (i.e., assuming s = 0 for 70% of configurations and s = 3.333 * observed mean for the remaining 30%), then we estimate[2] that around 25% of the observed variance in cuing scores for individual configurations would have to be attributable to variance in s rather than e. Put simply, the single-system model can only predict correlations approaching zero if almost all of the variance comes from e and this can only happen if s is generally invariant across configurations.

To demonstrate this, we created an additional version of the model ('subset' model), where the s values for a random 30% of configurations were 3.333 times the effect size estimated in Eqs. 6 and 7 (i.e., 183.3, 0.6) and zero for the other 70%, with the remaining variance attributed to e, hence sampled with a normal distribution with mean = 0 and SD = 145. This 'subset' model provides a simplified means of simulating the conditions in which only a small proportion of the configurations were actually learned but is conservative in the sense that it maximizes the variance that can be attributed to e by assuming that s can only take one of two values for any given configuration.

For each simulated participant in each model, we then split the configurations according to whether $d \leq 0$ (at-or-below chance) or $d > 0$ (above chance) and averaged RT for each subset of configurations, yielding two values for each participant. We ran paired-samples *t* tests comparing these two RT scores against each other. As with the previous simulations, we did this for 10,000 experiments at each of a range of sample sizes varying from N = 12 to N = 1,000; the results are summarized in Fig. 2, which shows the mean RT score for configurations with $d > 0$ and $d \leq 0$, and the probability of obtaining $p \geq 0.05$ on the comparison of these two RT scores. When either of these two models use a common s, as proposed by Vadillo et al. (2016), the mean RT effect for configurations with $d > 0$ is consistently predicted to be larger than the mean RT effect for configurations with $d < 0$. Furthermore, the probability of finding a significant difference between the two increases with sample size, as one might reasonably expect. In comparison, the probability of the difference being significant when independent s is assumed hovers at the type-I error rate of .05 and reflects the fact that on average, the model predicts equal RT effects for configurations with $d > 0$ and $d < 0$.

Empirical tests from our three experiments

To test these possibilities in our three experiments, we first classified individual configurations according to whether their recognition scores fell above or below chance and calculated two cuing scores for every participant averaging cuing for the above chance configurations and cuing scores for the below chance configurations.

For Experiments 1 and 3, which used a 2AFC test, this involved calculating a recognition score out of four for each repeated configuration based on the number of correct responses to that particular configuration in the recognition test. Configurations with scores <2 out of 4 constituted below chance performance, while those >2 out of 4 constituted above chance performance. We used confidence ratings to classify configurations with scores of 2 as above or below chance. Here a recognition composite was calculated for each configuration by allocating a positive confidence score to correct responses and a negative confidence score to incorrect responses in the recognition test and then averaging these scores. Configurations with scores of 2 and confidence composite score below zero were classified as below chance, whereas those with confidence composite scores above zero were classified as above chance. In Experiment 2, which used an old/new judgement task, individual configurations were judged as having above chance recognition if the proportion of old judgements for that configuration was higher than the proportion of old judgments for all new trials. One participant in Experiment 1, two participants in Experiment 2, and 13 participants in Experiment 3 were removed from this analysis because they either had no configurations classified as above chance or no configurations classified as below chance.

The mean number of configurations classified as above and below chance for each Experiment are shown in Table 2, along with the mean cuing scores for these two classifications. The most striking feature of this analysis is that configurations classified as producing below chance recognition actually have larger cuing scores than those classified as producing above chance recognition. This difference was statistically significant in Experiments 1 and 3 but fell short of significance in Experiment 2. This pattern of results clearly contradicts the simulated patterns in a single memory source model in which learning approximately 30% of configurations occurs ('subset' model). In this case, studies with sample size of n = 60 (Experiments 1 and 2) would yield less than 33% false negatives, but those with n = 600 (Experiment 3) would yield virtually no false negatives (less than 1 in 10,000). Even the 'all cues' model with some s on 100% of configurations, which assumes that most of the variance comes from e as outlined in Eqs. 6 and 7, still predicts that false negatives would be relatively rare (<10%) when n = 600. Both single-system models predict that a significant *negative* effect as observed in Experiments 1 and 3 should be extremely rare (≤0.15%) even with N = 60. In contrast, a model assuming full independence of s in the RT and *d* measures naturally predicts that null and near-zero results (positive or negative) should occur most of the time. The presence of a significant negative difference in cuing comparing configurations above and below chance in recognition is problematic for the independence model too, though the probabilities of these occurring under the independence model are much higher than

[2] Assuming variance from s will approximate $.3(3.333*s - s)^2 + .7(0 - s)^2$

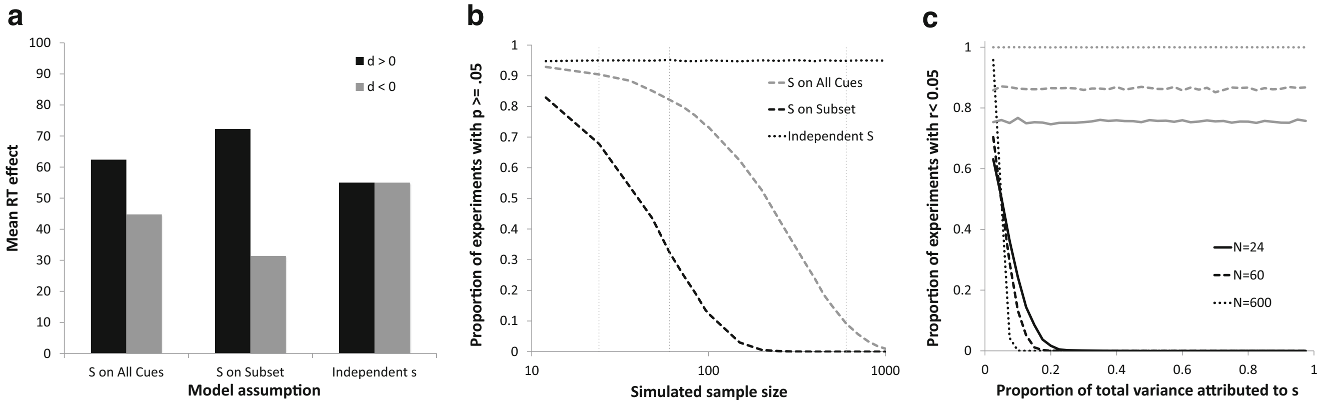

Fig. 2 Results of simulated experiments that vary the sample size between 12 and 1000 and examine two hypothetical measures, RT and *d*. Knowledge (s) for 8 configurations was simulated for each 'participant' for two single-system models, one with some learning on 'all cues' and on with learning on only a 'subset' of cues, as well as an independent s model. Panels A shows mean RT cuing for configurations with $d > 0$ and $d \leq 0$ for each of these models. Panel B shows proportion of simulated experiments that result in a failure to reject the null ($p > 0.05$) as a function of sample size for a dependent samples *t* test comparing RT cuing for configurations with $d > 0$ to RT cuing for configurations with $d < 0$ (note that sample size is displayed on a log scale, dotted vertical lines cross at N = 24, N = 60 and N = 600). Panel C shows the proportion of experiments yielding a near-zero or negative correlation (r < 0.05) between RT and *d* assuming there is learning on 'all cues' across the range of possible variance that could be attributed to s. Black lines show results with a common s (single-system model) and grey lines show results for independent s

under the single-system model (as indicated by the *p* values reported in Tables 2 and 3).

While dichotomising data according to awareness measures (i.e., aware vs. unaware comparisons) are heavily relied upon in the literature, one could argue that such procedures may obscure important variance in the relationship between cuing and recognition. To rule out the possibility that dichotomising did obscure a genuine positive relationship, we also compared the cuing effect across two continuous measures of recognition. The first measure was simply the 'N Hits' tallying number of correct recognition choices in the 2AFC in Experiments 1 and 3 or the number of correct 'Old' judgements for a repeated configuration in Experiment 2. These produced a score ranging from 0-4 for each configuration. The second measure was a recognition composite score, which as described above, involved averaging the confidence ratings for each recognition judgement with correct 2AFC choices positively and incorrect 2AFC choices scored negatively. We created a similar composite score for Experiment 2 using familiarity ratings, where the mean rating for new configurations was subtracted from the mean rating for the repeated configuration to create a composite score for each configuration. Both types of composite score could range from -100 to 100, with zero representing chance responding. We then tested the correlation between cuing and recognition with individual configurations as observations. These data are presented in Table 3. In all three experiments and for both recognition measures, the correlations were always small and negative (range r = −0.02 to −0.09). Albeit very small, the *p* values for these correlations hovered around *p* = 0.05 with approximately half being statistically significant and the other half not reaching statistical significance. The correlational analysis of continuous recognition was therefore consistent with the dichotomised data in terms of demonstrating the absence of any positive relationship between cuing and recognition and suggesting that if any relationship does exist, that it is a negative one.

As such, our data clearly indicate the absence of positive relationship between cuing and recognition, even when assessed at the configuration level *and* irrespective of whether recognition is treated as dichotomous or continuous. This suggests that the cuing effect is not driven by recognition and reflects learning that is independent of a simple measure of awareness. It seems quite reasonable, then, to assume cuing in these experiments reflects a form of nonconscious learning.

Table 2 Mean cuing effects for 'aware' and 'unaware' configurations

		Above chance			At or below chance			Above vs below		
	N	n/8	Mean	SD	n/8	Mean	SD	t	*p*	d_z
Ex 1	62	4.44	38.1	83.4	3.56	71.4	87.6	−2.05	0.040	−0.26
Ex 2	82	4.30	48.3	94.0	3.70	74.0	105.7	−1.74	0.082	−0.19
Ex 3	753	4.41	49.6	95.4	3.57	60.7	99.1	−2.36	0.026	−0.09

Table 3 Mean (SD) cuing effects for continuous measures of recognition, namely by recognition score (**A**) and recognition composite score (**B**). For recognition composite score, the mean recognition composite score (recog) and relevant cuing effect are reported ordered by rank of the recognition composite score within-subjects. This is for representational purposes with the correlational analysis being conducted with individual configurations as the observations

A. Cuing score as function of N Hits (range 0-4)

	N		0	1	2	3	4	r	*p*
Ex 1	504	Cuing (SD)	67.9 (134)	65.2 (148)	49.6 (159)	31.4 (184)	37.4 (162)	−0.07	0.10
Ex 2	672	Cuing (SD)	90.5 (161)	73.5 (174)	52.2 (156)	45 (170)	48.9 (145)	−0.08	0.03
Ex 3	6128	Cuing (SD)	70.5 (103)	62.5 (166)	53.7 (165)	52.4 (181)	58.3 (156)	−0.02	0.10

B. Cuing score as function of recognition composite score (range −100 to 100)

	N		1st	2nd	3rd	4th	5th	6th	7th	8th	r	*p*
Ex 1	504	Recog(SD)	35 (19.7)	24.2 (17.2)	16.6 (15.6)	9.5 (12.9)	0.9 (11.6)	-6.2 (12.5)	-14.1 (14.1)	-27.4 (15.9)	−0.09	0.054
		Cuing (SD)	31.4 (161)	46.8 (192)	6.8 (163)	16.5 (158)	93 (166)	47 (163)	38.3 (140)	91.9 (145)		
Ex 2	672	Recog(SD)	18.9 (11.5)	12.2 (8.2)	8.3 (7.3)	4 (6.4)	0.6 (5.2)	-2.8 (6.0)	-6.2 (5.7)	-12.1 (8.8)	−0.08	0.04
		Cuing (SD)	31.9 (146)	42.7 (148)	47.3 (167)	84.4 (167)	60.2 (175)	71.1 (157)	70.7 (146)	63.6 (143)		
Ex 3	6128	Recog(SD)	37.5 (20.3)	24.9 (16.4)	15.8 (14.6)	7.5 (13.4)	-0.3 (13.1)	-8.6 (13.9)	-19.1 (16.7)	-33.3 (20.9)	−0.02	0.09
		Cuing (SD)	50.3 (181)	47.7 (181)	54.7 (174)	47.5 (190)	71.8 (174)	53.5 (183)	67.6 (173)	59.8 (169)		

Limitations to NHST when attempting to prove the null and Bayes Factor alternatives

Theoretical considerations

The combination of the theoretical and empirical analysis above, suggest evidence against a positive relationship between cuing and recognition. However, any attempt to provide positive evidence for learning without awareness faces an inherent difficulty in that it essentially requires demonstration of a null effect, whether it be a failure to find evidence of awareness or a failure to find a relationship between cuing and recognition, which null hypothesis significance testing (NHST) is not well equipped to test (Nickerson, 2000).

Thus, another persuasive argument made by Vadillo et al. (2016) is that the existing preponderance of null results using NHST on underpowered and insensitive awareness tests is not particularly informative about the presence or absence of awareness in contextual cuing. One of their suggested solutions is to use Bayes factors (BF) to compare the evidence for the null against the evidence for an alternative. This suggestion is appealing for many reasons, particularly because it takes into account the quality of the evidence for different hypotheses in a way that is not captured by traditional NHST approaches. The real sticking point with BFs, and the greatest source of flexibility in using them, is how to define the hypotheses that are to be compared. Vadillo et al. use an approach advocated by Rouder, Speckman, Sun, Morey, & Iverson (2009) in which the alternative hypothesis comprises a Cauchy distribution of possible effect sizes that differ from zero. In some respects, this approach is relatively agnostic about the theoretical assumptions on which the alternative hypothesis might be based. That is, the alternative hypothesis is simply that the effect size is non-zero, probably small but possibly quite large. In contrast, Gallistel (2009) has suggested an approach in which researchers can derive an alternative hypothesis that is more precisely bounded by theoretically meaningful values, where appropriate. For instance in considering the comparison of RT cuing effects in subsamples of participants who are above and at-or-below chance on an awareness test, the single-system model makes a very clear prediction that the cuing effect for unaware participants should be positive but no larger than the cuing effect observed for aware participants (as described in detail above). Thus a very plausible and principled alternative hypothesis for the size of cuing in unaware participants is one bounded by zero and the mean for the aware participants. We describe next the results of this BF analysis of our data.

Applying Bayes Factor analysis to our three experiments (Participant level)

Following the method suggested by Gallistel (2009), we used an alternative prior consisting of a uniform distribution between zero and the mean of the aware participants, convolved with the likelihood function of the cuing effect shown by the aware participants. This is illustrated in Fig. 3 for Experiment 2 and essentially assumes a positive relationship between cuing and recognition. According to the single-system model, if there is a relationship between cuing and recognition then it should clearly be a positive one. Therefore, as a further point of comparison, we also computed BFs for the opposite hypothesis that there is a negative relationship between cuing and recognition, that is, cuing for the unaware participants is *larger* than cuing for aware participants. We calculated a 'negative' alternative prior with the exact same size as the 'positive' prior but ranging from the mean cuing effect for the aware participants to double the mean cuing effect for the

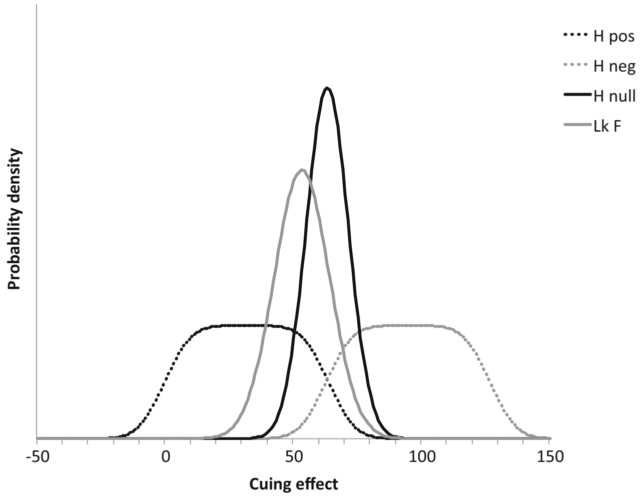

Fig. 3 An example of the hypotheses used in the Bayes Factor analyses reported for the participant level relationship between cuing and recognition (in this case from Experiment 2). H null reflects the normalised likelihood function derived from the aware participants (the null being that the unaware participants will display the same mean cuing effect). H pos reflects a hypothesis that cuing and recognition are positively related, thus the mean cuing effect for the unaware participants will fall somewhere between zero and the cuing effect observed for the aware participants. H neg reflects a hypothesis that cuing and recognition are negatively related, quantified to match H pos. Lk F is the (normalised) likelihood function of the mean cuing effect for the unaware participants

aware participants. We calculated BFs comparing each of these positive and negative alternative hypotheses against the null and also a third BF comparing them against each other. Here, the null hypothesis essentially reflects the predictions of the independence model (as does the null hypothesis in the preceding NHST analysis, but with the inherent limitations to NHST for testing null hypotheses as discussed above).

Bayes factors comparing a positive hypothesis against the null were in favour of the null for Experiments 1 and 3 (BF = 4.60 and 6.95 respectively) and relatively inconclusive for Experiment 2 (BF = 1.85 in favour of the null). The BFs comparing negative hypothesis against the null were in favour of the null for Experiments 2 and 3 (BF = 5.92 and 11.77 respectively) and were relatively inconclusive for Experiment 1 (BF = 1.50 in favour of the alternative). The BFs comparing positive and negative hypotheses were mixed, in favour of the negative for Experiment 1 (BF = 6.89), in favour of the positive for Experiment 2 (BF = 3.20) and inconclusive for Experiment 3 (BF = 1.69 in favour of the positive). The first two sets of BFs generally come down in favour of the null, that there is no relationship between cuing and recognition at the participant level. The point of these last BFs is to ask, if we assume that there *is* a relationship between cuing and recognition, is there clear evidence that this relationship is positive? Since the results across the experiments vary, the answer appears to be no. Thus the general conclusion from the participant level BF results is that a relationship between cuing and recognition is improbable and even if there is one, it is not necessarily a positive one.

Applying Bayes Factor analysis to our experiments (Configuration level)

However, recall that the evidence suggests it is quite plausible that participants only learn about a small subset of configurations – roughly 30% of the typical 8-12 repeated configurations. This means that data aggregated across all configurations may obscure a genuine relationship between cuing and recognition. We therefore applied a similar Bayes factor analysis for the configuration level analysis. For each experiment, we took the mean cuing effect for those configurations that received above chance recognition (as per Table 2) and calculated a prior based on a window from 0 to this number. This 'positive' prior reflects the hypothesis that the cuing for the above chance configurations will be larger than cuing for the below chance configurations by a difference that lies somewhere between zero (cuing score the same for both) and the actual size of the cuing effect for above chance configurations (cuing score effectively zero for below-chance configurations). Again, we calculated a 'negative' prior of the same magnitude reflecting the hypothesis that below chance configurations will actually display more cuing than the above chance configurations. We ran the same series of three BFs for each experiment comparing the positive hypothesis and negative hypothesis against the null and against each other. BFs comparing the positive and null hypotheses were in favour of the null for all experiments (BFs = 5.54, 6.66, 20.61 for Experiments 1-3 respectively). BFs comparing the negative and null hypotheses were in favour of the negative hypothesis, though weakly in Experiment 3 in particular (BFs = 5.51, 3.12, 2.96 in favour of the negative hypothesis for Experiments 1-3 respectively). The BFs comparing positive and negative hypotheses were strongly in favour of the negative for all three experiments (BFs = 30.55, 20.78, 61.04 for Experiments 1-3 respectively). These BFs strongly suggest that there is no positive relationship between cuing and recognition at the configuration level. If anything, they hint towards a negative relationship, although the evidence for this negative relationship is somewhat more equivocal than the evidence against a positive relationship.

Fitting the single-system model to the observed correlations

As noted earlier, the single-system approach is free to assume that, for a given procedure, the proportion of variance attributable to s (which we will refer to as var_s) can vary from 0 to 1. In our final analysis, we sought to determine which value of this parameter best captures the relationships that we observed in our three experiments, assuming the single-system model is accurate. To this end, we produced multiple simulations with the same N as each of the three experiments, varying the proportion of variance attributed to s in steps of .01 (i.e., 1% of variance) starting at 0%. We calculated correlations between cuing and awareness measures for both the participant- and configuration level analyses, tallied the frequency of observed correlations, and then compared these to the observed correlations from the three experiments. For the configuration level analysis, this was done using the 'all cues' model since the 'subset' model assumes a minimum of approximately 25% of variance is attributable to s, at which point null or negative correlations are already extremely unlikely. We then calculated the likelihood of the data given var_s using six experimental observations: the participant level correlations from Experiments 1-3 and the configuration level correlations from Experiments 1-3.

Computing this likelihood, it was clear that the data were most likely when $var_s = 0$. This is to be expected since one of the participant level correlations and all three of the configuration level correlations were negative. The model with $var_s = 0$ is a special case of the single-system model because it is indistinguishable from a full independence model. The assumption in this special case is that cuing and recognition share no variance at all, which is similar to the assumption of a full independence model. However, unlike the full independence model, the single s model with $var_s = 0$ assumes that the contribution of learning (s) to the dependent variables is essentially identical for every single configuration and for every single participant, with the observed variance instead attributed to other factors specific to the measure or unsystematic error. We argue that as an instance of a single-system model of learning, the var_s=0 case is neither useful nor plausible. This is because the var_s=0 case suggests that the contribution of learning to cuing is invariant for all configurations even though they differ in terms of the location of the target and also in terms of the spatial arrangement of distractors. Likewise this case assumes that variance in the participants' attention, motivation, and conditions of testing make no difference to learning on either measure. This seems highly unlikely. For these reasons, the distinction between the single-system model that attributes zero variance to s and one that predicts at least some shared variance is an important one. The model with the next lowest shared variance from s that we considered (and the one that provides the next best fit) assumed $var_s = 0.01$. A comparison of the likelihood of the data given $var_s = 0$ and $var_s = 0.01$ yields a BF = 9.16 in favour of $var_s = 0$. Nevertheless using $var_s = 0.01$ gives us the best fitting model that is actually distinguishable from using strictly independent sources of s, and so we illustrate the results from this version below in Fig. 4.

With this very small proportion of shared variance, the expected correlations between cuing and recognition are close to zero but still positive. Figure 4 plots the frequency distributions for simulations using $var_s = 0.01$, using the Ns of each experiment, and for both participant level and configuration level analyses. The observed correlations for each experiment are shown as symbols positioned on those distributions in the figure. The participant level correlations are reasonably consistent with what the model predicts using this very low level of shared variance. However, what should be clear is that even when this very small proportion of shared variance is assumed, the observed correlations between cuing and recognition at the configuration level are conspicuously more negative than would be expected from the model. For each experiment, we observed a configuration level correlation equal to or less than the experimental observation on less than 3% of simulated experiments, even when assuming $var_s = 0.01$.

It remains an open question whether a model assuming $var_s = 0.01$ is really more plausible than one assuming $var_s = 0$ as this model still assumes that s is close to being invariant across participants and configurations. Even if it were, this incarnation of the single-system model, which assumes 99% of the variance in learning is attributable to task-related factors that are independent of learning, fails to account for the negative correlations we observed across our three experiments. Further, the exercise suggests that it may be beneficial for future experiments on contextual cuing to provide ways of constraining model assumptions. For instance, it should be possible to vary the task parameters experimentally so that, in principle, variations in s are controlled and should be measurable on both cuing and awareness tasks.

Implications for the status of contextual cuing as nonconscious learning

Vadillo et al. (2016) make several persuasive arguments for why the body of literature on contextual cuing does not constitute strong evidence for learning without awareness. Our analysis, based on a simple extension of Vadillo et al.'s statistical assumptions as well the data from three large samples of participants, is consistent with several of their charges. For instance, our analysis also suggests that most studies are not sufficiently powered to assess whether participants possess explicit knowledge nor whether subsamples of participants above and below chance differ in the magnitude of the contextual cuing that they display. In relation to evidence of awareness overall and evidence of cuing amongst 'unaware' participants, both a single-system model and a model with two independent sources of learning predict that the probability of failing to reject the null decreases as sample size increases. However, that same single-system model also predicts that with a large enough N, the difference in the strength of cuing when comparing aware and unaware participants should be significant most of the time. Furthermore, although the predicted correlation between cuing and awareness is relatively modest even according to a single-system model, as sample size increases, this correlation should be consistently greater than zero. In contrast, a model that assumes complete

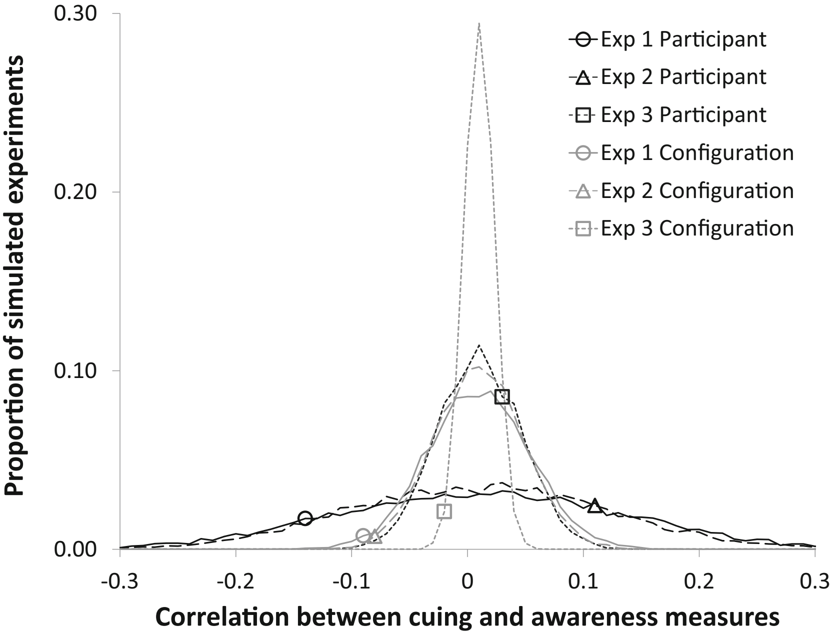

Fig. 4 Simulated distributions of participant-level and configuration level correlations between cuing and awareness when assuming that only 1% of variance on each measure is attributable to s. Symbols indicate the participant-level (circles) and configuration level (squares) correlations observed in Experiments 1-3 (placed vertically on the simulated distribution for ease of comparison)

statistical independence between the learning that serves recognition and the learning that serves contextual cuing does not make these last two predictions. Thus, contrary to Vadillo et al.'s argument, such analyses may still serve as important avenues to test the models. However, three caveats to this argument must be noted. First, such tests are only meaningful if researchers use much larger sample sizes than are conventionally relied upon on the contextual cuing literature. Second, because only a sub-set of configurations are likely to be learnt about (as little as 30% of the configurations), the analyses must include examination of the relationship between cuing and recognition at the configuration level. Third, the evidence from these tests needs to be assessed using a combination of analyses that should include Bayes Factors or another means of comparing the evidence in favour of each hypothesis rather than relying solely on traditional null hypothesis testing. On these points, we are again in complete agreement with Vadillo et al.

Applying these considerations to three contextual cuing experiments with large to very large samples, we found strong evidence of above chance performance on two types of recognition test (a 2AFC test in Experiments 1 and 3 and an old/new judgment test in Experiment 2). Although these effects were small in magnitude, their reliability is beyond reasonable doubt. It seems clear, then, that at least some participants become aware of a number of repeated configurations. Nevertheless, even with these large and very large samples, the correlations between cuing and recognition were extremely weak, and there was no reliable difference in the magnitude of cuing for participants who were above chance and those who were at or below chance. These two results are conspicuous specifically because the samples were large and the tests were clearly sensitive enough to find reliable evidence of explicit recognition. As such, our results favour a model in which there is substantial independence between the learning that supports contextual cuing and the learning that supports recognition, thereby providing evidence that this effect reflects nonconscious learning.

The most interesting result in the current study, however, came from the analysis of learning effects at the individual configuration level. If anything, our results suggest that better recognition of a repeated configuration was associated with weaker, not stronger contextual cuing for that configuration. This was evident in all three experiments, though statistically the effect was not particularly strong even with large samples. According to conventional tests, the relationship was marginal in all three experiments, with some falling just under and others just over the conventional threshold of statistical significance. Bayes factors support evidence in favour of a negative relationship over no relationship in all three experiments but it should be noted that the alternative (negative relationship) hypothesis we used was computed over a small expected effect size, based on theoretically principled limits as suggested by Gallistel (2009). In comparison with the modest evidence for a negative relationship over the null, BF tests in all three experiments indicated that either a null hypothesis or a negative relationship hypothesis were far more likely given the data than a positive relationship. To the extent that recognition does genuinely assess a conscious manifestation of learning about the configurations, this is an intriguing finding as it suggests that becoming aware of a particular cue does not facilitate and may even impede the contextual cuing effect.

Clearly, this pattern of results is very difficult to explain via any single-system account that assumes learning is mediated by conscious processes. Further, with its parameters free to vary, the best fit of our data was a single-system model with zero variance attributable to the common learning sources, rendering it indistinguishable from a full independence model. However, this is not to say that the results are easily explained by a dual systems account that assumes that some forms of nonconscious learning precede explicit learning and retrieval processes. Some dual theorists have proposed that awareness may arise as a result of automatic associative learning (see Mitchell et al., 2009 for a review), making an inverse relationship between the two difficult to explain even by this account. Similarly, whereas a full independence model, where the two systems are entirely separate, can obviously explain a lack of relationship between cuing and recognition, such a model would struggle to account for a reliable negative relationship. Of course, a negative relationship is much more likely to be observed under the full independence model than a single-system one. Nonetheless, the probability of observing such a relationship is still reasonably small.

A number of possibilities might explain this inverse relationship. Poorer learning in the visual search task for a particular configuration could provide additional opportunity to study that configuration, thereby increasing the probability of recalling it on test. This explanation assumes that participants spend more time and effort searching for the target on the repeated trials that are not learned as effectively and that the initial learning is divorced from the retrieval process that mediates recognition. Another possibility is that recognition of a repeated pattern during the visual search task slows down performance because additional mental processes are engaged. This explanation adequately accounts for the results but must assume that the cuing effect is not dependent on explicit memory.

As a third alternative, the negative statistical association between cuing and recognition might be based on a less direct causal relationship. For instance, some additional factor that varies across the repeated configurations might systematically facilitate recognition but also impede cuing (or vice versa). This could create systematic and task-specific sources of variance that are negatively correlated. This explanation does not necessarily require any independence in the learning and retrieval processes mediating cuing and recognition. For example, a distinctive feature of a particular repeated configuration could facilitate explicit recognition and at the same time impair a cuing effect, if it diverts attention away from the target location. Of some relevance to this possibility, Conci and von Muhlenen (2011) found that segmenting search displays based on colour or stimulus size can impair the cuing effect. However, we expect that it is unlikely that these sources of variance would be so strong not only to conceal but also to reverse an underlying positive relationship between the two. That is, given the current pattern of results involving a negative statistical association, we propose that it is far more plausible that there is no relationship between cuing and recognition in this task compared with the possibility of a latent positive relationship.

Neuroimaging studies may be able to shed some light on the viability of these and other accounts. For example, Geyer, Baumgartner, Müller, Pollmann (2012) recently found that BOLD activation in the medial temporal lobe was reduced on configurations classified as unaware and enhanced on those classified as aware and this effect only became evident after the cuing effect had been established, which they interpreted as indicative of differences in retrieval as opposed to learning. Furthermore, it is worth emphasizing that we specifically examined the relationship between cuing and recognition for the repeated configurations as they were presented during the cuing task. Performance on other measures of explicit knowledge, specifically generation tests, may be based on different knowledge and may hold a different relationship with the cuing effect. Generation tasks test for knowledge of the relationship between the pattern of distractors and the spatial location of the target predicted by that pattern. These tasks also necessarily involve a small, but potentially significant disruption of the configuration, because the target must be either replaced by a distractor or omitted in order for the participant to be able to guess its location. In contrast, the recognition tests we used involved presenting the configurations being presented in exactly the same visual layout as in cuing and thus should be a relatively sensitive test.

Conclusions

This current theoretical and empirical analysis suggests that when large samples are used in conjunction with Bayesian analysis at the configuration level, there is strong evidence against a positive relationship between cuing and recognition. In our view, this provides convincing evidence that contextual cuing reflects nonconscious learning. Importantly, this conclusion is in no way inconsistent with Vadillo et al.'s (2016) arguments. Instead, their arguments provided a clear set of situations in which such evidence could be found, which we have extended here. Indeed, the theoretical considerations highlighted by them and developed here provide the framework for ways in which future empirical studies that seek to examine whether a given learning phenomenon results from nonconscious learning.

Acknowledgments Ben Colagiuri was a recipient of a University of New South Wales Vice-Chancellor's Postdoctoral Research Fellowship (RG104440-1) while conducting the first experiment in this report. This research was supported by Discovery Grant DP120102190 from the Australian Research Council.

References

Berry, C. J., Shanks, D. R., & Henson, R. N. (2008a). A single-system account of the relationship between priming, recognition, and fluency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 97–111. doi:10.1037/0278-7393.34.1.97

Berry, C. J., Shanks, D. R., & Henson, R. N. (2008b). A unitary signal-detection model of implicit and explicit memory. *Trends in Cognitive Science, 12,* 367–373. doi:10.1016/j.tics.2008.06.005

Berry, C. J., Shanks, D. R., Speekenbrink, M., & Henson, R. N. (2012). Models of recognition, repetition priming, and fluency: exploring a new framework. *Psychological Review, 199,* 40–79. doi:10.1037/a0025464

Chun, M. M., & Jiang, Y. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology, 36,* 28–71. doi:10.1006/cogp.1998.0681

Chun, M. M., & Jiang, Y. (2003). Implicit, long-term spatial contextual memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 224–234.

Chun, M. M., & Phelps, E. A. (1999). Memory deficits for implicit contextual information in amnesic subjects with hippocampal damage. *Nature Neuroscience, 2,* 844–847. doi:10.1038/12222

Colagiuri, B., Livesey, E. J., & Harris, J. H. (2011). Can expectancies produce placebo effects for implicit learning? *Psychonomic Bulletin and Review, 18,* 399–405.

Conci, M., & Muhlenen, A. (2011). Limitations of perceptual segmentation on contextual cueing in visual search. *Visual Cognition, 19,* 203–233.

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review, 116,* 439–453.

Goujon, A., Didierjean, A., & Thorpe, S. (2015). Investigating implicit statistical learning mechanisms through contextual cueing. *Trends in Cognitive Sciences, 19,* 524–533.

Geyer, T, Baumgartner, F., Müller, H. J., & Pollmann, S. (2012). Medial temporal lobe-dependent repetition suppression and enhancement due to implicit vs. explicit processing of individual repeated search displays. *Frontiers in Human Neuroscience,6:* article 272. doi:10.3389/fnhum.2012.00272

Howard, J. H., Jr., Howard, D. V., Dennis, N. A., Yankovich, H., & Vaidya, C. J. (2004). Implicit Spatial Contextual Learning in Healthy Aging. *Neuropsychology, 18,* 124–134.

Jimenez, L., & Vazquez, G. A. (2011). Implicit Sequence Learning and Contextual Cueing Do Not Compete for Central Cognitive Resources. *Journal of Experimental Psychology: Human Perception and Performance, 37,* 222–235.

Lovibond, P. F., & Shanks, D. R. (2002). The role of awareness in Pavlovian conditioning: empirical evidence and theoretical implications. *Journal of Experimental Psychology: Animal Behavior Processes, 28,* 3–26.

Manns, J. R., & Squire, L. R. (2001). Perceptual learning, awareness, and the hippocampus. *Hippocampus, 11,* 776–782. doi:10.1002/hipo.1093

Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning. *Behavioral and Brain Sciences, 32,* 183–198. doi:10.1017/S0140525X09000855

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5,* 241–301. doi:10.1037/1082-989X.5.2.241

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. doi:10.1093/cercor/bhm245

Rausei, V., Makovski, T., & Jiang, Y. V. (2007). Attention dependency in implicit learning of repeated search context. *The Quarterly Journal of Experimental Psychology, 60,* 1321–1328. doi:10.1080/17470210701515744

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t-Tests for Accepting and Rejecting the Null Hypothesis. *Psychonomic Bulletin & Review, 16,* 225–237. doi:10.3758/PBR.16.2.225

Shanks, D., & St John, M. F. (1994). Characteristics of dissociable human learning systems. *Behavioural and Brain Sciences, 17,* 367–447.

Shanks, D., & Berry. (2012). Are there multiple memory systems? Tests of models of implicit and explicit memory. *Quarterly Journal of Experimental Psychology, 65,* 1449–1474.

Smyth, A., & Shanks, D. (2008). Awareness in contextual cuing with extended and concurrent explicit tests. *Memory and Cognition, 36*(2), 403–415. doi:10.3758/mc.36.2.403

Vadillo, M. A., Konstantinidis, E., & Shanks, D. R. (2016). Underpowered samples, false negatives, and unconscious learning. *Psychonomic Bulletin & Review, 23,* 87–102.

Vaidya, C. J., Huger, M., Howard, D. V., & Howard, J. H., Jr. (2007). Developmental Differences in Implicit Learning of Spatial Context. *Neuropsychology, 21*(4), 497–506.