Running Head: Probabilistic cuing of visual attention

Revision of XGE-2018-1004R1 as invited by the action editor, Nelson Cowan

Unconscious or underpowered? Probabilistic cuing of visual attention

Miguel A. Vadillo[1], Douglas Linssen[2], Cristina Orgaz[3],
Stephanie Parsons[2], & David R. Shanks[2]

[1]*Departamento de Psicología Básica, Universidad Autónoma de Madrid, Spain*
[2]*Department of Experimental Psychology, University College London, United Kingdom*
[3]*Departamento de Psicología Básica, Universidad Nacional de Educación a Distancia, Madrid, Spain*

Word count: 12,501 (Abstract + Main text)

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

Abstract

Recent debate about the reliability of psychological research has raised concerns about the prevalence of false positives in our discipline. However, false negatives can be just as concerning in areas of research that depend on finding support for the absence of an effect. This risk is particularly high in unconscious learning experiments, where researchers commonly seek to demonstrate that people can learn to perform a task in the absence of any explicit knowledge of the information that drives performance. The fact that some unconscious learning effects are typically studied with small samples and unreliable awareness measures makes false negatives especially likely. In the present article we focus on a popular unconscious learning paradigm, probabilistic cuing of visual attention, as a case study. Firstly, we show that, at the meta-analytic level, previous experiments reveal positive signs of participant awareness, although individual studies are severely underpowered to detect this. Secondly, we report the results of two empirical studies in which participants' awareness was tested with alternative and more sensitive dependent measures, both of which manifest positive evidence of awareness. We also show that, based on the predictions of a formal model of probabilistic cuing and given the reliabilities of the dependent measures collected in these experiments, any statistical test aimed at detecting a significant correlation between learning and awareness is doomed to return a non-significant result, even if at the latent level both constructs are actually related and participants' knowledge is completely explicit.

Keywords: computational modeling; false negatives; implicit learning; meta-analysis; probabilistic cuing; unconscious learning.

In the few years that have passed between the "crisis of confidence" in psychological research (Pashler & Wagenmakers, 2012) and "psychology's renaissance" (Nelson, Simmons, & Simonsohn, 2018) we have learned that many of the scientific findings that we took for granted might actually reflect false positives, entirely attributable to the selective publication of significant findings, questionable research practices, and a poor understanding of statistical methods (Franco, Malhotra, & Simonovits, 2014; Kerr, 1998; Simmons, Nelson, & Simonsohn, 2011). What this debate has largely overlooked, though, is that researchers are not always interested in establishing the reliability of a positive effect. In many areas of psychology, key findings depend on reporting the *absence* of an effect (Rouder, Speckman, Sun, Morey, & Iverson, 2009).

For instance, researchers might be interested in showing that the different groups of an experiment do not differ in a crucial confounding variable or in their pre-test scores (e.g., Hilgard, Engelhardt, Bartholow, & Rouder, 2017), or that suicide rates associated with a drug treatment for depression are no higher than in a placebo control group (Fergusson et al., 2006). Finding out whether these null results are true negatives is just as important as assessing the reliability of positive findings (Hartgerink, Wicherts, & van Assen, 2017), perhaps even more so, given that Null Hypothesis Significance Testing (NHST), as regularly implemented in psychological research, is poorly suited to assess the plausibility of the null hypothesis (Dienes, 2011, 2015; Hoekstra, Finch, Kiers, & Johnson, 2016). Furthermore, given the average low power of psychological research (Sedlmeier & Gigerenzer, 1989; Smaldino & McElreath, 2016), finding a null result is hardly surprising and often uninformative as to the veracity of the null hypothesis.

The present study addresses the scale of this problem in an area of research that relies extensively on garnering support for the null hypothesis, namely, implicit (or unconscious) learning (Berry & Dienes, 1993; Greenwald & De Houwer, 2017; Shanks, 2005). Although

different implicit learning paradigms are based on a range of different methods, a common strategy to show that a learning effect is unconscious consists in finding positive evidence of learning in a behavioural task together with the absence of evidence of awareness of the stimuli or the statistical regularities that drive performance (Dienes, 2015). For instance, in a classical implicit learning effect known as contextual cuing (Chun & Jiang, 1998; Chun & Turk-Browne, 2008), participants are presented with a series of visual search displays where they have to find a rotated T among a series of L-shaped distractors. Some of the search displays are presented repeatedly across the experiment while others are random arrangements of distractors that occur just once (participants are not told about this manipulation). Search times usually reveal that participants learn something about the repeating patterns, as they find the target much more easily in repeated than in random search displays. However, when directly asked to identify the repeating patterns, participants' performance, it is claimed, does not differ significantly from chance. This dissociation between significant evidence of learning and non-significant evidence of explicit recognition is interpreted as strong support for unconscious learning (Chun & Jiang, 2003).

In a recent meta-analysis of the literature on contextual cuing (Vadillo, Konstantinidis, & Shanks, 2016), we showed that, in truth, participants' performance in these awareness tests is actually above chance on average, although the effect is too small (Cohen's $d_z = 0.31$) to be detected with the typical samples tested in most of these experiments (usually, $N = 16$). Furthermore, there are good reasons to suspect that the small effect size of awareness is related to methodological limitations of the measurement procedure rather than to an intrinsically weak degree of awareness. In our review, studies that assessed awareness using longer tests with many trials were more likely to find significant evidence of awareness (see also Smyth & Shanks, 2008). In other words,

attributing contextual cuing to unconscious learning processes might be the product of a statistical illusion created by the use of poor measures of awareness in combination with underpowered samples.

As explained above, reliance on NHST to establish the absence of awareness is not exclusive to contextual cuing studies and there are good reasons to think that unreliable measures and underpowered samples may affect other implicit cognition paradigms in the same manner. Chance-level performance in a recognition or discrimination test is a typical criterion for unconsciousness in other implicit learning and unconscious perception paradigms (e.g., Hedger, Gray, Garner, & Adams, 2016; Reed & Johnson, 1994; Shang, Fu, Dienes, Shao, & Fu, 2013) and our meta-analysis on contextual cueing is not the only one to suggest that studies relying on this strategy tend to be underpowered (e.g., Hedger et al., 2016). The goal of the present article is to explore this problem in another implicit learning task, probabilistic cuing of visual attention (Druker & Anderson, 2010; Geng & Behrmann, 2002; Jiang, 2018), which is in many ways closely related to contextual cuing.

In a typical probabilistic cuing experiment, participants are instructed to perform a visual search task, superficially identical to the one used in contextual cuing studies. Participants are asked to find a visual target among a number of distractors in a series of search displays. Unlike in contextual cuing experiments, however, the key experimental manipulation is not that some displays are repeated, but that the target tends to appear more frequently in a specific region of the search display (usually a quadrant, such as top left) than in the other regions. Across trials, participants become faster at finding the target when it appears in the ‘rich’ region than in any of the ‘sparse’ regions. However, this learning effect is considered implicit because, after the experiment, participants do not report any awareness of this spatial regularity and, when asked to guess which region contained the target most frequently, the total number of participants selecting the rich

region is not significantly above chance (e.g., Jiang, Capistrano, Esler, & Swallow, 2013; Jiang, Swallow, Rosenbaum, & Herzig 2013).

As in the case of contextual cuing, this analytic strategy is statistically misleading: The fact that the number of participants selecting the rich region is not significantly different from chance does not mean that it is genuinely at chance, particularly if this inference is based on a small number of participants and few observations per participant. In this sense, the standard awareness test used in probabilistic cuing experiments is even more problematic than the comparable test in contextual cuing experiments. While in the latter awareness is usually assessed by means of an explicit recognition test comprising several trials (typically, around 24), in probabilistic cuing awareness is assessed with a single-item test: Participants are invited just once to select the region that they think contained the target most often, guessing if necessary. Extrapolating from the results of our meta-analysis in the domain of contextual cuing, we think there are good reasons to suspect that false negatives are particularly likely under these conditions.

Other peculiarities of probabilistic cuing experiments could also undermine the informativeness of awareness tests conducted at the end of the experiment. It is well known that visual search is influenced by target location repetitions. When the target appears in two spatially close locations on consecutive trials, reaction times are usually faster on the second trial, an effect known as repetition priming (Maljkovic & Nakayama, 1996). This effect is particularly problematic in probabilistic cuing experiments because, unless careful measures are taken, repetition priming and location probability will naturally be confounded with each other: If the target appears more frequently in one region than in the others, then there will also be more repetitions in the former than in the latter (Walthew & Gilchrist, 2006).

The confounding contribution of repetition priming can be cancelled out statistically. For instance, researchers can show that probabilistic cuing is observed independently in trials with and without repetitions (Jiang & Swallow, 2013). An alternative solution is to use a two-stage design in which participants are first trained in the visual search task with the targets appearing more frequently in one region and then tested in a subsequent stage in which the targets appear with the same probability in all regions (Jiang, Capistrano et al., 2013). If reaction times are faster for the rich region even in the unbiased testing stage, this cannot be attributed to repetition priming and, consequently, it must be attributed to a learned attentional bias developed over the initial training stage.

Although this approach is methodologically sound when it comes to establishing the reliability of probabilistic cuing, it can negatively affect the results of any awareness test conducted at the end of the experiment. If participants are first exposed to a training stage and then to an unbiased testing stage, it is likely that the attentional biases learned during the former will become diluted or vanish entirely across the latter (e.g., Jiang, Capistrano et al., 2013; Jiang & Won, 2015, Experiment 1). When participants are then asked to select the region that contained the target most frequently, it is natural that some of them will fail to select the correct response, either through confusion or through unlearning. After all, the target was equally likely to appear anywhere during the immediately preceding stage. In more general terms, this approach fails to meet what Newell and Shanks (2014) called the 'immediacy criterion': A good measure of awareness should take place concurrently to the learning task or as soon as possible afterwards to avoid forgetting, interference, or, in this case, unlearning.

Another potential problem in probabilistic cuing experiments is that what participants learn during the training stage might not overlap perfectly with the experimental manipulation that is addressed in the awareness test (Shanks & St. John,

1994). Imagine a cuing experiment in which the target is more likely to appear in the top-right quadrant than in the remaining quadrants. A participant might fail to detect this regularity in full detail, but she might notice that the target appears more frequently on the right-hand side of the screen than on the left-hand side. Knowing this would be sufficient to speed up visual search in the rich region, but would not guarantee good performance in the awareness test, as according to her hypothesis the top-right and the bottom-right quadrants are equally valid responses.

In the present study we explore these problems in more detail using a combination of meta-analytic methods, empirical studies, and computational modeling. First, in the following section we report the results of a systematic review and meta-analysis exploring the awareness tests of published probabilitistic cuing studies. Then, we report the results of two new empirical studies where we test the sensitivity of alternative measures of awareness that address some of the shortcomings identified in the previous paragraphs. We also show that, based on the predictions of a formal model of probabilistic cuing and given the reliabilities of the dependent measures collected in these experiments, any statistical test aimed at detecting a significant correlation between learning and awareness is doomed to return a non-significant result, even if at the latent level both constructs are actually related and participants' knowledge is completely explicit.

Meta-analysis of Previous Probabilistic Cuing Studies

Individual studies with small sample sizes are poorly suited to make inferences about the absence of an effect because non-significant results can be easily attributed to a simple lack of statistical power. Meta-analysis, in contrast, allows researchers to achieve a high degree of statistical power by collating data from large numbers of participants tested in different studies. If the null results typically observed in the awareness tests of probabilistic

cuing experiments reflect true negatives, then a meta-analytic integration of all the experiments published so far should provide a highly precise effect size estimate that would not depart significantly from zero. Additionally, meta-analytic methods can also be used to assess the potential impact of some of the methodological artefacts noted in the previous paragraphs. For instance, it is possible that interpolating an unbiased testing stage between training and the awareness check produces some amount of unlearning, reducing the likelihood of observing significant levels of awareness. If so, one would expect to find greater awareness effects in experiments that did not include such a testing stage than in experiments that did include it. Similarly, awareness should be greater in studies for which there is no positive evidence of unlearning than in studies where the statistical analyses suggest that unlearning occurred. The present meta-analysis puts these hypotheses to the test.

Method

Literature search

Y. Jiang is the most prolific and active expert in probability cuing of visual attention. Consequently, we decided to use her work as the starting point of our literature search strategy. On November 10th 2017, we explored the website of her research team at the University of Minnesota and the Web of Science to find all her published research on probability cuing. On the basis of the titles and abstracts we selected 18 studies by Jiang and colleagues that were assessed for inclusion in the meta-analysis. Then, we searched for additional papers by inspecting the reference sections of these articles and also by searching for papers citing them in the Web of Science. This led to 27 additional articles that were assessed for inclusion.

Eligibility criteria

Studies were only included in the meta-analysis if they complied with four selection criteria. Firstly, we selected only studies exploring probabilistic cuing of visual attention (Criterion 1). By this, we understood any task in which participants had to locate a visual target in a search display, with the targets being more likely to appear in one specific region than in the others. Secondly, studies were only included in the meta-analysis if they administered an awareness test (Criterion 2). Furthermore, our meta-analysis focuses only on the most popular type of awareness test, in which at the end of the experiment each participant is asked to guess which region contained the target most often and above chance performance at the group level is taken as significant evidence of awareness. Studies including only other types of awareness tests, such as unstructured interviews or subjective ratings, were excluded.

Thirdly, we excluded studies in which participants were given some explicit instruction or presented with exogenous cues that would orient their attention, either hindering or promoting attention to the rich region (Criterion 3). For instance, experiments in which participants were explicitly instructed to pay attention to one quadrant or were presented with arrows directing their attention towards specific quadrants were excluded from the meta-analysis. Finally, studies were included in the meta-analysis only if participants showed statistically significant evidence of probabilistic cuing at some point in the experiment (Criterion 4). Table A1 in Appendix A lists all the studies that were excluded according to these criteria. The studies that passed all the eligibility criteria are reported in Table 1 and marked with an asterisk in the reference list.

Computation of effect sizes

In all the studies included in this meta-analysis the main dependent variable was the proportion of participants selecting the rich region in the awareness test. If this proportion

was higher than chance, then participants were assumed to be aware of the biased distribution of targets in the search displays. A simple and straightforward means to meta-analyze these data would be to collate these proportions. Unfortunately, proportions on their own carry little informative value, because the definition of chance-level performance differed from one experiment to another. While in most studies the search display was divided into four regions, one of them containing the target more frequently than the other three, in an important subset of studies the search display was divided into just two regions, one of them acting as the rich region and the other as the sparse region. Consequently, chance-level performance was .25 in some experiments, but .50 in others. Bearing this in mind, simply knowing that the proportion of participants guessing the correct region was, say, .35, does not provide sufficient information to assess whether participants were aware or not. And, for the same reason, raw proportions cannot be collated in a single meta-analysis for studies with different numbers of regions.

To overcome this problem, we conducted all analyses using Cohen's *h* as the default effect size estimate (Cohen, 1977, Hedger et al., 2016). Cohen's *h* measures the difference between two proportions, in our case, the proportion of participants selecting the rich region at test (p_1) and the proportion that would be expected by chance (p_2). Cohen's *h* is computed as $2 \cdot \arcsin\sqrt{p_1} - 2 \cdot \arcsin\sqrt{p_2}$ and its value can be interpreted following the same standards as the popular Cohen's *d*: values of .20, .50, and .80 are considered small, medium and large, respectively. For comparisons of a one-sample proportion against a theoretical proportion, the variance of Cohen's *h* is simply the inverse of the number of participants.

In a small number of studies, not all participants were invited to guess which region contained the target most frequently. In some studies (e.g., Smith, Hood, & Gilchrist, 2010) only participants who thought that the targets had been randomly distributed were

invited to guess which was the rich region. In contrast, in other studies (e.g., Umemoto, Scolari, Vogel, & Awh, 2010) only participants who thought that the target was not randomly distributed were invited to select a region. In yet other studies (Jiang, Swallow, & Rosenbaum, 2013) some of the sample were invited to select a region unconditionally and others were excluded from this test if they thought that the target was equally likely to appear anywhere in the search display. In all these cases, we computed the effect size on the basis of the proportion of participants who correctly selected the rich region out of the total number of participants who were invited to make this selection. That is to say, in our analyses the sample size is the total number of participants who took part in the region-guessing test, which is not always the same as the total number of participants included in the experiment.

Similarly, in a small number of studies there was some ambiguity regarding which region should be considered the rich one. For instance, in Jiang, Swallow, and Sun (2014) the search display lay on a table and participants moved to a different side of the table between training and testing (or the display was rotated between one stage and the other). In this situation, at test one can distinguish a viewer-rich region and a scene-rich region, depending on whether the reference frame is centered in the viewer or in the scene. In cases like this, we considered that the correct response was selecting the region for which there was positive evidence of learning. For example, in Experiment 1 of Jiang, Swallow et al. (2014), participants showed probabilistic cuing only for the viewer-rich quadrant, but not for the scene-rich quadrant. Consequently, we considered that selecting the viewer-rich quadrant was the correct response in the awareness test.

Coding of study characteristics

As explained above, before starting the literature search, we suspected that two features of the studies could have an important impact on effect sizes. Firstly, we hypothesized that studies that tested probabilistic cuing with an unbiased stage after training would give rise to some amount of unlearning or "extinction" of any learned search bias that would reduce the chances of observing above-chance performance in the subsequent awareness test. Therefore, for each study we coded whether or not the design included an unbiased testing stage before the awareness test. For similar reasons, we hypothesized that the results of the awareness test would only be above chance if there was positive evidence of learning immediately before the awareness test. That is, even if the magnitude of cuing had been significant at some stage of the experiment, if there was evidence that probabilistic cuing had vanished before the awareness test (e.g., because of unlearning in an unbiased testing stage), we expected to find chance-level performance in the awareness test. Consequently, evidence of learning at the end of the experiment was coded as a potential moderator of effect sizes.

In addition to these two moderators, while we were assessing the articles for inclusion we detected important methodological differences across studies that we thought deserved to be explored as potential moderators. For instance, while in most studies the rich region was one of the quadrants of the search display, in other studies the rich region was one half of the search display. We hypothesized that the biased distribution of targets would be easier to notice in the latter and, consequently, we coded this feature as a potential moderator. Similarly, although most studies were conducted using a computer screen or other electronic device to present the search displays, a small number of experiments were conducted in large-scale settings, such as a park or a room. Research in other implicit learning paradigms has shown that learning tends to be more robust when search displays comprise natural scenes than when employing artificial stimuli such as

rotated Ls and Ts (Brockmole & Henderson, 2006; Brockmole & Vo, 2010). Consequently, we hypothesized that the evidence of awareness would be larger in studies conducted in large-scale settings. Finally, as noted above, not all participants were invited to select a region in all the experiments. We were unsure about the potential impact of this feature on the results of the awareness test. For exploratory purposes, we decided to code whether all participants or just a subset of them were invited to select a region in the awareness test and we assessed the impact of this variable on effect sizes.

Results and Discussion

The effect sizes and variances for all the 44 studies that met the criteria for inclusion in the meta-analysis are shown in Table 1. All the analyses reported in this section were conducted with the 'metafor' R package (Viechtbauer, 2010). Across studies, the random-effects meta-analytic effect size was 0.35, with 95% confidence interval (CI) from 0.21 to 0.49, $z = 4.81$, $p < .001$, providing strong evidence that participants were able to identify the rich region in the awareness test. The meta-analysis also revealed a substantial and statistically significant amount of heterogeneity across studies, $I^2 = 71.24\%$, $Q(43) = 143.11$, $p < .001$.

To assess whether this heterogeneity could be accounted for by the different characteristics that we coded for each study, we conducted five independent meta-regressions. The results are shown in Table 2. As can be seen, the only moderator that explained a significant proportion of heterogeneity was the experimental setting: Although awareness was statistically significant both when the search was conducted on a computer/tablet screen and when it was conducted in a large-scale setting (i.e., a park or a room), effect sizes tended to be noticeably larger in the latter case.

None of the other potential moderators reached statistical significance. Even so, it is interesting to note that in all cases the ordering of the means is consistent with our predictions. For instance, on average, effect sizes tended to be larger when the awareness test was conducted immediately after the training stage than when an unbiased testing stage was interpolated between them, suggesting that the lack of awareness observed in some experiments might be due to some degree of unlearning taking place during the unbiased testing stage or confusion in some participants about which stage the experimenter is probing them about. Similarly, effect sizes tended to be larger in the subset of studies in which there was clear evidence of learning immediately before the awareness test than in the subset in which learning seemed to have diminished before the awareness test. In fact, only the former subset of studies produced statistically significant evidence of awareness.

In any case, given that these moderator analyses failed to reach statistical significance, the value of these patterns of results can only be considered suggestive. The fact that some subsets of studies include only 4-6 effect sizes limits considerably the power of the moderation tests and our ability to draw any firm conclusions. An additional shortcoming of these analyses is that the moderators themselves are correlated with each other. In Table 3 we report the results of a series of χ^2 tests exploring the relationships between moderators. As can be seen, across studies there is a significant correlation between whether or not they included an unbiased testing stage before the awareness test and (1) the experimental setting of the study (computer vs. large-scale) and (2) the number of regions into which the search display was divided.

Although 4 of the moderator analyses failed to confirm our predictions, our results do show quite convincingly that at the meta-analytic level participants' performance in the awareness tests is significantly above chance. For a typical experiment in which

participants have to guess the rich quadrant out of 4 candidate regions, an effect size of h = 0.35 is equivalent to a proportion of .42 participants correctly selecting the rich quadrant. As in our previous analysis of contextual cuing (Vadillo et al., 2016), we conclude that the failure to detect significant evidence of awareness in many of these studies, taken in isolation, must be attributed to a simple lack of statistical power. In fact, a power analysis with the 'pwr' R package shows that with the median sample size of these studies (N = 16), the statistical power to detect an effect of size h = 0.35 in a two-tailed test is barely .29. Stated differently, a study requires 64 participants in order to reach .80 power to detect a significant awareness effect in a two-tailed test or 51 participants for an equivalent one-tailed test, far greater than the typical sample size.

It is worth noting that the results of the present meta-analysis are in perfect agreement with a recent reanalysis of data from more than 300 participants who have taken part in many of the studies conducted at Jiang's laboratory over the last seven years (Jiang, Sha, & Sisk, 2018). Across these studies, the proportion of participants correctly guessing the rich quadrant in the awareness test was exactly .42, which was also significantly higher than chance, $\chi^2(1) = 53.40$, $p < .001$.

Relationship between Measures of Performance and Awareness

We have focused to this point on the analysis of the proportion of participants selecting the rich region in the awareness test because finding at-chance performance on this variable is, by far, the most common basis in past studies for inferring unconscious probabilistic cuing. Table S1 in the Supplementary Material reports the verbatim quotations from each article where the authors justify their reasons for concluding that learning was unconscious. As can be seen, failure to perform above chance in the awareness test at the group level was the main argument in most cases. However, not all

the studies included in the meta-analysis relied just on this line of reasoning. For instance, researchers sometimes compared the size of probabilistic cuing in participants who selected the rich region and participants who did not (e.g., Salovich, Remington, & Jiang, 2018; Twedell, Koutstaal, & Jiang, 2017; see also Jiang, Sha et al., 2018). The logic behind this analysis is that, if learning is based on explicit memory, then one would expect cuing to be larger among participants who show evidence of awareness. However, a moment's thought reveals that the problems of statistical power that we have detected for the more typical analysis become even worse when this alternative approach is adopted.

It is useful to consider the issue from a psychometric perspective. Testing for an interaction between awareness and cuing is essentially an attempt to detect a correlation between two variables and, as such, is affected by measurement error. Even if two variables, *x* and *y*, measure exactly the same latent construct, if their reliabilities, r_{xx} and r_{yy}, are less than perfect, the observed correlation between them will usually be lower than 1 (Spearman, 1904). Specifically, the attenuation factor is given by $\sqrt{r_{xx}} \cdot \sqrt{r_{yy}}$. How small can the observed correlations become for the particular case of probabilistic cuing and awareness?

At present, we simply do not have sufficient information to provide even a tentative answer to this question. It is well known that most experimental tasks provide remarkably unreliable measures, especially when participants' performance is measured on the basis of differences in reaction times under two different conditions (Enkavi et al., 2019; Hedge, Powell, & Sumner, 2018). To the best of our knowledge, no previous study in this literature has reported the reliability of probabilistic cuing or of the awareness tests employed. However, the articles that have done so in other implicit learning paradigms have reported disappointingly low values. For instance, West, Vadillo, Shanks, and Hulme (2018) tested 7-8 year-old children in several implicit learning tasks, including different

versions of contextual cuing, serial reaction time, and a Hebb serial order learning task. Most of the reliabilities of these tasks were below .50. In the same vein, Kaufman, DeYoung, Gray, Jiménez, Brown, and Mackintosh (2010) found a reliability of .44 for the serial reaction time task. Regarding the measurement of awareness, Smyth and Shanks (2008) reported a reliability of .46 for an extended version of a test typically used to measure explicit knowledge in contextual cuing. For the standard (shorter) test, the reliability was just .09.

Although at present we ignore the reliability of the different measures of performance and awareness that have been used to explore probabilistic cuing, it is easy to see that if these values are not substantially higher than those reported for similar implicit learning paradigms, then any attempt to detect a correlation between cuing and awareness is doomed to fail even in very large samples. If these reliabilities are, for instance, in the order of .40, then even if there is a medium-sized correlation of .30 between learning and awareness at the latent level, the observed correlation will drop to just .12. With this effect size, 725 participants are needed to detect a significant correlation with 90% power.

Instead of exploring the correlation between learning and awareness, some studies have concluded that probabilistic cuing is unconscious because it could be detected in the subsample of participants who did not select the rich region in the awareness test (e.g., Addleman, Tao, Remington, & Jiang, 2018; see also Jiang, Sha et al., 2018). The psychometric stance outlined in the previous paragraphs also highlights the shortcomings of this approach. Because the awareness test, like any other psychological variable, is subject to measurement error, the set of participants who fail the awareness test will necessarily include participants who were actually aware (see Shanks, 2017b). And these participants may exhibit a significant cuing effect due to entirely conscious processes. The

scale of the problem depends on the reliability of both dependent variables, which, as explained above is as yet unknown for probabilistic cuing.

Although none of the following experiments were originally designed with this purpose in mind, the resulting data and modeling can be used to illuminate the question of whether probabilistic cuing and explicit knowledge can be expected to correlate with each other and to what extent probabilistic cuing can be expected to occur even among participants who apparently lack any explicit knowledge.

Experiment 1

While the previous meta-analysis shows that participants' performance in the awareness test is clearly above chance on average, we cannot deny that these effects are small by statistical standards. The mismatch between the large effects observed in search times and the small effects in awareness measures may be seen, per se, as compelling evidence of unconscious learning. However, an alternative explanation is that the method used to measure awareness may not be as sensitive as the method used to measure probabilistic cuing. While the former is usually measured with a single item, the latter is typically measured on the basis of hundreds and hundreds of visual search trials. This factor alone could suffice to explain the differences in effect sizes.

Furthermore, as we argued in the introduction, it is also possible that the awareness test fails to capture awareness of the specific information that participants have actually learned during the training stage. Over the course of the experiment, participants may develop all kinds of hypotheses about the location of the targets. Some of those hypotheses may depart substantially from the experimental manipulation that the researcher has in mind, but even so they may still speed up visual search in the rich region: in Dulany's (1961) terms, they are 'correlated hypotheses'. A good example of this would be a

participant believing that the target appears more frequently on the right side of the screen when, in truth, it appears more frequently specifically in the top-right quadrant. The type of awareness tests included in the meta-analysis are almost guaranteed to yield small effect sizes in they ignore this possibility.

In the present study, we explored the sensitivity of an alternative test in which participants were asked to rank the four quadrants according to their likelihood of being the rich region. Specifically, participants were first asked to guess which quadrant was most likely to be the rich region. This response, on its own, is comparable to the dependent measure taken in previous studies. But, in addition to this, participants were then also asked to guess which was the second most likely quadrant to be the rich region, and then the third quadrant. Even if some participants failed to rank the rich quadrant first, we hypothesized that they would still be more likely to rank it second than third or fourth, which would confirm that even participants who fail the traditional test show some degree of awareness. Following this logic, we expected that analysing the whole 4-quadrant ranking provided by each participant would provide clear evidence of awareness, even if the simple proportion of participants ranking the rich quadrant first failed to be significantly above chance.

Method

Participants

Thirty-two participants from the UCL participant pool took part in Experiment 1 in exchange of £4. The meta-analysis presented in the previous section had not been completed by the time Experiment 1 was programmed. Therefore, sample size could not be determined taking into account the results of the meta-analysis. Given that the present experiment included an improved awareness test and that there was no unbiased testing

stage that could give rise to unlearning, we estimated that a sample size twice as large as the typical probabilistic cuing experiment should provide sufficient statistical power to detect awareness. All participants had normal or corrected to normal vision and completed the experimental task individually in isolated cubicles.

Materials and Apparatus

The experimental task was programmed in MATLAB, using Cogent 2000 and Cogent Graphics (The MathWorks, Inc., Natick, MA; www.vislab.ucl.ac.uk/cogent.php) to collect participants' responses and present stimuli on a 17-in TFT computer screen set at a resolution of 1280 × 1024. Each trial began with the brief presentation of a black, 8 × 8 mm fixation cross, followed by a unique search display comprising 11 L-shaped distractors and 1 T-shaped target in random locations, each of them 14 × 14 mm. Distractors and targets were presented on a 12 × 12 grid, invisible to participants, covering an area of 240 × 240 mm in the center of the screen. Distractors could be rotated 0, 90, 180 or 270 degrees. Targets could only be rotated 90 or 270 degrees. Distractors and targets could be randomly presented in yellow, red, green or blue, against a grey background. Participants entered their responses by pressing keys z and m on a standard computer keyboard.

Procedure and Design

At the beginning of the experiment, participants were told that they would have to perform a visual search task. They were instructed to find a rotated T target among a series of distractors and to press key z if the stem of the T pointed to the left and key m if it pointed to the right. They were asked to find the target as quickly as possible but without making errors. After seeing two examples, they started the training stage.

The training stage consisted of 40 blocks, each comprising six trials. In each block the target was located in the rich quadrant on three trials and in each of the sparse quadrants on the remaining three trials, in random order. The rich quadrant was randomly selected for each participant. To reduce fatigue effects, participants could take a brief rest after trials 60, 120, and 180.

Immediately after completing the training stage, participants' explicit knowledge of the distribution of targets was assessed through a ranking test. Participants were told that, even though perhaps they had not noticed it, the target tended to appear more frequently in one quadrant than in the others. Then they were asked to select which quadrant was most likely to contain the target. After entering their response, they selected which was the quadrant second most likely to contain the target, and then third most likely. Once they had entered these three responses, they were given an opportunity to review and change their ranking if they wished.

Results and Discussion

Training Stage

Reaction times (RT) from trials with incorrect responses were removed from the analyses. Similarly, we removed reaction times shorter than 100 ms or longer than 10,000 ms, RTs from trials immediately following a resting break, and all RTs that were three standard deviations faster or slower than each participant's mean RT. To reduce random noise, the 40 blocks of training were collapsed into 20 epochs, each comprising 2 blocks. Figure 1A depicts participants' RTs as a function of the location of the target (in the rich or the sparse region of the search display) and epoch. RTs decreased smoothly over the course of the training stage, but the decline was clearly steeper when the target appeared in the rich than in the sparse region. A 2 (target location: rich vs. sparse) × 20 (epoch: 1-20)

repeated-measures analysis of variance (ANOVA) yielded significant effects of location, $F(1, 31) = 100.72$, $p < .001$, $\eta^2_G = .17$, and epoch, $F(19, 589) = 12.27$, $p < .001$, $\eta^2_G = .11$, as well as a significant interaction, $F(19, 589) = 1.77$, $p = .024$, $\eta^2_G = .01$.

As mentioned in the introduction, a methodological problem in probabilistic cuing experiments is that simply comparing RTs to the target in rich and sparse regions ignores the potential contribution of repetition priming (Walthew & Gilchrist, 2006). Responses tend to be faster if the target appears in the same region as in the previous trial. In our experiment, as in previous ones, repetitions were more likely to happen when the target appeared in the rich region, which means that, in principle, the pattern of results observed in the previous analyses could be entirely due to repetition priming. To discount this possibility, we repeated the analysis removing from the data all trials in which the target appeared in the same quadrant as on the previous trial. The dotted lines in Figure 1A denote RTs excluding those repetitions. A 2 (target location: rich vs. sparse) × 20 (epoch: 1-20) repeated-measures ANOVA on these reaction times again yielded main effects of target location, $F(1, 31) = 58.68$, $p < .001$, $\eta^2_G = .11$, and epoch, $F(19, 589) = 8.80$, $p < .001$, $\eta^2_G = .09$. In contrast, the location × epoch interaction failed to reach statistical significance, $F(19, 589) = 1.39$, $p = .124$, $\eta^2_G = .01$.

Awareness Test

Figure 1B shows the results of the awareness test. 16/32 participants ranked the rich quadrant first in the awareness test, 8 participants ranked it second, 5 ranked it third, and only 3 ranked it in the fourth and last position. This distribution of responses departs significantly from chance performance, $\chi^2(3) = 12.25$, $p = .007$. The observed number of participants ranking the rich quadrant first is significantly greater than chance, represented by the dotted line in Figure 1B, binomial test $p = .002$. From the 16 participants who did

not rank the rich quadrant first, one would expect only one third (5.33) of them to rank it second, but 8 of them did. This proportion, however, was not significantly greater than chance, binomial $p = .126$. Similarly, from the 8 participants who did not rank the rich quadrant first or second, 5 ranked it first, although this failed to be significantly higher than the expected number of 4, binomial $p = .363$. Even though these two comparisons miss statistical significance, the fact that participants who failed to rank the rich quadrant first were numerically more likely to rank it second or third than fourth suggests that they were not totally unaware of the biased distribution of targets across training trials.

Experiment 2

Experiment 2 sought to replicate and extend the findings of the previous experiment. Although the results together with the meta-analysis strongly imply that probabilistic cuing of visual attention is associated with above-chance performance in the awareness test, this conclusion is restricted to circumstances in which a substantial degree of learning has taken place. In Experiment 1, for example, participants received over 200 search trials in which they developed a high degree of search bias, as shown in Figure 1A. Although this number of trials is typical of the studies included in the meta-analysis, it is possible that search bias develops more rapidly than awareness and that Experiment 1 (and previous experiments) have failed to identify the 'sweet spot' at which search bias, but not awareness, has developed. It is possible that implicit learning is acquired faster than explicit knowledge (Bechara, Damasio, Tranel, & Damasio, 1997; Goujon, Didierjean, & Poulet, 2014). Although previous research has generally failed to find clear evidence of different learning rates (e.g., Konstantinidis & Shanks, 2014; Maia & McClelland, 2004; Perruchet & Amorim, 1992), it is important to ask whether the trajectories of learning of

search bias and awareness are similar or different. Thus, in Experiment 2 we test awareness at different time points during search bias training.

The second major contribution is that we introduce a new and alternative method for assessing awareness. In Experiment 1 we employed an innovative improvement to the standard test (asking participants to rank all quadrants rather than just stating which one they believed to be the richest). Here we include a betting test in which on each trial the target was hidden from view and participants wagered on its most likely location.

Method

Participants, Materials, and Apparatus

Given that we succeeded at detecting awareness in Experiment 1, we planned to test a roughly similar number of participants in each of the groups included in Experiment 2. One hundred and thirty-four psychology students from Universidad Nacional de Educación a Distancia, Madrid (UNED) volunteered to take part in the experiment. Random allocation of participants to experimental conditions resulted in 35, 33, 31, and 35 participants in Groups 1, 2, 3, and 4, respectively. As in Experiment 1, all participants completed the experimental task individually in isolated cubicles. Except for the facts that all instructions were in Spanish and distractors and targets were located on a 10×10 grid, invisible to participants, the search displays, and the experimental program in general, were otherwise identical to those of Experiment 1.

Design and Procedure

As in Experiment 1, immediately after reading the instructions, participants began the training stage. The general design of Experiment 2 is summarized in Table 4. As can be seen, there are two crucial differences with respect to Experiment 1. Firstly, training was divided into four stages, some of them separated by an awareness test. Each training stage

consisted of 60 visual search trials, half of them containing the target in the rich quadrant (randomly selected for each participant) and the other half containing the target in the other (sparse) quadrants. In these trials, participants were asked to find the target as quickly as possible using keys z and m to report its orientation. Secondly, the experiment included two types of awareness tests, administered at different points during the experimental session depending on group assignment.

At the end of the experiment, all participants completed a single-item quadrant-guessing test similar to the one used in previous studies of probabilistic cuing. Specifically, participants were invited to select which quadrant they thought had contained the target most often during the training stage. In addition, after some training stages participants completed a 24-trial betting test designed to provide an alternative measure of awareness. In each of these trials, participants were presented with a search display identical to the ones used during the training stage, except that the T-shaped target was replaced by an additional L-shaped distractor. Participants were instructed that their task was to guess which quadrant contained the hidden target, using four different keys (T, Y, G, and H) to select a quadrant. They did not receive any feedback about the accuracy of their responses, but they were told that the program would record whether they provided the correct response or not. Groups 1-4 differed only in the number of betting tests they completed during the experimental session. The group designation (1, 2, 3, 4) refers to the number of betting tests administered during the experiment. Varying the number of tests in this way permits us to measure awareness at different points during the development of probability cuing, while also measuring any quantitative effect that taking an awareness test has on probability cuing.

Results and Discussion

Training Stages

RTs were filtered using the same procedure as in Experiment 1. Figure 2 shows participants' RTs across conditions and epochs (each of them comprising 12 trials, for consistency with Figure 1A), separately for each group. As in Experiment 1, RTs decreased smoothly over the course of the training stage and this decline was stronger when the target appeared in the rich quadrant than when it appeared in any of the sparse quadrants. As can be seen in Figure 2, overall RTs for each condition tended to be similar across groups, except that, unsurprisingly, RTs became slower in the few trials immediately following an awareness test. To reduce the impact of these fluctuations, in the following analyses we aggregated trials at the stage (instead of epoch) level. A 4 (group: 1-4) × 2 (target location: rich vs. sparse) × 4 (stage: 1-4) mixed ANOVA on RTs yielded significant main effects of target location, $F(1, 130) = 404.83$, $p < .001$, $\eta^2_G = .15$, and stage, $F(3, 390) = 73.54$, $p < .001$, $\eta^2_G = .08$, and a significant interaction between these factors, $F(3, 390) = 5.04$, $p = .002$, $\eta^2_G = .002$. The remaining effects and interactions were nonsignificant, all Fs < 1.32. As in Experiment 1, to discount the possibility that these effects were due to repetition priming, we ran the same analysis removing from the sample all trials in which the target appeared in the same quadrant as in the preceding trial. Again, only the main effects of target location, stage and their interaction were statistically significant, all ps < .001.

Taking into account all valid trials, all groups showed significant evidence of probabilistic cuing at epoch 5, all ts > 3.27, ps < .003, d_z's > 0.37, except Group 1, $t(34) = 1.72$, $p = .095$, $d_z = 0.29$. At epochs 10, 15, and 20 all groups showed significant evidence of cuing, ts > 3.48, ps < .002, d_z's > 0.63. To control for repetition priming, we repeated the same analyses excluding trials with repetitions. None of the results changed substantially, except that cuing at epoch 5 in Group 2 also became statistically non-

significant, $t(32) = 1.61$, $p = .117$, $d_z = 0.28$. From these analyses we can conclude that cuing was robust by the end of all training stages, with the possible exception of epoch 5 in Groups 1 and 2.

Awareness Tests

In this experiment, participants' awareness was assessed in two different ways: In the betting tests interpolated between the training stages (and also at the end) and in the traditional quadrant guessing test, conducted only at the end of the experiment (see the design summary in Table 4). The results of these two tests are shown in Figure 3. Panel A shows the average proportion of times that each participant bet on the rich quadrant. As can be seen, this proportion was always numerically above chance, although the 95% confidence intervals excluded .25 in only half of the tests. Participants' performance tended to improve over the experiment, but this cannot be solely due to repeated experience with the betting task, as participants in Group 1, who performed the task just once, showed above chance performance by the end of training (Epoch 20).

These impressions were confirmed by a comparison of linear mixed-effects models. Firstly, we fitted a linear mixed model predicting the proportion of bets to the rich quadrant from group, stage and their interaction, adding a random intercept for participants. This model did not perform significantly better than an otherwise identical model without the group × stage interaction, $\chi^2(1) < 0.01$, $p = .950$. The model with just the two main effects of group and stage did not perform better than an identical model with just stage, $\chi^2(1) = 1.50$, $p = .220$. However, it did perform better than an identical model without stage, $\chi^2(1) = 4.21$, $p = .040$. This pattern of results suggests that the proportion of bets to the rich quadrant did increase over the course of the experiment, to a similar extent in all groups.

Figure 3B shows the proportion of participants who chose the rich quadrant in the traditional quadrant guessing test. As can be seen, although individually not all participants made the correct choice, at the group level performance is clearly above chance. Binomial tests conducted separately for each group showed that the proportion of participants selecting the rich quadrant at test was significantly higher than .25 in all cases, largest p = .027.

Updated meta-analysis

Some of the dependent variables gathered in Experiments 1 and 2 are directly comparable to ones garnered from studies included in the initial meta-analysis reported in the present article. For instance, the proportion of participants ranking the rich quadrant first in Experiment 1 can be compared to the equivalent proportion of participants selecting the rich quadrant in experiments using the traditional single-trial quadrant guessing test. Similarly, the final test included at the end of Experiment 2 is also analogous to the traditional test. Given these similarities, it is possible to update the results of the meta-analysis with the data collected in the present experiments. Figure 4 shows the results of the present experiments, along with a summary of the initial meta-analysis, a meta-analysis of the present Experiments 1 and 2, and a comprehensive meta-analysis including all the effect sizes of the original meta-analysis together with those of Experiments 1 and 2.

While the number of participants selecting the rich quadrant was somewhat larger in Groups 3 and 4 of Experiment 2, the meta-analysis of the present studies reveals very little evidence of heterogeneity, $Q(4) = 473$, $p = .316$, $I^2 = 14.71\%$. In other words, the amount of variation seen across the different groups in Experiments 1 and 2 is not higher than expected by mere chance. In contrast, the updated overall meta-analysis reveals a

substantial amount of heterogeneity, $Q(48) = 155.13$, $p < .001$, $I^2 = 70.82\%$, although this result is barely different from the original meta-analysis excluding the present experiments.

We also repeated the moderator analyses reported in Table 2 with the updated set of effect sizes. The results changed very little, except that the presence or absence of an unbiased testing stage now did reliably moderate the results, $Q(1) = 4.11$, $p = .042$: Studies without a testing stage yielded larger effect sizes, $h = 0.52$, 95% CI [0.32, 0.72], than those with one, $h = 0.26$, 95% CI [0.11, 0.41]. This confirms our earlier hypothesis that an unbiased testing stage inserted between the initial probabilistic cuing training stage and the awareness test can dilute the attentional biases learned during the former and hence lead to underestimation of awareness. For the rest of the moderators listed in Table 2, the results were virtually identical after including the present experiments.

Relation between Visual Search and Awareness in Experiments 1 and 2

As explained previously, some studies have concluded that probabilistic cuing is unconscious because the size of the effect tends to be similar for participants who show some signs of awareness and those who do not (Jiang, Sha et al., 2018; Salovich, Remington, & Jiang, 2018; Twedell, Koutstaal, & Jiang, 2017). This correlational approach to assessing the unconscious character of cognitive processes is also quite common in other areas of research (e.g., Colagiuri & Livesey, 2016). We therefore now ask whether there was any correlation between participants' performance in the awareness test and the size of the probability cueing effect in Experiments 1 and 2.

For Experiment 1, we only included RTs from the second half of the experiment (to ensure that RTs reflect asymptotic performance) and trials without repetitions (to ensure that they are not biased by repetition priming). Figure 5A shows the size of the cuing effect observed for each participant (i.e., average RTs to targets in the sparse quadrants minus

average RTs to targets in the rich quadrant), as a function of the rank that each participant gave to the rich quadrant. Positive scores are indicative of probabilistic cuing. In principle, one would expect to find the largest probabilistic cuing effect among participants who ranked the rich quadrant first. However, if anything we found a small and non-significant trend in the opposite direction, Spearman's rho = .08, p = .652.

For Experiment 2, we explored the correlation between cuing and awareness considering only trials without repetitions (again, to eliminate repetition priming effects) from Stage 4. As shown in Figure 5B, participants who betted more frequently on the rich quadrant tended to show somewhat larger probabilistic cuing effects, although the relation is far from reaching statistical significance, r = .04, p = .631. A multiple regression of cuing on group and proportion of bets failed to find a significant main effect of group (p = .217) or a significant group × proportion of bets interaction (p = .100). Therefore, although participants showed clear evidence of both cuing and awareness at the end of the experiment, there was no obvious correlation between them in any experimental group.

As explained previously, the fact that measures of learning and awareness fail to correlate with each other, even in large samples, is not completely unexpected. Even if there is a positive correlation between learning and awareness at the latent level, the observed correlation will necessarily be attenuated if their measures are unreliable. Is there any reason to suspect that the present correlations are attenuated by measurement error? In an attempt to answer this question, we first estimated the split-half reliability of probabilistic cuing by computing, for each participant, a separate measure of probabilistic cuing (RTs in sparse quadrants minus RTs in the rich quadrant) from odd and even trials, after excluding invalid trials, quadrant repetitions, and trials from the first half of the training stage. The correlations between those two measures across participants were .30 and .33 for Experiments 1 and 2, respectively, with a meta-analytic average of .32.

Applying the Spearman-Brown correction to this value, the split-half reliability of probabilistic cuing can be estimated as .49. This value is quite close to the reliability of other implicit learning tasks, such as the serial reaction time task (Kaufman et al., 2010; Siegelman & Frost, 2015).

We cannot compute a similar split-half reliability index for the measure of awareness employed in Experiment 1, because the quadrant ranking test involved a single trial. However, it is possible to estimate the split-half reliability of the betting test used in Experiment 2. As in the case of visual search times during the training stage, we divided the 24 trials of the last betting test into two identical data sets based on an odd/even trial split and then correlated the proportion of bets to the rich quadrant in both halves. The results were remarkably similar to those of the cuing effect: We obtained a split-half correlation of .34, which became .50 after applying the Spearman-Brown correction. Again, this value is close to the reliability estimate of the awareness test used in a related implicit learning paradigm (Smyth & Shanks, 2008).

Given these reliabilities, which would be regarded as 'unacceptable' in most psychometric contexts (Cicchetti & Sparrow, 1981), we can conclude that the observed correlation between probabilistic cuing and awareness will be roughly half the size (i.e., attenuated by a factor of $\sqrt{.49}\cdot\sqrt{.50}$ = .49) of the true correlation between their corresponding constructs at the latent level. In practical terms, this means that to detect a medium-size correlation of r = .30 at the latent level (.15 at the observed level) with 90% power, a sample of at least 462 participants would be needed. Detecting a smaller, but yet meaningful, correlation of .20 at the latent level would require more than 1,000 participants.

To make things worse, there are good reasons to suspect that the reliability of the betting test we used in Experiment 2 is substantially higher than the measures of awareness

used in previous research, usually comprising a single quadrant-guessing test. In a way, we can see each of the trials in our betting test as an independent single-trial quadrant-guessing test. Following this logic, if the reliability of 24-trial betting test is .50, using Spearman-Brown's prediction formula, we can estimate that the reliability of a test comprising just one trial must be .04.

Converging evidence for this pessimistic conclusion comes from a recent reanalysis of probabilistic cuing experiments conducted by Jiang, Sha et al. (2018). Before being asked to guess the rich quadrant, the participants tested in these experiments were also asked whether they thought that the target had appeared in some locations disproportionally often. With the data reported in the reanalysis, it is possibly to reconstruct the 2 × 2 contingency matrix of responses to both questions. The relationship between the two responses, both of which are assumed to be measures of awareness, is not statistically significant, $\chi^2(1) = 2.70$, $p = .100$, and corresponds to a tetrachoric correlation of .15. To the extent that both questions are intended to measure the same construct (i.e., awareness of the target distribution) these figures suggest that their reliability and validity is minimal.

Even assuming that .04 and .15 are gross underestimations of the reliability of the traditional quadrant-guessing test, it is clear that any plausible estimate falls very far short of any psychometric standard for correlational research, however lenient. Detecting a significant correlation between learning and awareness with these dependent measures is simply impossible, unless thousands of participants are tested. It is worth adding that the alternative approach of testing for probabilistic cuing in the subset of participants showing no signs of awareness is bound to be misleading (Shanks, 2017b). Given the low reliabilities of awareness tests, many participants who possess some explicit knowledge

will necessarily be misclassified has being unaware and, consequently, the true level of awareness in the subset of participants will be underestimated.

Modeling Single- and Dual-Process Accounts of Probabilistic Cuing

The analysis of Experiments 1 and 2 revealed that, overall, participants' performance tended to be systematically above chance on all measures of awareness, although this trend did not reach statistical significance in all cases. In contrast, we failed to detect a significant correlation between performance and awareness, an outcome we attribute at least in part to the low reliability of our dependent measures. To further explore whether performance in both tasks might be driven by a common underlying representation, we developed several competing models of probabilistic cuing, one assuming a single memory representation for cuing and awareness and two assuming different representations for the two tasks, and compared their relative fits to the data gathered in Experiment 2.

Model 1 assumes that the proportion of bets in the awareness test and RTs in the visual search task are driven by a common memory trace. In this model, parameter ω_R represents the weight or strength of the rich quadrant, which affects behaviour in both the awareness tests and in search times. This parameter can be regarded as a latent variable representing participants' perception of the tendency for targets to appear in the rich quadrant. In our implementation, ω_R can take values between 0 and 1. Each sparse quadrant also has a weight denoted by ω_S, which is computed as $(1-\omega_R)/3$, so that the sum of the weights of all four quadrants is 1. Participants' performance in the betting task is assumed to be a direct translation of these weights, so that on any given trial, the probability that they will bet on the rich quadrant is exactly ω_R while the probability that they will bet on any of the three sparse quadrants is $3\omega_S$.

Participants' search times in each quadrant during the training stages are assumed to be inversely related to the weights. Specifically, RTs are modelled as an ex-Gaussian distribution, which has shown a good fit to visual search times in previous research (e.g., Palmer, Horowitz, Torralba, & Wolfe, 2011). This distribution is the convolution of a normal and an exponential distribution and has three parameters: the mean of the normal component (μ), the standard deviation of the normal component (σ), and a third parameter determining the mean and standard deviation of the exponential component (τ). Model 1 assumes that RTs are sampled from different ex-Gaussian distributions when the target is presented in the rich quadrant and when it is presented in any of the sparse quadrants. Specifically, parameters σ and τ are constrained to adopt equal values in all conditions, but μ adopts different values depending on the location of the target and on repetition priming. For trials in which the target is in the rich quadrant μ_R is computed as $\mathbf{a}+\mathbf{b}/\omega_R$, while for trials in which the target is in the sparse quadrants μ_S is computed as $\mathbf{a}+\mathbf{b}/\omega_S$. In other words, μ is a linear transformation of the inverse of each quadrant's weight, with intercept **a** and slope **b**, so that RTs will be faster for quadrants with larger weights. Importantly, parameters **a** and **b** are fixed across all trials, so that differences in RTs across conditions can only be attributed to differences in ω_R and ω_S. Additionally, to model repetition priming, a constant value **p** is subtracted from μ_R or μ_S whenever the target is in the same quadrant as on the previous trial. Therefore, Model 1 has six free parameters to be estimated from the empirical data: ω_R, **a**, **b**, **p**, σ, τ.

Model 2A is identical to Model 1, except that ω_R is assumed to determine only RTs during the visual search task. Performance in the awareness test is assumed to be completely random. That is to say, on any given trial, the probability that participants will bet on the rich quadrant is assumed to be exactly .25, regardless of the specific value of ω_R. This assumption is akin to the expectation that participants' performance in the awareness

test will tend to be at chance, even if search times are faster when the target is located in the rich quadrant. Model 2A thus has the same number of free parameters as Model 1.

Unlike Model 2A, Model 2B assumes that performance in the awareness test is driven by an independent parameter, γ_R, with values from 0 to 1, which determines the propensity of each participant to bet on the rich quadrant and which is completely independent from ω_R. The probability of betting on any of the sparse quadrants is $3\gamma_S$, where γ_S is computed as $(1-\gamma_R)/3$. Therefore, Model 2B has seven free parameters: the same parameters as Models 1 and 2A plus the additional parameter γ_R. Model 1 can be seen as a special case of Model 2B in which $\gamma_R = \omega_R$, and Model 2A can be seen as a special case of Model 2B in which $\gamma_R = .25$.

We fitted the three models independently to each participant using individual-trial data from stages 2-4, including both training stages and betting tests. RTs from training stages were trimmed following exactly the same procedure as in Experiments 1 and 2. This resulted in a somewhat different number of valid trials for each participant. Additionally, not all groups were exposed to the same number of betting tests. Consequently, the number of valid observations from each participant ranged from 190 to 251, with an average of 227.07 (*SD* = 19.83). In total we fitted 30,428 data points. We optimized the fit of each model using the Nelder-Mead algorithm to find the combination of parameter values with the lowest negative log likelihood. As starting values for all models, we set ω_R, **a** and **p** to .25, 0, and 0, respectively. To select suitable starting values for the remaining parameters, we first fitted an ex-Gaussian distribution to all the RTs using the dexgauss function of the 'retimes' R package. The results were used to inform the selection of starting values for **b**, σ, and τ. The empirical probability of betting on the rich quadrant was used as the starting value for γ_R in Model 2B.

Table 5 shows the mean best-fitting values for all models and Table 6 shows the correspondence between observed RTs in the visual search task and the predictions made by each model. As can be seen all models predicted RT data with reasonable accuracy, except in the case when the target was located in the same sparse quadrant as in the previous trial. This is a normal consequence of the fact that this condition contributed fewer data points to the model fit. Table 6 also shows the correlation between performance in the awareness test and the predictions of Models 1 and 2B. (This correlation is irrelevant for Model 2A as it predicts that the proportion of bets to the rich quadrant is .25.) In this case, the fit of Model 2B is perfect, while Model 1 achieves only a modest .39 correlation between observed and expected data. Again, this discrepancy is logical, given that Model 2B is saturated because its predictions depend on parameter γ_R, which only needs to fit the awareness data, while in Model 1 these predictions depend on parameter ω_R, which is constrained to fit both awareness data and RTs.

To compare the goodness of fit of each model we computed the Akaike Information Criterion (AIC) for each participant (Lewandowsky & Farrell, 2010). Table 6 shows the sum of AICs across participants for each model and how many individual participants were best fitted by each model. As can be seen, at the aggregate level, Model 1 provides the best fit to the empirical data and also yields the best fit for most individual participants, although roughly 55% of them were better fit by Models 2A or 2B. Model recovery analyses, presented in Appendix B, show that our model fitting procedure described in the preceding analyses usually retrieves the true underlying model, except, logically, when the predictions of Model 2B overlap with those of Model 1 (i.e., when $\gamma_R = \omega_R$) or Model 2A (i.e., when $\gamma_R = .25$).

Perhaps most interestingly, Model 1 not only fits reasonably well participants' performance in the visual search task and in the awareness test, it also predicts extremely

low correlations between these dependent variables, even though, at the latent level, both RTs and bets are assumed to be driven by a common latent variable, ω_R (and by no other learning-dependent factors). Figure 5D shows an example of data generated from Model 1 using the best-fitting parameters for each participant. To mirror the empirical data depicted in Figure 5B, where each data point was computed on the basis of 60 visual search trials (ignoring the removal of invalid RTs and repeated trials) and 24 betting-test trials, we simulated for each participant 30 visual search trials with the target in the rich quadrant, 30 visual search trials with the target in the sparse quadrants, and 24 betting trials in the awareness test. As can be seen, although the model does predict a positive correlation between performance and awareness, this correlation is rather weak. The bell-shaped curve in Figure 5F summarizes the distribution of correlation coefficients obtained across 10,000 iterated simulations like the one represented in Figure 5D. On average, these simulations produce a correlation of .12 between probabilistic cuing and the proportion of bets to the rich quadrant. The diamond shape represents the 95% CI of the observed correlation in the empirical data depicted in Figure 5B. Although the observed correlation is clearly lower than the correlation predicted by the model, the two distributions are by no means inconsistent with each other, and 85.91 % of simulated correlations fall within the 95% CI of the observed data.

Given that Experiment 1 included an awareness test with just one trial, it is not feasible to fit Model 1 to participants' data, as the resulting best-fitting parameters would be driven almost exclusively by the (hundreds of) RTs collected during the visual search task. However, it is possible to use the model fits of Experiment 2 to simulate how those participants would have performed in an awareness test like the one conducted in Experiment 1. In the following simulations, we assumed that the probability of ranking the rich quadrant first in the awareness test would be given by ω_R. Among participants who

did not rank the rich quadrant first, the probability of ranking it second would be equal to ω_R divided by the sum of the weights of all the quadrants that remained unselected, and so on for the third and fourth ranking positions. In this case, we simulated 60 rich quadrant trials and 60 sparse quadrant trials, to equate the number of trials included in the empirical data shown in Figure 5A. Figure 5C shows the results of a simulation based on this approach. As expected, the model predicts a (negative) correlation between learning and awareness, such that participants who tend to rank the quadrant first show, on average, larger cuing effects. However, the correlations are remarkably small. Figure 5E show the distribution of rho correlation coefficients obtained across 10,000 iterations. Again, these results are not inconsistent with the empirical results of Experiment 1, summarized by the diamond-shaped figure. In this case 99.22% of simulated correlations fall within the 95% CI of the observed data.

General Discussion

The idea that implicit processes permeate much of perception and cognition has become a central plank of contemporary psychology (Greenwald & Banaji, 2017; Hassin, 2013). Yet much of the evidence supporting this viewpoint is controversial (Shanks, 2017a). One of the reasons for this state of affairs is that some of the most frequent approaches to test the unconscious character of cognitive processes are methodologically problematic, particularly when strong inferences are made on the basis of null results obtained with small samples and unreliable measures. This approach is extremely popular in many areas of implicit cognition research, including probabilistic cuing of visual attention. In fact, meta-analytic integration of previous research in this domain shows that participants do perform significantly above chance in the standard region-guessing test devised to measure awareness, although the size of this effect is relatively small and hence

unlikely to be detected reliably with a small sample. Our own Experiments 1 and 2, which partially replicate previous research, confirm this conclusion. We also found converging evidence from alternative measures of awareness, such as the ranking test in Experiment 1 and the betting test in Experiment 2.

Consistent with previous studies, we also failed to detect a significant correlation between the size of probabilistic cuing and participants' measured awareness. However, we found that these dependent measures show rather low split-half reliabilities, even though each betting test in Experiment 2 comprised a much larger number of trials than any previous study (24 trials, as opposed to the traditional single-trial quadrant-guessing test). Given these reliabilities, it is unsurprising that previous research has failed to detect correlation between learning and awareness. Furthermore, we showed that a simple model in which probabilistic cuing and performance in the awareness test are based on a single latent variable also predicts remarkably small correlations between learning and awareness, albeit different from zero. Both lines of thought converge to the same conclusion: It is naïve to expect a significant correlation between these dependent variables, unless thousands of participants are tested.

The fact that we question the conclusions drawn on the basis of these approaches does not mean that we are necessarily equating probability cuing with explicit or goal-driven attention. In fact, some results of the present study are consistent with the idea that unconsious processes may contribute to probabilistic cuing to some extent in some participants. For instance, our modeling suggests that Models 2A and 2B, which assume independent memory traces for search times and awareness, do provide a better fit for 74/134 participants. In other words, for many participants there was a dissociation between awareness and probabilistic cuing that may indicate the presence of a truly unconscious learning effect. The fact that the predictions of Model 1 did not correlate perfectly with

participants' performance also dovetails with this conclusion. Therefore, although our analyses suggest that the implicit character of probabilistic cuing may have been exaggerated in previous research, we do not reject categorically the possibility that probabilistic cuing can be unconscius for some participants. Instead, what we want to contend is that the dominant approach to testing the role of awareness in probablistic cuing, based on null results in underpowered tests, is methologically flawed and misleading, and should be abandoned in favor of alternative methods. Our own modeling approach can be considered a first step towards the development of suitable methods for the categorization of aware and unaware participants (see also Berry, Shanks, Speekenbrink, & Henson, 2012; Rouder, Morey, Speckman, & Pratte, 2007).

Recommendations for future research

There are some obvious solutions to the shortcomings of current research practices in implicit learning research, of which probabilistic cuing is just an example. In the present article we have advocated the use of a modeling approach, but researchers can consider alternative and more easily-implemented solutions. For instance, if they wish to show that participants are truly at chance on some dependent measure, they can employ a Bayes Factor analysis, which unlike NHST can be used to quantify evidence in favor of the null hypothesis (Dienes, 2011, 2015; Rouder et al., 2009; Sand & Nilsson, 2016). Even without stepping out of the NHST framework, much can be gained by simply calibrating more carefully the significance threshold according to the relative risks of making Type I or Type II errors (Lakens et al., 2018). Setting α to .05 and β to .20 may make sense when Type I errors are more likely or more important than Type II errors. But for the particular case of researchers wishing to conclude that some dependent variable, such as awareness,

is at chance, it makes more sense to reverse these values, setting α to .20 and β to .05. To make this strategy even more effective, it would be advisable to use one-tailed tests.

Interestingly, if anything, current research practices in implicit learning research go in the opposite direction. For instance, in their famous study on unconscious instrumental learning, Pessiglione, Petrovich, Daunizeau, Palminteri, Dolan, and Frith (2008) used one-tailed tests for all the analyses assessing learning, but two-tailed tests for all the analyses assessing awareness. This strategy increases power to detect learning and decreases power to detect awareness. Finding "unconscious" learning is hardly surprising in these conditions. Perhaps more interestingly, researchers can also test awareness using equivalence tests (Lakens, 2017). These tests require researchers to define what is the smallest effect size that they would consider of practical significance. If the effect size (e.g., for awareness) they find is significantly smaller than this value and non-significantly different from zero, then it is concluded that, for pragmatic purposes, the effect can be considered equivalent to zero. Interestingly this approach is helpful not only to assess whether performance in the awareness test is at chance; It can also be used to assess whether the correlation between learning and awareness is actually indistinguishable from zero given some basic information about the reliabilities of both dependent variables. If, for instance, we consider *a priori* that any correlation below .10 would be of no practical significance and at the same time we also know that the reliability of our dependent variables is around .50, then we should choose .05, instead of .10, as the smallest effect size of interest.

A further simple but important recommendation is that researchers should endeavor to formally measure the reliability of their explicit and implicit measures, either by using test-retest designs or (as done here) the split-half method. Without reliability estimates, interpretation of correlation coefficients is fraught with danger.

To sum up, although these approaches are relatively unconventional, there are simple statistical methods, both within and beyond the traditional NHST framework, that can be used to improve the measurement and analysis of awareness in implicit learning studies and ameliorate the risk of Type II errors.

Context of the Research

The meta-analysis and empirical studies reported in the present article are part of a larger research program aimed at identifying methodological shortcomings in implicit cognition research and suggesting alternative approaches. In previous articles we have highlighted how widespread practices, such as inferring the absence of awareness from statistically null results in explicit memory tests, can lead to a largely misleading picture of the role of unconscious processes in human cognition (Vadillo et al., 2016). Our research also addresses how measurement error can contribute to these problems by making awareness even more difficult to detect, diluting the observed correlations between measures of learning and awareness, and giving rise to new methodological problems such as regression to the mean (Shanks, 2017b). The fact that implicit learning measures tend to be very unreliable also questions the significance of alleged null correlations between these measures and problems in language and reading acquisition (West et al., 2018). In future studies, we plan to offer specific guidelines to overcome these and other methodological problems in implicit cognition research.

References

*Addleman, D. A., Tao, J., Remington, R. W., & Jiang, Y. V. (2018). Explicit, goal-driven attention, unlike implicitly learned attention, spreads to secondary tasks. *Journal of Experimental Psychology: Human Perception and Performance, 44,* 356-366.

Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (1997). Deciding advantageously before knowing the advantageous strategy. *Science, 275,* 1293-1295.

Beck, M. R., Angelone, B. L., Levin, D. T., Peterson, M. S., & Varakin, D. A. (2008). Implicit learning for probable changes in a visual change detection task. *Consciousness and Cognition, 17,* 1192-1208.

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, UK: Lawrence Erlbaum.

Brockmole, J. R., & Henderson, J. M. (2006). Using real-world scenes as contextual cues for search. *Visual Cognition, 13,* 99-108.

Brockmole, J. R., & Vo, M. L.-H. (2010). Semantic memory for contextual regularities within and across scene categories: Evidence from eye movements. *Attention, Perception, & Psychophysics, 72,* 1803-1813.

Chang, T.-Y., Little, D. R., & Yang, C.-T. (2016). Selective attention modulates the effet of target location probability on redundant signal processing. *Attention, Perception, & Psychophysics, 78,* 1603-1624.

Chua, K.-W., & Gauthier, I. (2016). Category-specific learned attentional bias to object parts. *Attention, Perception, & Psychophysics, 78,* 44-51.

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.

Chun, M. M., & Turk-Browne, N. B. (2008). Associative learning mechanisms in vision. In S. J. Luck & A. Hollingworth (Eds.), *Visual memory* (pp. 209–245). New York: Oxford University Press.

Chukoskie, L., Snider, J., Mozer, M. C., Krauzlis, R. J., & Sejnowski, T. J. (2013). Learning where to look for a hidden target. *Proceedings of the National Academy of Sciences, 110,* 10438-10445.

Cicchetti, D. V., & Sparrow, S. A. (1981). Developing criteria for establishing interrater reliability of specific items: Applications to assessment of adaptive behavior. *American Journal of Mental Deficiency, 86,* 127-137.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences.* New York, NY: Routledge.

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.

Dienes, Z. (2011). Bayesian versus Orthodox statistics: Which side are you on? *Perspectives on Psychological Sciences, 6,* 274-290.

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: Oxford University Press.

*Druker, M., & Anderson, B. (2010). Spatial probability aids visual stimulus discrimination. *Frontiers in Human Neuroscience, 4,* 63.

Dulany, D. E. (1961). Hypotheses and habits in verbal "operant conditioning". *Journal of Abnormal and Social Psychology, 63,* 251-263.

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.

Fecteau, J. H., Korjoukov, I., & Roelfsema, P. R. (2009). Location and color biases have different influences on selective attention. *Vision Research, 49,* 996-1005.

Fergusson, D., Doucette, S., Cranley Glass, K., Shapiro, S. Healy, D., & Hutton, B. (2006). Association between suicide attempts and selective serotonin reuptake inhibitors: systematic review of randomised controlled trials. *British Medical Journal, 330,* 396.

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science, 345,* 1502-1505.

Geng, J. J., & Behrmann, M. (2002). Probability cuing of target location facilitates visual search implicitly in normal participants and patients with hemispatial neglect. *Psychological Science, 13,* 520-525.

Geng, J. J., & Behrmann, M. (2005). Spatial probability as an attentional cue in visual search. *Perception & Psychophysics, 67,* 1252-1268.

Goschy, H., Bakos, S., Müller, H. J., & Zehetleitner, M. (2014). Probability cueing of distractor locations: Both intertrial facilitation and statistical learning mediate interference reduction. *Frontiers in Psychology, 5,* 1195.

Goujon, A., Didierjean, A., & Poulet, S. (2014). The emergence of explicit knowledge from implicit learning. *Memory & Cognition, 42,* 225-236.

Greenwald, A. G., & Banaji, M. R. (2017). The implicit revolution: Reconceiving the relation between conscious and unconscious. *American Psychologist, 72,* 861-871.

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.

Hartgerink, C. H. J., Wicherts, J. M., & van Assen, M. A. L. M. (2017). Too good to be false: Nonsignificant results revisited. *Collabra: Psychology, 3,* 9.

Hassin, R. R. (2013). Yes it can: On the functional abilities of the human unconscious. *Perspectives on Psychological Science, 8,* 195-207.

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., Gray, K. L. H., Garner, M., & Adams, W. J. (2016). Are visual threats prioritized without awareness? A critical review and meta-analysis involving 3 behavioral paradigms and 2696 observers. *Psychological Bulletin, 142,* 934-968.

Hilgard, J., Engelhardt, C. R., Bartholow, B. D., & Rouder, J. N. (2017). How much evidence is $p > .05$? Stimulus pre-testing and null primary outcomes in violent video games research. *Psychology of Popular Media Culture, 6,* 361-380.

Hoekstra, R., Finch, S., Kiers, H. A. L., & Johnson, A. (2016). Probability as certainty: Dichotomous thinking and the misuse of *p* values. *Psychonomic Bulletin & Review, 13,* 1033-1037.

Hoffmann, J., & Kunde, W. (1999). Location-specific target expectancies in visual search. *Journal of Experimental Psychology: Human Perception and Performance, 25,* 1127-1141.

Jiang, Y. V. (2018). Habitual versus goal-driven attention. *Cortex, 102,* 107-120.

*Jiang, Y. V., Capistrano, C. G., Esler, A. N., & Swallow, K. M. (2013). Directing attention based on incidental learning in children with autism spectrum disorder. *Neuropsychology, 27,* 161-169.

*Jiang, Y. V., Koutstaal, W., Twedell, E. L. (2016). Habitual attention in older and young adults. *Psychology and Aging, 31,* 970-980.

Jiang, Y. V., Sha, L. Z., & Sisk, C. A. (2018). Experience-guided attention: Uniform and implicit. *Attention, Perception, & Psychophysics, 80,* 1647-1653.

*Jiang, Y. V., Sha, L. Z., & Remington, R. W. (2015). Modulation of spatial attention by goals, statistical learning, and monetary reward. *Attention, Perception, & Psychophysics, 77,* 2189-2206.

*Jiang, Y. V., & Swallow, K. M. (2013). Spatial reference frame of incidentally learned attention. *Cognition, 126,* 378-390.

*Jiang, Y. V., & Swallow, K. M. (2014). Changing viewer perspectives reveals constraints to implicit visual statistical learning. *Journal of Vision, 14,* 3.

*Jiang, Y. V., Swallow, K. M., & Capistrano, C. G. (2013). Visual search and location probability learning from variable perspectives. *Journal of Vision, 13,* 13.

*Jiang, Y. V., Swallow, K. M., & Rosenbaum, G. M. (2013). Guidance of spatial attention by incidental learning and endogenous cuing. *Journal of Experimental Psychology: Human Perception and Performance, 39,* 285-297.

*Jiang, Y. V., Swallow, K. M., Rosenbaum, G. M., & Herzig, C. (2013). Rapid acquisition but slow extinction of an attentional bias in space. *Journal of Experimental Psychology: Human Perception and Performance, 39,* 87-99.

Jiang, Y. V., Swallow, K. M., & Sun, L. (2014). Egocentric coding of space for incidentally learned attention: Effects of scene context and task instructions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40,* 223-250.

*Jiang, Y. V., Swallow, K. M., Won, B.-Y., Cistera, J. D., & Rosenbaum, G. M. (2015). Task specificity of attention training: The case of probability cuing. *Attention, Perception, & Psychophysics, 77,* 50-66.

*Jiang, Y. V., & Won, B.-Y. (2015). Spatial scale, rather than nature of task or locomotion, modulates the spatial reference frame of attention. *Journal of Experimental Psychology: Human Perception and Performance, 41,* 866-878.

*Jiang, Y. V., Won, B.-Y., & Swallow, K. M. (2014). First saccadic eye movement reveals persistent attentional guidance by implicit learning. *Journal of Experimental Psychology: Human Perception and Performance, 40,* 1161-1173.

*Jiang, Y. V., Won, B.-Y., Swallow, K. M., & Mussack, D. M. (2014). Spatial reference frame of attention in a large outdoor environment. *Journal of Experimental Psychology: Human Perception and Performance, 40,* 1346-1357.

Jones, J. L., & Kaschak, M. P. (2012). Global statistical learning in a visual search task. *Journal of Experimental Psychology: Human Perception and Performance, 38,* 152-160.

Kabata, T., & Matsumoto, E. (2012). Cueing effects of target location probability and repetition. *Vision Research, 73,* 23-29.

Kabata, T., Yokoyama, T., Noguchi, Y., & Kita, S. (2014). Location probability learning requires focal attention. *Perception, 43,* 344-350.

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.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2,* 196-217.

Konstantinidis, E., & Shanks, D. R. (2014). Don't bet on it! Wagering as a measure of awareness in decision making under uncertainty. *Journal of Experimental Psychology: General, 143,* 2111-2134.

Lakens, D. (2017). Equivalence tests: A practical primer for t-tests, correlations, and meta-analyses. *Social Psychological and Personality Science, 8,* 355-362.

Lakens, D., Adolfi, F.G., Albers, C., Anvari, F., Apps, M.A.J., Argamon, S.E., ... & Zwaan, R. (2018). Justify Your Alpha. *Nature Human Behavior, 2,* 168-171.

Leganes-Fonteneau, M., Scott, R., & Duka, T. (2018). Attentional responses to stimuli associated with a reward can occur in the absence of knowledge of their predictive values. *Behavioural Brain Research, 341,* 26-36.

Lewandowsky, S., & Farrell, S. (2010). *Computational modeling in cognition: Principles and practice.* Sage Publications.

Liu, C.-L., Chiau, H.-Y., Tseng, P., Hung, D. L., Tzeng, O. J. L., Muggleton, N. G., & Juan, C.-H. (2010). Antisaccade cost is modulated by contextual experience of location probability. *Journal of Neurophysiology, 103,* 1438-1447.

Maia, T. V., & McClelland, J. L. (2004). A reexamination of the evidence for the somatic marker hypothesis: What participants really know in the Iowa gambling task. *Proceedings of the National Academy of Sciences, 101,* 16075-16080.

Maljkovic, V., & Nakayama, K. (1996). Priming of pop-out: II. The role of position. *Perception & Psychophysics, 58,* 977-991.

Miller, J. (1988). Components of the location probability effect in visual search tasks. *Journal of Experimental Psychology: Human Perception and Performance, 14,* 453-471.

Müller, H. J., & Findlay, J. M. (1987). Sensitivity and criterion effects in the spatial cuing of visual attention. *Perception & Psychophysics, 42,* 383-399.

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology, 69,* 17.1-17.24.

Newell, B. R., & Shanks, D. R. (2014). Unconscious influences on decision making: A critical review. *Behavioral and Brain Sciences, 37,* 1-19.

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.

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science, 7,* 528-530.

Pellicano, E., Smith, A. D., Cristino, F., Hood, B. M., Briscoe, J., & Gilchrist, I. D. (2011). Children with autism are neither systematic nor optimal foragers. *Proceedings of the National Academy of Sciences, 108,* 421-426.

Perruchet, P., & Amorim, M. A. (1992). Conscious knowledge and changes in performance in sequence learning: Evidence against dissociation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 785-800.

Pessiglione, M. Petrovich, P., Daunizeau, J., Palminteri, S., Dolan, R. J., & Frith, C. D. (2008). Subliminal instrumental conditioning demonstrated in the human brain. *Neuron, 59,* 561-567.

Rabitt, P., Cumming, G., & Vyas, S. (1979). Modulation of selective attention by sequential effects in visual search tasks. *Quarterly Journal of Experimental Psychology, 31,* 305-317.

Reder, L. M., Weber, K., Shang, J., & Vanyukov, P. M. (2003). The adaptive character of the attentional system: Statistical sensitivity in a target localization task. *Journal of Experimental Psychology: Human Perception and Performance, 29,* 631-649.

Reed, J., & Johnson, P. (1994). Assessing implicit learning with indirect tests: determining what is learned about sequence structure. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 585-594.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Pratte, M. S. (2007). Detecting chance: A solution to the null sensitivity problem in subliminal priming. *Psychonomic Bulletin & Review, 14*, 597-605.

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.

*Salovich, N. A., Remington, R. W., & Jiang, Y. V. (2018). Acquisition of habitual visual attention and transfer to related tasks. *Psychonomic Bulletin & Review, 25,* 1052-1058.

Sand, A., & Nilsson, M. E. (2016). Subliminal or not? Comparing null-hypothesis and Bayesian methods for testing subliminal priming. *Consciousness and Cognition, 44,* 29-40.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105,* 309-316.

Shang, J., Fu, Q., Dienes, Z., Shao, C., &, Fu, X. (2013) Negative affect reduces performance in implicit sequence learning. *PLOS ONE, 8,* e54693.

Shanks, D. R. (2005). Implicit learning. In K. Lamberts & R. L. Goldstone (Eds.), *Handbook of cognition* (pp. 202–220). London, UK: Sage.

Shanks, D. R. (2017a). Misunderstanding the behavior priming controversy: Comment on Payne, Brown-Iannuzzi, and Loersch (2016*). Journal of Experimental Psychology: General, 146,* 1216-1222.

Shanks, D. R. (2017b). Regressive research: The pitfalls of post hoc data selection in the study of unconscious mental processes. *Psychonomic Bulletin & Review, 24,* 752-775.

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

Shaqiri, A., & Anderson, B. (2012). Spatial probability cuing and right hemisphere damage. *Brain and Cognition, 80,* 352-360.

Shaw, M. L. (1978). A capacity allocation model for reaction time. *Journal of Experimental Psychology: Human Perception and Performance, 4,* 586-598.

Shaw, M. L., & Shaw, P. (1977). Optimal allocation of cognitive resources to spatial locations. *Journal of Experimental Psychology: Human Perception and Performance, 3,* 201-211.

Siegelman, N., & Frost, R. (2015). Statistical learning as an individual ability: Theoretical perspectives and empirical evidence. *Journal of Memory and Language, 81,* 105-120.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22,* 1359-1366.

Simonsohn, U. (2014). *No-way interactions* [Blog post]. Retrieved from http://datacolada.org/17

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science, 3,* 160384.

*Smith, A. D., Hood, B. M., & Gilchrist, I. D. (2010). Probabilistic cuing in large-scale environmental search. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 605-618.

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. *American Journal of Psychology, 15,* 72-101.

*Stankevich, B. A., & Geng, J. J. (2014). Reward associations and spatial probabilities produce additive effects on attentional selection. *Attention, Perception, & Psychophysics, 76,* 2315-2325.

*Twedell, E. L., Koutstaal, W., & Jiang, Y. V. (2017). Aging affects the balance between goal-guided and habitual spatial attention. *Psychonomic Bulletin & Review, 24,* 1135-1141.

*Umemoto, A., Scolari, M., Vogel, E. K., & Awh, E. (2010). Statistical learning induces discrete shifts in the allocation of working memory resources. *Journal of Experimental Psychology: Human Perception and Performance, 36,* 1419-1429.

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

van Lamsweerde, A. E., & Beck, M. R. (2011). The change probability effect: Incidental learning, adaptability, and shared visual working memory resources. *Consciousness and Cognition, 20,* 1676-1689.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36,* 1-48.

Walthew, C., & Gilchrist, I. D. (2006). Target location probability effects in visual search: An effect of sequential dependencies. *Journal of Experimental Psychology: Human Perception and Performance, 32,* 1294-1301.

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.

*Won, B.-Y., & Jiang, Y. V. (2015). Spatial working memory interferes with explicit, but not probabilistic cuing of spatial attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41,* 787-806.

*Won, B.-Y., Lee, H. J., & Jiang, Y. V. (2015). Statistical learning modulates the direction of the first head movement in a large-scale search task. *Attention, Perception, & Psychophysics, 77,* 2229-2239.

Autor note

All data and materials related to this study are publicly available on The Open Science Framework (https://osf.io/xup9t/). Parts of the present research were presented at the joint meeting of the Experimental Psychology Society and the Spanish Society for Experimental Psychology (Oxford, UK, July 2016). MV was supported by grants 2016-T1/SOC-1395 from Comunidad de Madrid (Programa de Atracción de Talento Investigador) and PSI2017-85159-P from Agencia Estatal de Investigación, Ministerio de Economía y Competitividad. DS and MV were supported by grant ES/P009522/1 from the Economic and Social Research Council. 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

Table 1. List of effect sizes and study characteristics

Study code	N_{valid}	$N_{correct}$	# Regions	Test stage?	Learning?	Setting	# Guess	*h*	*v*
DA10.1	12	10	2	No	Yes	Computer	All	0.73	0.08
SHG10.2	2	2	2	No	Yes	Natural	Subset	1.57	0.50
SHG10.5	8	7	2	No	Yes	Natural	Subset	0.85	0.13
USVA10.1A	7	2	4	No	Yes	Computer	Subset	0.08	0.14
USVA10.1B	15	1	4	No	Yes	Computer	Subset	-0.52	0.07
USVA10.1C	8	2	4	No	Yes	Computer	Subset	0.00	0.13
USVA10.2	20	12	4	No	Yes	Computer	All	0.72	0.05
USVA10.3	16	11	4	No	Yes	Computer	All	0.91	0.06
JCES13.asd	14	4	4	Yes	No	Computer	All	0.08	0.07
JCES13.ctrl	15	6	4	Yes	No	Computer	All	0.32	0.07
JS13.1	24	7	4	Yes	Yes	Computer	All	0.09	0.04
JS13.2	12	2	4	Yes	Yes	Computer	All	-0.21	0.08
JS13.3	15	5	4	Yes	Yes	Computer	All	0.18	0.07
JSC13.3	16	3	4	Yes	No	Computer	All	-0.15	0.06
JSR13.1	6	3	4	Yes	Yes	Computer	Subset	0.52	0.17
JSRH13.1	1	1	4	Yes	Yes	Computer	Subset	2.09	1.00
JSRH13.2	3	3	4	Yes	Yes	Computer	Subset	2.09	0.33
JSRH13.3	5	2	4	Yes	Yes	Computer	Subset	0.32	0.20
JSRH13.4	8	4	4	Yes	Yes	Computer	All	0.52	0.13
JS14.4A	16	3	4	Yes	Yes	Computer	All	-0.15	0.06
JS14.4B	16	3	4	Yes	Yes	Computer	All	-0.15	0.06

JSS14.1	7	1	4	Yes	Yes	Computer	All	-0.27	0.14
JSS14.2	13	1	4	Yes	Yes	Computer	All	-0.49	0.08
JWS14.1	12	3	4	Yes	Yes	Computer	All	0.00	0.08
JWSM14.1	16	11	4	No	Yes	Natural	All	0.91	0.06
JWSM14.2	16	14	4	No	Yes	Natural	All	1.37	0.06
JWSM14.3	16	5	4	No	Yes	Natural	All	0.14	0.06
SG14.1	12	6	2	No	Yes	Computer	All	0.00	0.08
JSR15.1A	16	6	4	Yes	Yes	Computer	All	0.27	0.06
JSR15.1B	16	6	4	Yes	Yes	Computer	All	0.27	0.06
JSWC15.1fs	48	13	4	Yes	No	Computer	All	0.05	0.02
JW15.1	15	10	4	Yes	No	Computer	All	0.86	0.07
JW15.3	16	3	4	No	Yes	Computer	All	-0.15	0.06
JW15.4	16	6	4	Yes	Yes	Computer	All	0.27	0.06
WJ15.1	18	5	4	No	Yes	Computer	All	0.06	0.06
WLJ15	16	13	4	Yes	Yes	Natural	All	1.10	0.06
JKT16.1y	16	5	4	Yes	Yes	Computer	All	0.14	0.06
JKT16.1o	16	6	4	Yes	Yes	Computer	All	0.27	0.06
JKT16.2y	24	16	4	Yes	Yes	Computer	All	0.86	0.04
JKT16.2o	24	8	4	Yes	Yes	Computer	All	0.18	0.04
ATRJ17.3	48	22	4	Yes	Yes	Computer	All	0.44	0.02
TKJ17.y	24	15	4	No	Yes	Computer	All	0.78	0.04
TKJ17.o	24	17	4	No	Yes	Computer	All	0.95	0.04
SRJ17	40	15	4	Yes	Yes	Computer	All	0.27	0.03

Note: In the leftmost column studies are coded according to the first letters of the names of the authors, followed by the year of publication. The numbers or characters that follow the period denote the experiment number or condition name. $\mathbf{N_{valid}}$ refers to how many participants were invited to guess which region of the display contained the target most frequently. $\mathbf{N_{correct}}$ refers to how many of them gave the correct response. **# Regions** refers to the number of different regions into which the search display was divided (2 for search displays divided into two halves or 4 for search displays divided into quadrants). **Test stage?** codes whether the experiment included an unbiased testing stage before the awareness test. **Learning?** codes whether there was clear (i.e., statistically significant) evidence of learning in the stage immediately preceding the awareness test. **Setting** denotes whether participants searched for the target on a computer display (or a similar device) or in a natural large-scale setting (e.g., a room or a park). **# Guess** codes whether all participants were invited to guess which region contained the target most frequently or only some (usually depending on their answer to a previous question). Finally, ***h*** and ***v*** are the effect size and variance computed for the meta-analytic synthesis.

Table 2. Results of moderation analyses

Moderator / Sub-group	***h***	**LL**	**UL**	***z***	***p***	***k***	***Q***	***df***	***p***
Unbiased test stage before the awareness test							2.37	1	.124
Without test stage***	0.50	0.22	0.77	3.56	<.001	16			
With test stage***	0.26	0.11	0.41	3.41	<.001	28			
Number of regions in the search display							1.12	1	.290
Two*	0.63	0.09	1.17	2.31	.021	4			
Four***	0.33	0.18	0.48	4.33	<.001	40			
Evidence of learning before the awareness test							0.44	1	.506
No	0.22	-0.11	0.54	1.30	.193	5			
Yes***	0.37	0.21	0.53	4.61	<.001	39			
*Experimental setting****							11.37	1	<.001
Computer***	0.26	0.13	0.40	3.88	<.001	38			
Large-scale setting***	0.94	0.53	1.35	4.48	<.001	6			
Number of participants guessing							0.45	1	.503
All***	0.33	0.19	0.48	4.46	<.001	35			
Subset*	0.59	0.03	1.15	2.07	.038	9			

Note: *h* = effect size. LL = lower limit of the 95% CI; UL = upper limit of the 95% CI; *z* = *z*-score associated with the *h* value in the same row; *p* = *p*-value associated with the *z*-score in the same row; *k* = number of effect sizes contributing to *g* in the same row; *Q* = result of the *Q*-test for moderation; *df* = degrees of freedom of the *Q*-test for moderation; *p* = *p*-value of the *Q*-test for moderation. * $p < .05$, ** $p < .01$, *** $p < .001$

Table 3. Relationship between moderators

	Number of regions in the search display		Evidence of learning before the awareness test		Experimental setting		Number of participants guessing	
	χ^2	*p*	χ^2	*p*	χ^2	*p*	χ^2	*p*
Unbiased test stage before the awareness test	**4.97**	**.026**	1.69	.193	**4.48**	**.034**	0.91	.340
Number of regions in the search display	-		< 0.01	.999	2.13	.145	0.79	.375
Evidence of learning before the awareness test	-		-		0.06	.801	0.38	538
Experimental setting	-		-		-		0.09	.767
Number of participants guessing	-		-		-		-	

Note: All χ^2 tests for independence had one degree of freedom. Bold characters denote statistically significant results at $\alpha = .05$.

Table 4. Design summary of Experiment 2

Group	**Stage 1**		**Stage 2**		**Stage 3**		**Stage 4**		**Final Test**
1	Training	-	Training	-	Training	-	Training	Betting test	Guessing Test
2	Training	-	Training	-	Training	Betting test	Training	Betting test	Guessing Test
3	Training	-	Training	Betting test	Training	Betting test	Training	Betting test	Guessing Test
4	Training	Betting test	Training	Betting test	Training	Betting test	Training	Betting test	Guessing Test

Table 5. Best fitting parameters to data from Experiment 2

	ω_R	γ_R	a	b	p	σ	τ
Model 1	.28	-	-129.15	222.01	78.19	105.23	493.88
	(.05)		(1,232.61)	(315.82)	(72.28)	(58.31)	(153.40)
Model 2A	.29	-	-8.44	191.59	81.05	105.80	494.22
	(.09)		(324.72)	(89.25)	(79.27)	(63.99)	(155.90)
Model 2B	.30	.29	6.61	188.37	75.43	110.71	492.56
	(.11)	(.10)	(353.88)	(95.88)	(69.68)	(59.49)	(164.23)

Note: Mean (and standard deviation) of best fitting parameters for Models 1, 2A, and 2B across participants.

Table 6. Performance of Models 1, 2A, and 2B

	Correlation between predicted and observed data						
	RT rich, repetition	**RT rich, no repetition**	**RT sparse, repetition**	**RT sparse, no repetition**	**Proportion of bets**	**∑AIC**	**N_{best}**
Model 1	.95	.95	.89	.98	.39	**352,056.00**	**60**
Model 2A	.95	.95	.89	.98	-	352,175.40	47
Model 2B	.94	.94	.87	.98	1.00	352,267.30	27

Note: Correlation between observed data and model predictions across participants. The first four columns report the correlation between each participant's mean RT in four conditions and the mean RT predicted for each of those participants by Models 1, 2A, and 2B with best fitting parameters. The fifth column reports the correlation between the proportion of bets to the rich quadrant made by each participant and the predictions made by each model with best fitting parameters. **∑AIC** denotes the sum of AICs across participants. **N_{best}** denotes the number of participants with lower AICs for that model than for the alternative models. Bold characters denote the lowest AICs and highest N_{best}.

Figure Captions

Figure 1. Results of Experiment 1. Panel A depicts reaction times across Epochs in the training stage, separately for trials in which the target appeared in the rich and sparse quadrants. The dotted lines denote reaction times excluding inter-trial repetitions of target quadrant. Panel B represents the number of participants ranking the rich quadrant first, second, third and fourth in the awareness test.

Figure 2. Results of Experiment 2: Probabilistic cuing. Each panel presents reaction times across Epochs in the training stage, separately for each experimental condition. The dotted lines denote reaction times excluding inter-trial repetitions of target quadrant.

Figure 3. Results of Experiment 2: Awareness tests. Panel A depicts the average proportion of bets to the rich quadrant in the betting tests, separately for each group and stage. Panel B shows the proportion of participants selecting the rich quadrant in the final awareness test. All error bars denote 95% CIs.

Figure 4. Updated meta-analysis. The top row reports the meta-analytic estimate and 95% confidence interval of the initial meta-analysis. The following rows denote the effect sizes and confidence intervals of awareness in the present series of experiments. The final rows represent the meta-analysis of the current studies and the combination of the present studies and the initial meta-analysis.

Figure 5. Correlation between learning and awareness. Panels A and B show the correlation between probabilistic cuing and performance in the awareness test in Experiments 1 and 2, respectively. Panels C and D show two simulations of the correlations predicted by Model 1 for Experiments 1 and 2, respectively. The bell-shaped curve in Panels E and F shows the distribution of correlations predicted by Model 1 for Experiments 1 and 2 over 10,000 iterations. The diamond denotes the mean and 95% CI of the correlation observed in the empirical data.

Appendix A

Characteristics of Excluded Studies and Reasons for Exclusion

Table A1. Excluded studies, in chronological order

Study	Reason for exclusion
Shaw & Shaw (1977)	Participants were explicitly informed about the probabilities of the target appearing in each region (Criterion 3) and there was no awareness test (Criterion 2).
Shaw (1978)	Participants were explicitly informed about the probabilities of the target appearing in each region (Criterion 3) and there was no awareness test (Criterion 2).
Rabitt, Cumming, & Vyas (1979)	Not a probabilistic cuing task (Criterion 1).
Müller & Findlay (1987)	On every trial, an arrow was presented to indicate the likely target position (Criterion 3) and there was no awareness test (Criterion 2).
Miller (1988)	No awareness test (Criterion 2).
Hoffmann & Kunde (1999)	Awareness was assessed with an unstructured interview (Criterion 2).
Geng & Behrmann (2002)	Participants were asked whether the target was equally likely to appear in all regions, but they were not asked to guess what was the rich region (Criterion 2).
Reder, Weber, Shang, & Vanyukov (2003)	Not a probabilistic cuing task (Criterion 1). Although the study included an awareness test, little information is reported (Criterion 2).
Geng & Behrmann (2005)	Participants were asked whether they had realised that one location was more likely to contain the target than the others, but they were not asked to guess what was the rich region (Criterion 2).
Walthew & Gilchrist (2006)	No awareness test (Criterion 2).
Beck, Angelone, Levin, Peterson, & Varakin (2008)	Experiments 1-3 and 5 do not use a probabilistic cuing task (Criterion 1). In Experiment 4 (and in the implicit condition of Experiment 6) participants rated regions instead of selecting one (Criterion 2). Beck (personal communication, May 27th 2017) reports that most participants gave equal ratings to all regions, possibly for consistency with Question 2 of the awareness test ("Do you think that any of the changes were more likely to occur in certain locations?"), where most participants replied "No". The explicit condition of Experiment 6 has the same shortcomings and,

	additionally, participants receive explicit information about the location of targets (Criterion 3).
Fecteau, Korjoukov, & Roelfsema (2009)	Details of the awareness test are missing (Criterion 2).
Druker & Anderson (2010), Experiment 2	Details of the awareness test are missing (Criterion 2). Anderson (personal communication, May 26th, 2017) has confirmed that participants in this experiment were not asked to select a region.
Liu, Chiau, Tseng, Hung, Tzeng, Muggleton, & Juan (2010)	Experiment 1 is not a probabilistic cuing task (Criterion 1). Experiments 2 and 3 apparently included an awareness test, but the authors did not report detailed information about the procedure or results (Criterion 2).
Smith, Hood, & Gilchrist (2010), Experiments 1, 3, 4, 6	There was no evidence of cuing in Experiments 1, 3, and 4 (Criterion 4). Experiment 6 is not a probabilistic cuing task (Criterion 1), given that colors (instead of sides or quadrants) are probabilistically related to the target.
Pellicano, Smith, Cristino, Hood, Briscoe, & Gilchrist (2011)	No awareness test (Criterion 2).
van Lamsweerde & Beck (2011)	Not a probabilistic cuing task (Criterion 1).
Jones & Kaschak (2012)	No awareness test (Criterion 2).
Kabata & Matsumoto (2012)	No awareness test (Criterion 2).
Shaqiri & Anderson (2012)	No awareness test (Criterion 2).
Chukoskie, Snider, Mozer, Krauzlis, & Sejnowski (2013)	No awareness test (Criterion 2).
Jiang, Swallow, & Capistrano (2013), Experiment 4	There was no evidence of cuing in Experiments 1 and 2 (Criterion 4). In Experiment 4, participants were explicitly told that all locations are equally likely at test (Criterion 3).
Jiang, Swallow, & Rosenbaum (2013), Experiments 2-5	Experiment 2 was not a probabilistic cuing task (Criterion 1). In Experiments 3-5 there was an endogenous cue directing participants' attention to specific quadrants at different stages, depending on the experiment (Criterion 3).
Goschy, Bakos, Müller, & Zehetleitner (2014)	Not a probabilistic cuing task (Criterion 1). The target was equally likely in all regions.
Jiang & Swallow (2014), Experiments 1A, 1B, 2, 3	There was no evidence of cuing in Experiments 1A, 1B, and 2 (Criterion 4). Experiment 3 did not use a probabilistic cuing task (Criterion 1).
Jiang, Swallow, & Sun (2014), Experiments 3-5	Participants were explicitly told where to find the target (Criterion 3). For instance, they were instructed to give equal priority to all

	quadrants, or to the sparse quadrants, where the target was not most likely to appear.
Jiang, Won, & Swallow (2014), Experiments 2-3	Participants were given explicit information about the likelihood that the target would appear in each quadrant in different stages (Criterion 3) and either there was no awareness test or its results were not reported (Criterion 2).
Kabata, Yokoyama, Noguchi, & Kita (2014)	No awareness test (Criterion 2).
Stankevich & Geng (2014), Experiment 2	Not a probabilistic cuing task (Criterion 1).
Jiang, Sha, & Remington (2015), Experiments 2-4	Not a probabilistic cuing task (Criterion 1).
Jiang, Swallow, Won, Cistera, & Rosenbaum (2015), Experiments 2-5	Awareness data were only reported for Experiment 1 (Criterion 2).
Jiang & Won (2015), Experiments 2, 5, 6	There was no evidence of cuing in Experiments 2, 5, and 6 (Criterion 4).
Won & Jiang (2015), Experiments 2-5	Unlike the rest of the experiments in the meta-analysis, in Experiments 2 and 3 there were two rich quadrants, one for load and one for the no-load trials (Criterion 2). Experiments 4 and 5 did not use a probability cuing task (Criterion 1).
Chua & Gauthier (2016)	Experiments 1 and 2 did not include an awareness test (Criterion 2). Experiment 3 included an awareness test that differs substantially from the rest of the experiments included in the meta-analysis (Criterion 2).
Chang, Little, & Yang (2016)	Experiments 2 and 3 did not include an awareness test (Criterion 2) and there was an obvious explicit component (Criterion 3). For instance, one of the four participants is an author of the paper and in Experiment 3 all participants were informed about the probabilities.
Addleman, Tao, Remington, & Jiang (2018)	Participants were explicitly informed about the probabilities of the target appearing in each region (Criterion 3) and there was no awareness test (Criterion 2).

Appendix B

Model Recovery Analysis

The analyses reported in the main text of the article rely on the Akaike Information Criterion (AIC) to compare the performance of Models 1, 2A, and 2B. This approach assumes that the model with the lowest AIC is more likely to be the true model that generated the empirical data. However, it is not impossible that data generated by one model are better fitted by an alternative model due to sampling error or to the relative flexibility of each model. Needless to say, our conclusions depend heavily on the ability of our procedure to recover the true model that generated the data. To confirm the validity of our analyses, we generated data from Models 1, 2A, and 2B using the best-fitting parameters of each participant in Experiment 2 and then we fitted all three models to the generated datasets. This allowed us to assess whether the true model was more likely to be selected than the alternative two models.

For each participant, we generated the same number of data points (i.e., reaction times in the rich and sparse conditions, with and without repetitions, and number of bets in the awareness test) that had been used to fit the models for that participant. Furthermore, to reduce the impact of sampling error, we generated 20 complete data sets for each participant and we fitted the three models to all of them using exactly the same procedure as in the analyses reported in the main text. For each participant, we registered (a) the average AIC for each model across the 20 data sets and (b) the proportion of times (out of 20) that each model yielded the lowest AIC compared to the other two models.

Table B1 shows the results of the model recovery analysis across participants. As can be seen, when data were generated by Models 1 and 2A, both the sum (across participants) and the mean proportion (across participants) of times each model was selected favoured the true model. In contrast, when data were generated by Model 2B, the true model was not

more likely to be selected than Models 1 and 2A. This is a natural consequence of the fact that Models 1 and 2A are special cases of Model 2B. When $\gamma_R = \omega_R$, Model 2B generates data indistinguishable from Model 1 and when $\gamma_R = .25$ it generates data indistinguishable from Model 2A. Because Model 2B includes an additional parameter and AICs put a penalty on the number of free parameters, under those conditions Models 1 and 2A are systematically preferred over Model 2B. Therefore, we repeated our recovery analysis of Model 2B excluding all participants for whom γ_R was very similar to ω_R or to .25. Specifically, we removed all participants for whom $|\gamma_R - \omega_R| > .05$ or γ_R was within the [.2, .3] interval. The final row in Table B1 reports the results of the recovery analysis with the subset of 66 participants who met these conditions. As can be seen, the true model was successfully recovered within this sample.

Table B1. Results of the Model Recovery Analysis

	Sum of Average AIC			**Mean Proportion N_{best}**		
True Model	**Model 1**	**Model 2A**	**Model 2B**	**Model 1**	**Model 2A**	**Model 2B**
Model 1	**351,010.30**	351,082.40	351,237.70	**.49**	.39	.12
Model 2A	350,702.76	**350,669.02**	350,914.93	.38	**.51**	.11
Model 2B	**351,402.58**	351,500.74	351,367.24	**.37**	.33	.30
Model $2B_{66}$	174,478.50	174,588.23	**174,345.29**	.31	.20	**.49**

Note: The analysis reported in the final row, labeled as Model $2B_{66}$, retained only 66 participants for whom the best fitting γ_R was out of the [.2, .3] interval and $|\gamma_R - \omega_R| > .05$. Bold characters denote the lowest sum of AICs and the highest proportion of N_{best} in each row.

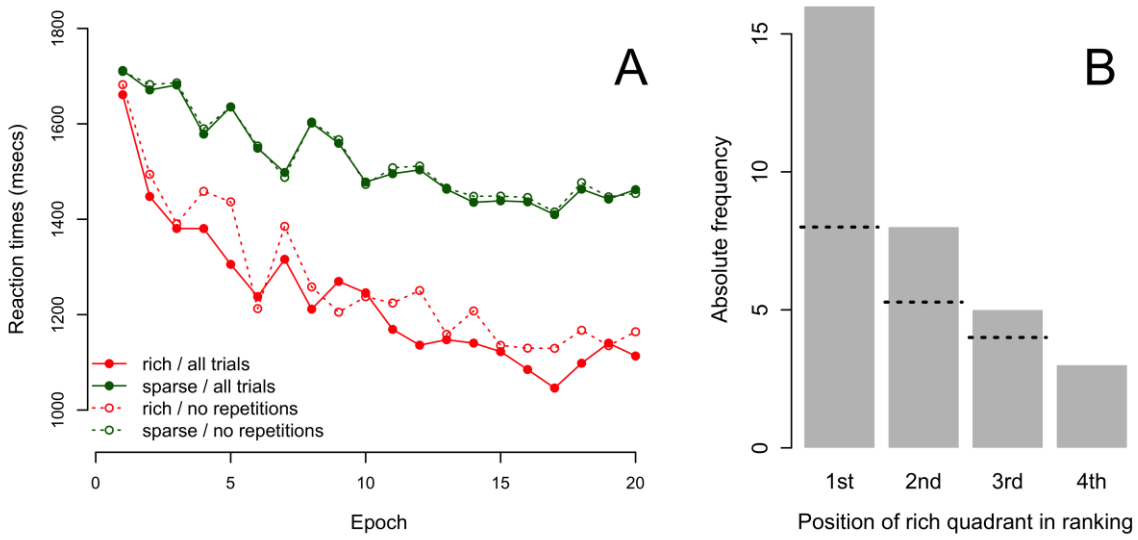
A
Reaction times (msecs)
1000
1200
1400
1600
1800
0
5
10
15
20
Epoch
rich / all trials
sparse / all trials
rich / no repetitions
sparse / no repetitions
B
Absolute frequency
0
5
10
15
1st
2nd
3rd
4th
Position of rich quadrant in ranking

Figure #1

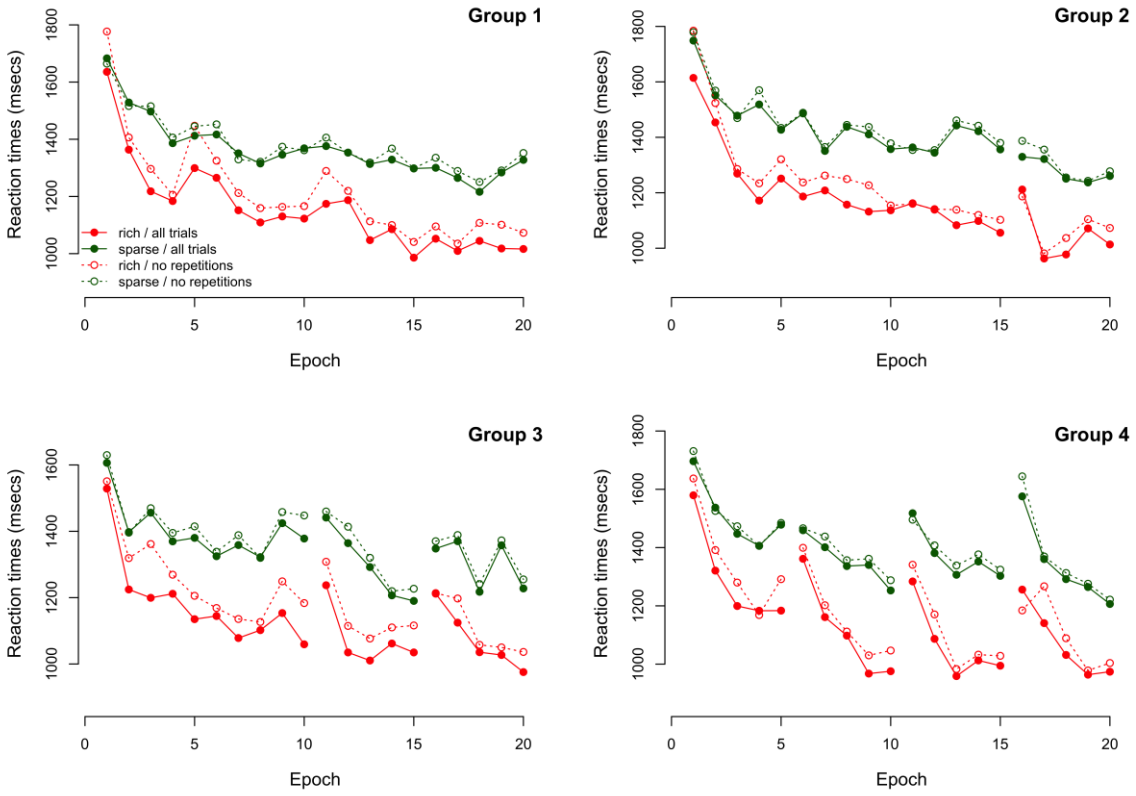
Group 1
Group 2
Group 3
Group 4
Reaction times (msecs)
Epoch
rich / all trials
sparse / all trials
rich / no repetitions
sparse / no repetitions

Figure #2

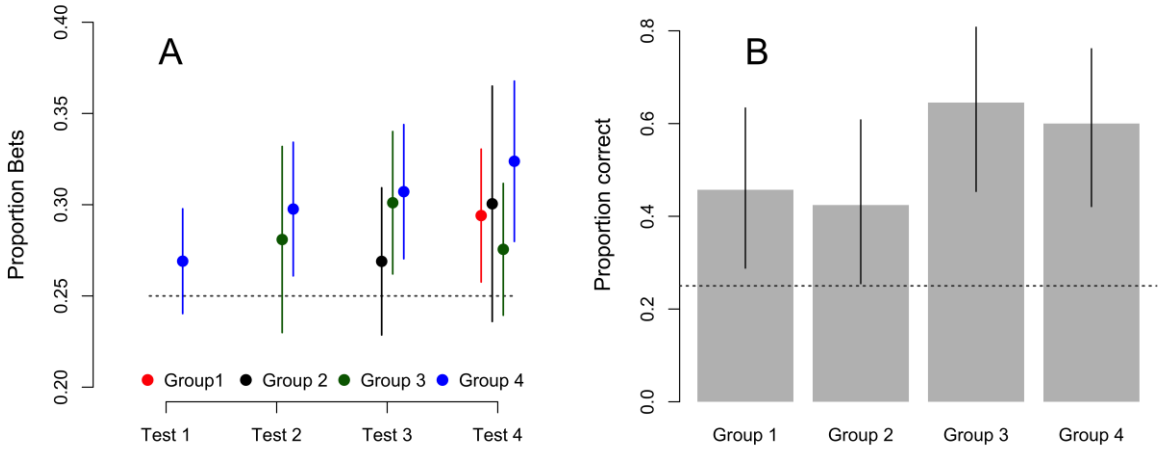
A
Proportion Bets
0.20
0.25
0.30
0.35
0.40
Group1
Group 2
Group 3
Group 4
Test 1
Test 2
Test 3
Test 4
B
Proportion correct
0.0
0.2
0.4
0.6
0.8
Group 1
Group 2
Group 3
Group 4

Figure #3

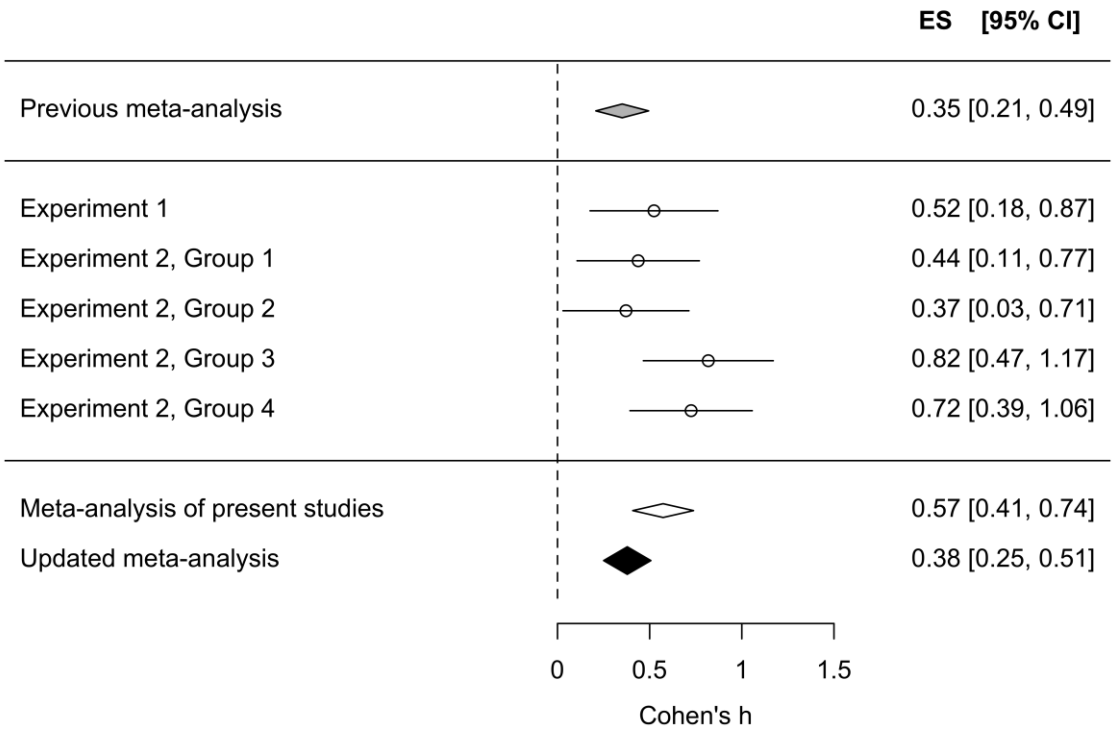
ES [95% CI]
Previous meta-analysis
0.35 [0.21, 0.49]
Experiment 1
0.52 [0.18, 0.87]
Experiment 2, Group 1
0.44 [0.11, 0.77]
Experiment 2, Group 2
0.37 [0.03, 0.71]
Experiment 2, Group 3
0.82 [0.47, 1.17]
Experiment 2, Group 4
0.72 [0.39, 1.06]
Meta-analysis of present studies
0.57 [0.41, 0.74]
Updated meta-analysis
0.38 [0.25, 0.51]
0
0.5
1
1.5
Cohen's h

Figure #4

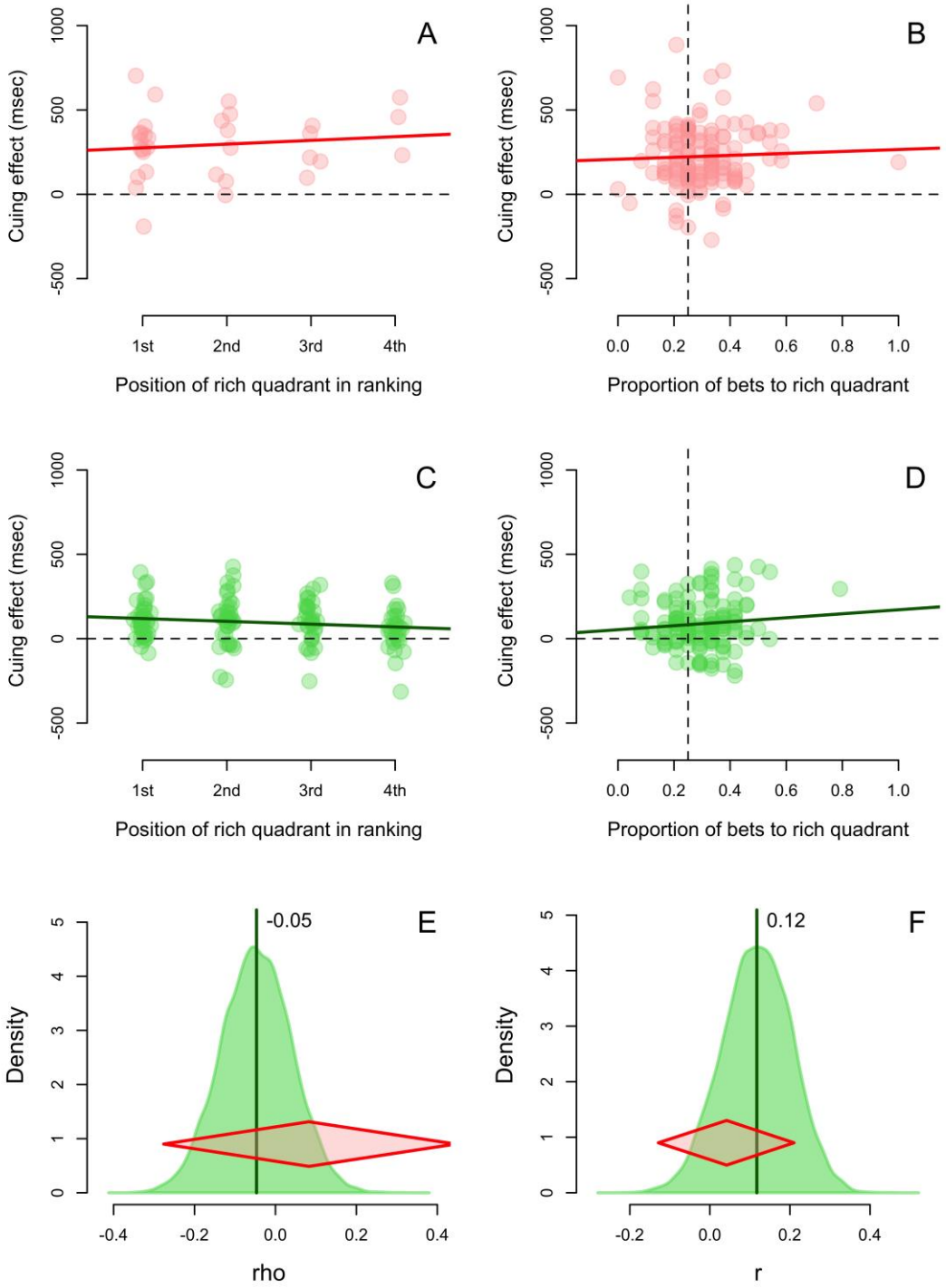
A
B
C
D
E
F
Cuing effect (msec)
1000
500
0
-500
1st
2nd
3rd
4th
Position of rich quadrant in ranking
0.0
0.2
0.4
0.6
0.8
1.0
Proportion of bets to rich quadrant
Density
0
1
2
3
4
5
-0.05
-0.4
-0.2
0.0
0.2
0.4
rho
0.12
r

Figure #5