Psychonomic Bulletin & Review (2021) 28:1848–1859
https://doi.org/10.3758/s13423-021-01901-4

THEORETICAL/REVIEW

Avoiding pitfalls: Bayes factors can be a reliable tool for post hoc data selection in implicit learning

M. Leganes-Fonteneau[1,2] · R. Scott[3,4] · T. Duka[3,5] · Z. Dienes[3]

Accepted: 10 February 2021 / Published online: 25 March 2021

Abstract

Research on implicit processes has revealed problems with awareness categorizations based on nonsignificant results. Moreover, post hoc categorizations result in regression to the mean (RTM), by which aware participants are wrongly categorized as unaware. Using Bayes factors to obtain sensitive evidence for participants' lack of knowledge may deal with nonsignificance being nonevidential, but also may prevent regression-to-the-mean effects. Here, we examine the reliability of a novel Bayesian awareness categorization procedure. Participants completed a reward learning task followed by a flanker task measuring attention towards conditioned stimuli. They were categorized as *B*_Aware and *B*_Unaware of stimulus–outcome contingencies, and those with insensitive Bayes factors were deemed *B*_Insensitive. We found that performance for *B*_Unaware participants was below chance level using unbiased tests. This was further confirmed using a resampling procedure with multiple iterations, contrary to the prediction of RTM effects. Conversely, when categorizing participants using *t* tests, *t*_Unaware participants showed RTM effects. We also propose a group boundary optimization procedure to determine the threshold at which regression to the mean is observed. Using Bayes factors instead of *t* tests as a post hoc categorization tool allows evaluating evidence of unawareness, which in turn helps avoid RTM. The reliability of the Bayesian awareness categorization procedure strengthens previous evidence for implicit reward conditioning. The toolbox used for the categorization procedure is detailed and made available. Post hoc group selection can provide evidence for implicit processes; the relevance of RTM needs to be considered for each study and cannot simply be assumed to be a problem.

Keywords Implicit learning and memory · Bayesian statistics

Introduction

Implicit processes take place when knowledge that the subject is unaware of possessing guides his or her responses. Such processes have been proposed in different cognitive domains, such as attention (Chun & Jiang, 1998), artificial grammar learning (Reber, 1989; Scott & Dienes, 2010), and conditioning (Hofmann, De Houwer, Perugini, Baeyens, & Crombez, 2010). However, the validity and reliability of measures and techniques used to assess the awareness of knowledge have been criticized (e.g., Mertens & Engelhard, 2020; Rebuschat, 2013). In this short report, we aim to scrutinize a previously published measure of implicit learning (Leganes-Fonteneau, Nikolaou, Scott, & Duka, 2019; Leganes-Fonteneau, Scott, & Duka, 2018) to assess its reliability, particularly regarding regression-to-the-mean effects. As described by Shanks (2017), measurement error in the determination of awareness scores may yield incorrect classifications of participants as unaware, potentially inflating evidence for the existence of implicit processes. By using Bayes factors to perform an awareness categorization we seek to circumvent regression-to-the-mean effects.

Conditioning tasks consist of the pairing of neutral stimuli with a rewarding or aversive outcome according to a contingency schedule. After repeated associations, conditioned

✉ M. Leganes-Fonteneau
mateoleganes@gmail.com

[1] Department of Kinesiology and Health, Rutgers University–New Brunswick, New Brunswick, NJ, USA

[2] Cardiac Neuroscience Laboratory, Center of Alcohol Studies, Rutgers University–New Brunswick, New Brunswick, NJ, USA

[3] School of Psychology, University of Sussex, Falmer BN1 9QH, UK

[4] Sackler Centre for Consciousness Science, University of Sussex, Brighton BN1 9RH, UK

[5] Sussex Addiction Research and Intervention Centre (SARIC), University of Sussex, Brighton, UK

stimuli (CS) obtain the properties of the outcome and can elicit a variety of behavioral and cognitive responses. In these tasks, a measure of awareness, the ability to predict the outcome associated with a CS, can be obtained, together with a measure of performance (e.g., reaction times) towards the CS. Implicit conditioning occurs when participants who are not aware of the contingencies still perform accurately on a learning test.

Experiments examining implicitly conditioned responses have been the target of methodological criticism for decades (Lovibond & Shanks, 2002; Mertens, Basci, & Engelhard, in press). A crucial limitation of implicit cognitive processing research in general, and implicit conditioning experiments in particular, is that researchers frequently rely on nonsignificant results when asserting that participants are unaware of contingencies (Dienes, 2015; Vadillo, Konstantinidis, & Shanks, 2016), thus accepting the null hypothesis that participants have not developed any awareness of their knowledge. Customarily, *t* tests are used to categorize participants as *t*_Aware if they score significantly above chance level, or *t*_Unaware if their accuracy in the detection of contingencies is either significantly below chance or is not significantly different from chance. But it is logically invalid to conclude that awareness is absent solely from a *t* test revealing a nonsignificant difference from chance. Nonsignificance by itself may arise because the data provide evidence for no awareness, or because there is too much noise in the data to conclude anything (Dienes, 2014).

Bayes factors allow examining evidence for the null hypothesis (Dienes, 2014) by quantifying how well the null hypothesis predicts the empirical data relative to a competing hypothesis. Such analyses can be applied to the study of unconscious processes (Dienes, 2015; Sand & Nilsson, 2016), providing evidence for the absence of knowledge. In the context of reward conditioning, a novel Bayesian awareness categorization technique (BACT) was recently developed, using signal detection theory and Type I and Type II measures of contingency awareness to sensitively determine the awareness state of each participant (Leganes-Fonteneau et al., 2018). This method makes it possible to categorize participants as Bayesian *B*_Aware or *B*_Unaware of the contingencies. Importantly, a third category, *B*_Insensitive, is available for participants for whom the result of the Bayes factor on contingency awareness is ambiguous and insensitive.

Using this methodology, it was possible to find participants for whom there was sensitive evidence for the absence of awareness of the relationship between CS and monetary rewards (*B*_Unaware). These participants, however, still showed significant performance in a testing phase, as the CS generated preferential attention in an emotional attentional blink task (Leganes-Fonteneau et al., 2018), and cognitive interference in a modified flanker task (Leganes-Fonteneau et al., 2019). This provided initial evidence for implicit learning effects.

However, as much as novel methodologies can be developed to sensitively determine the absence of awareness in implicit paradigms, a further crucial limitation needs to be addressed. Shanks (2017) showed how performing post hoc categorizations of participants in different groups according to their level of awareness is a form of extreme group selection. Measured awareness scores are composed of a true score for each case and a random error with a mean of zero. However, as shown in simulation studies (León & Suero, 2000; Shanks, 2017), randomly selecting cases with a highly positive (or negative) score increases the likelihood that this score is in fact due to a larger positively (or negatively) biased measurement error. For that reason, when groupings of cases with extreme values are created, the mean measurement error of each group will be different from zero. The estimated mean score of that group will not reflect actual true scores, but rather the biased distribution of errors within that group.

Because of this "statistical inevitability" (Shanks, 2017), some participants deemed unaware will have been added to this group because of larger negatively biased measurement error rather than an actual null true score; thus, as a group, the selected subjects will in reality have greater awareness than indicated by the mean of the very variable they were selected with. That is, their actual awareness will be closer to the mean awareness of all participants ("regression to the mean"), and this awareness may then be responsible for learning (e.g., attentional responses), wrongly providing evidence for implicit learning.

On that basis, Shanks (2017) also rightly predicted that subsequent tests of awareness would regress to the mean. If participants were categorized as Unaware on an initial Test 1, and their awareness state tested again on Test 2, then both tests would yield roughly similar scores only if they had a random unbiased error. However, because Test 1 generates a categorization based on biased measurement error, Shanks found for the data he considered that on a Test 2 participants' awareness scores regressed to the mean. Additionally, in his example, a reliability analysis showed a poor test–retest reliability of the categorization, implying a large amount of measurement error in the categorization tool. Note the problem is a specific form of selective inference (Davenport & Nichols, 2020; Leeb & Potscher, 2006), an issue we take up in the Discussion.

Therefore, in order to obtain strong evidence for the absence of awareness in participants, it is useful to examine the results of an awareness test over two different measurement phases, verifying that participants do not score above chance level on the second unbiased test (i.e., their measured awareness does not regress to the mean). Moreover, as performance scores are typically computed as the aggregate mean of aware and unaware participants, it is necessary not only to verify that individual Test 1 and Test 2 measures are congruent, but also that the mean accuracy on Test 2 for unaware participants remains not above chance level.

Regression to the mean occurs because measurement includes random error. The less the error, the more reliable the measure, the less regression to the mean occurs. Bayes factors can help because they can make a three-way distinction: evidence for awareness, evidence for no awareness; and not much evidence either way. The threshold for evidence for no awareness can be adjusted in making this three-way distinction. For example, one can take a Bayes factor less than 1/3 as evidence for no awareness; if this was not strong enough, one could take less than 1/6, and so on (though this also involves a more extreme subgroup selection). In effect, measurement error, when large, can be partially hived off in the cases classified as showing not much evidence, reducing regression to the mean. By contrast, a *p* value does not indicate evidence for no effect; so, adjusting the threshold does not adjust the evidence for no effect in the same clear way as for Bayes factors. A nonsignificant result does not discriminate evidence for no awareness from no evidence. A $p > .09$ will include cases where there was not much evidence either way, as well as cases where there may have been good evidence for no awareness. In sum, a Bayes factor can indirectly increase the reliability of the measurement compared with a *t* test, by separating out cases with different amounts of measurement error. This procedure however, is not guaranteed to do so sufficiently; it depends on how Bayes factors are used. For example, Sand and Nilsson (2016) showed that subgroup selection using Bayes factors can still show sufficient regression to the mean to produce spurious evidence for implicit cognition. Nonetheless, the use of Bayes factors opens the possibility of regression to the mean not being a problem for implicit cognition research because the threshold for evidence can be adjusted.

This report focuses on the awareness measure collected in Leganes-Fonteneau et al. (2019), examining its susceptibility to the limitations identified by Shanks (2017); specifically, evidence of regression to the mean for post hoc awareness categorization. First, we will test whether the aggregate means of each awareness group according to the Bayes factors (precisely as used in Leganes-Fonteneau et al., 2018) are different from chance level. We will then further examine the Bayesian categorization using a resampling procedure on split-half random trial selection. This will also allow to compute a test–retest reliability analysis and odds ratio for the categorization. A group boundary optimization procedure will be used to examine the behavior of the BACT under different Bayes factor thresholds. Finally, we will test predictions for regression to the mean effects using *t* tests to categorize participants instead of Bayes factors.

Methods

Previous analysis

Conditioning task

For this analysis we used data published in Leganes-Fonteneau et al. (2019). In that experiment, participants (n = 49) completed a task-irrelevant reward conditioning paradigm. Participants were presented with two categories of geometrical shapes (octagons vs. squares) as conditioned stimuli (CS) associated with high (high reward–90%) or low (low reward–10%) probabilities of earning money. A green or yellow square was overlaid on top of the CS, and participants had to press a key depending on the color of the square (Yokoyama, Padmala, & Pessoa, 2015). Therefore, responses were irrelevant to the stimulus–reward contingencies.

On 50% of the conditioning trials, participants indicated whether they expected to get money or not (Yes/No) and their confidence about their response on a 5-point Likert scale (1 = *completely guessing*, 2 = *more or less guessing*, 3 = *fairly sure*, 4 = *almost certain*, 5 = *completely certain*).

Bayesian awareness categorization

The original Bayesian awareness categorization used in Leganes-Fonteneau et al. (2018) and (2019) was based on the use of Type I and Type II *d′* scores to obtain sensitive evidence for the lack of awareness. Because we were using a novel conditioning paradigm, we did not have a prior study to inform expectations of effect size for Type I awareness, and we therefore designed a methodology in which constraints defined by aspects of each participant's performance could be used as constraints for other aspects (see Dienes, 2015, 2019, for this general approach). In other cases, where prior studies exist or a norming study is done, Type I (or raw accuracy) scores from a prior study can be used to constrain Type I performance in order to model H1.

Type I refers to the ability to discriminate states of the world (e.g., octagons are followed by reward); Type II refers to the ability to discriminate the accuracy of one's knowledge (being more confident in accurate responses). A Bayes factor requires a model of the range of effects predicted (the model of H1). The model of H1 can be a uniform [0, max], indicating that the effect can be anything from 0 to a maximum. A Type II *d′* typically is not higher than the Type I *d′*; thus, the analysis utilizes participants' own Type I knowledge in the model of H1 for Type II *d′*; specifically, the Bayes factor for Type II *d′* ($B_{d2'}$) used the uniform [0, Type I *d′* for that participant]. Participants with $B_{d2'} < 1/3$ were categorized as Metacognitively Unaware, whereas those with a $B_{d2'} > 3$ were considered Metacognitively Aware, and the rest ($1/3<B_{d2'} <3$) as Insensitive.

Presuming that unconscious Type I knowledge would not be more than the Type I knowledge that is metacognitively conscious, the mean Type I score of participants with Type II knowledge (M1) was used as a maximum for a uniform Bayes factor to model H1. This model was used for each participant to determine the existence of contingency awareness, obtaining a Bayes factor for their Type I scores ($B_{d1'}$). Participants with $B_{d1'} < 1/3$ were categorized as *B*_Unaware of the contingencies, whereas those with a $B_{d1'} > 3$ were considered *B*_Aware, and the rest as *B*_Insensitive (see the

Appendix or Leganes-Fonteneau et al., 2018. for a detailed description of the Bayesian categorization procedure).

After the conditioning task, participants completed a modified flanker task (Nikolaou, Field, & Duka, 2013) in which CS acted as task-irrelevant distractors. It was found that high-reward CS generated a higher flanker effect than low reward CS and control trials. Importantly, this effect occurred only in participants who were categorized as Unaware of contingencies (see Leganes-Fonteneau et al., 2019, for a detailed description of the experimental procedures, analyses, and results). Figure 1 shows the awareness scores of each participant for each awareness group, plotted against their performance in the flanker task (flanker score for high reward minus low reward CS). For *B*_Aware participants, the mean flanker effect was −3.54 ms, *SD* = 35.934, and a Bayes factor (modelling H1 according to Nikolaou, Field, & Duka, 2013) showed the scores were insensitively different from 0, $B_{\mathrm{U}[0,\ 10]}$ = 1.0146. For *B*_Unaware participants, the mean flanker effect was 21.09 ms, *SD* = 30.69, $B_{\mathrm{U}[0,\ 10]}$ = 10.057. For *B*_Insensitive participants, the mean flanker effect was −7.84 ms, *SD* = 21.58, and the Bayes factor was insensitive, $B_{\mathrm{U}[0,\ 10]}$ = 1.93.

Current analysis

Bayesian categorization

The first step in the analysis was to categorize participants using the Bayesian categorization methodology described above. This yielded 16 *B*_Aware participants, with Type I knowledge of the contingencies; 20 *B*_Unaware, with sensitive evidence for their lack of contingency awareness; and 13 *B*_Insensitive for contingency awareness, for which there is little evidence about their conscious state in either direction. Using one-sample *t* tests, comparing awareness scores to chance level, we found that all *B*_Unaware and all *B*_Insensitive participants except one would be considered as *t*_Unaware, and the rest would be categorized as *t*_Aware.

Then, for *B*_Aware, *B*_Unaware, and *B*_Insensitive groups, we estimated their aggregated Type I awareness scores (proportion correct, where 0.5 is chance level). We report sample means and 95% confidence intervals, taken as 95% credibility intervals with uniform priors. For *B*_Aware participants, mean Type I awareness was 0.87, *SD* = 0.085, 95% CI [0.82, 0.91]. For *B*_Unaware participants, mean Type I awareness was 0.46, *SD* = 0.039, 95% CI [0.44, 0.48]. For *B*_Insensitive participants, mean Type I awareness was 0.55, *SD* = 0.043, 95% CI [0.53, 0.57].

Iteration analysis

A resampling procedure was performed to examine regression to the mean effects and the internal reliability of the categorization procedure. For each participant, 1,000 iterations were run. In each iteration, half of the response trials of that participant (X-half) were randomly selected, and the BACT was applied, basing M1 on the mean Type I score of Type II aware participants obtained in the categorization (2.373 over all trials and participants). Therefore, each iteration could be deemed

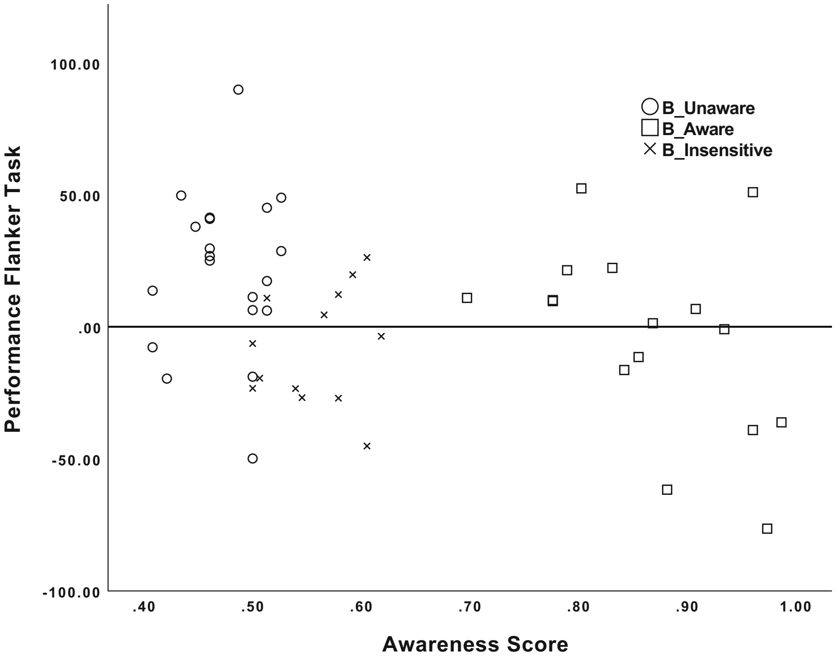

Fig. 1 Performance scores on the flanker task (high reward − low reward CS) and awareness score for each awareness group according to the Bayesian categorization. This graph shows how only participants in the *B*_Unaware group displayed significant responses towards high reward CS on the Flanker task measuring learning

*B*_Aware, *B*_Unaware, or *B*_Insensitive depending on the outcome of the Bayesian analysis. Next, the mean Type I awareness for the X-half and for the remaining half of the trials on that iteration (Y-half) were also computed and stored.

After the resampling procedure, mean cumulative X-half and Y-half Type I awareness scores for *B*_Aware, *B*_Unaware, and *B*_Insensitive iterations were computed, and the means were weighted according to the number of iterations for each categorization outcome. For *B*_Aware iterations, the mean Type I awareness on Y trials was 0.85, *SD* = 0.13, 95% CI [0.842, 0.846]. For *B*_Unaware iterations, the mean Type I awareness on Y trials was 0.498, *SD* = 0.082, 95% CI [0.497, 0.499]. For *B*_Insensitive iterations, the mean Type I awareness on Y trials was 0.524, *SD* = 0.086, 95% CI [0.523, 0.526]. Figure 2 presents density plots for X and Y halves of each *B*_Awareness group, showing how *B*_Unaware iterations on the X-half remain below chance level on the Y-half and how cumulative Type I awareness scores on the Y-half for *B*_Insensitive iterations are above chance level.

A split-half reliability analysis was performed, the odds ratio for the *B*_Aware/*B*_Unaware categorization across X/Y-halves was computed, *OR* = 298.31, 95% CI [11.31, 6983.20], as well as for the *B*_Unaware/*B*_Insensitive categorization, *OR* = 4.882, 95% CI [1.06, 22.41].

Group boundary optimization

For the main analysis, we chose a cutoff for the Bayes factor of 3-1/3 because it roughly corresponds to the amount of evidence reflected by the usual significance level used in psychological science, $p < .05$ (Dienes, 2014; Jeffreys, 1939). However, it is possible to adjust this cutoff in a group boundary optimization procedure to determine whether regression to the mean occurs at more stringent thresholds. In this procedure, the BACT produces the iteration analysis presented above starting with a cutoff of *B* = 6-1/6. If the iteration analysis yields regression-to-the-mean effects (i.e., cumulative awareness scores on the Y-halves above chance level), then the iteration analysis is repeated using a more liberal cutoff (i.e., *B* = 5.5-1/5.5) until no regression to the mean is observed. This allows obtaining an estimate of which criterion allows maximizing group categorization while still preventing the occurrence of regression to the mean.

In this case, at *B* = 6-1/6, we did find regression-to-the-mean effects, as awareness scores on Y-halves were above chance level. The most stringent cutoff generates a very low cumulative awareness score on *B*_Unaware X-halves, which contrasts with an above chance level mean on Y-halves. However, a cutoff of 5-1/5 did not generate regression to the mean. In fact, as Fig. 3 illustrates, applying progressively more lenient cutoffs for the Bayes factors decreases the difference between the cumulative scores on X and Y halves, and in fact smooths out the distribution of scores in both halves while increasing the *B*_Unaware/*B*_Insensitive odds ratio, signifying a higher reliability in the classification.

For the purpose of this reliability analysis, we applied the maximum cutoff for which no regression to the mean was observed (5-1/5) to the original categorization procedure. With this cutoff, 17 participants were deemed *B*_Unaware, 16 *B*_Aware, and 16 *B*_Insensitive. That is, because the threshold for sensitive/insensitive categorization was higher, unsurprisingly, three participants who would be deemed *B*_Unaware with a conventional 3-1/3 threshold were this time considered insensitive. Even with the ensuing decrease in degrees of freedom due to the lesser amount of *B*_Unaware participants, the learning measure in the flanker task was significant, *t*(16) = 2.118, *p* = .050.

Using *t* tests as a categorization tool

The resampling procedure was also performed categorizing participants on X-half trials using t-tests instead of Bayes factors. This corresponds to the customary procedure used in implicit learning literature, by which a participant is categorized as *t*_Aware if their performance is significantly above chance level, but as *t*_Unaware if the *t* test is nonsignificant. For each iteration, a one-sample t test compared accuracy on the X-half of trials to 0.5 (chance level) and the iteration was deemed *t*_Aware or *t*_Unaware if the one-sample t test on the X-half of trials was significantly above 0.5 or not. Mean awareness scores for the X-half and for the remaining half of the trials (Y-half) were also computed and stored. After the resampling procedure, mean cumulative X-half and Y-half awareness scores for *t*_Aware and *t*_Unaware iterations were computed, and the means were weighted according to the number of iterations for each categorization outcome. For *t*_Aware iterations, the mean Type I awareness on Y trials was 0.851, *SD* = 0.12, 95% CI [0.862, 0.864]. For *t*_Unaware iterations, the mean Type I awareness on Y trials was highly probably above 0.5, if only by a small amount, 0.510, *SD* = 0.085, 95% CI [0.510, 0.512].

Figure 4 presents density plots for X and Y halves of each *t*_Awareness group, showing how *t*_Unaware iterations both on the X and Y halves were above chance level, with awareness scores on *t*_Unaware Y-halves regressing even more towards the mean.

Discussion

This report examined the reliability of a measure of awareness used in an implicit conditioning task. We were able to demonstrate, as previously predicted (Shanks, 2017), that using standard significance tests to categorize participants as *t*_Aware or *t*_Unaware of contingencies yields sufficient regression to the mean on a subsequent test to undermine the

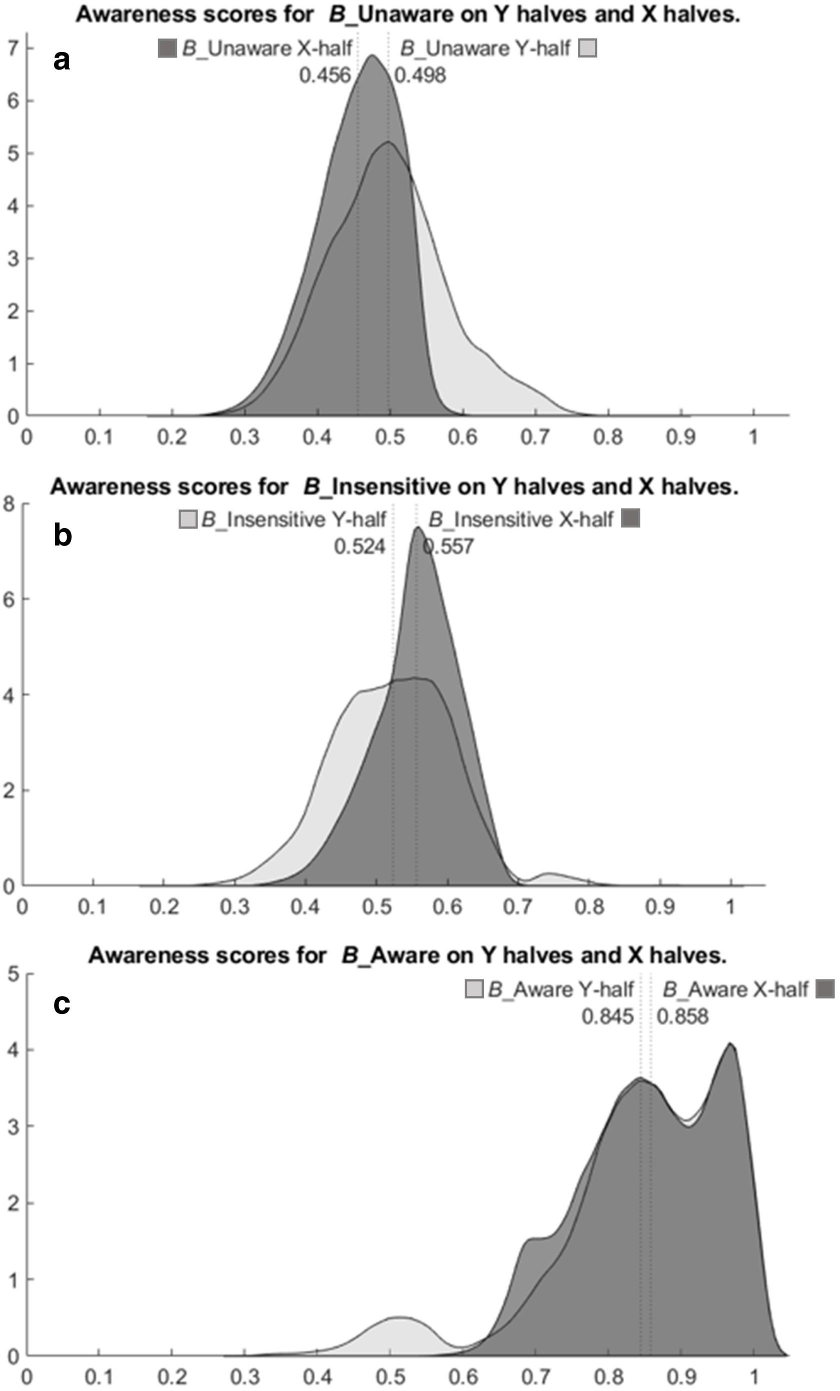

Fig. 2 Density plots on the resampling procedure on awareness scores for X and Y halves in *B*_Unaware (**a**), *B*_Insensitive (**b**), and *B*_Aware (**c**) iterations. Iterations resulting in a sensitive Bayes factor for lack of awareness on the X-halves (**a**) produced cumulative awareness scores on the Y-halves that remained below chance level. Iterations for which the Bayes factor was insensitive (**b**) produced cumulative awareness scores on the Y-halves that were above chance level

original classification. That is, determining the absence of knowledge based on the absence of significance of an awareness measure (score not different than chance level), eventually leads to participants being incorrectly categorized as unaware. This is a crucial consideration as learning effects shown by participants wrongly categorized as unaware will falsely increase evidence for the existence of implicit responses, and this has been the customary approach in the study of implicit processes.

However, using Bayes factors as a categorization tool, regression-to-the-mean effects in *B*_Unaware participants were too small to be able to undermine the classification.

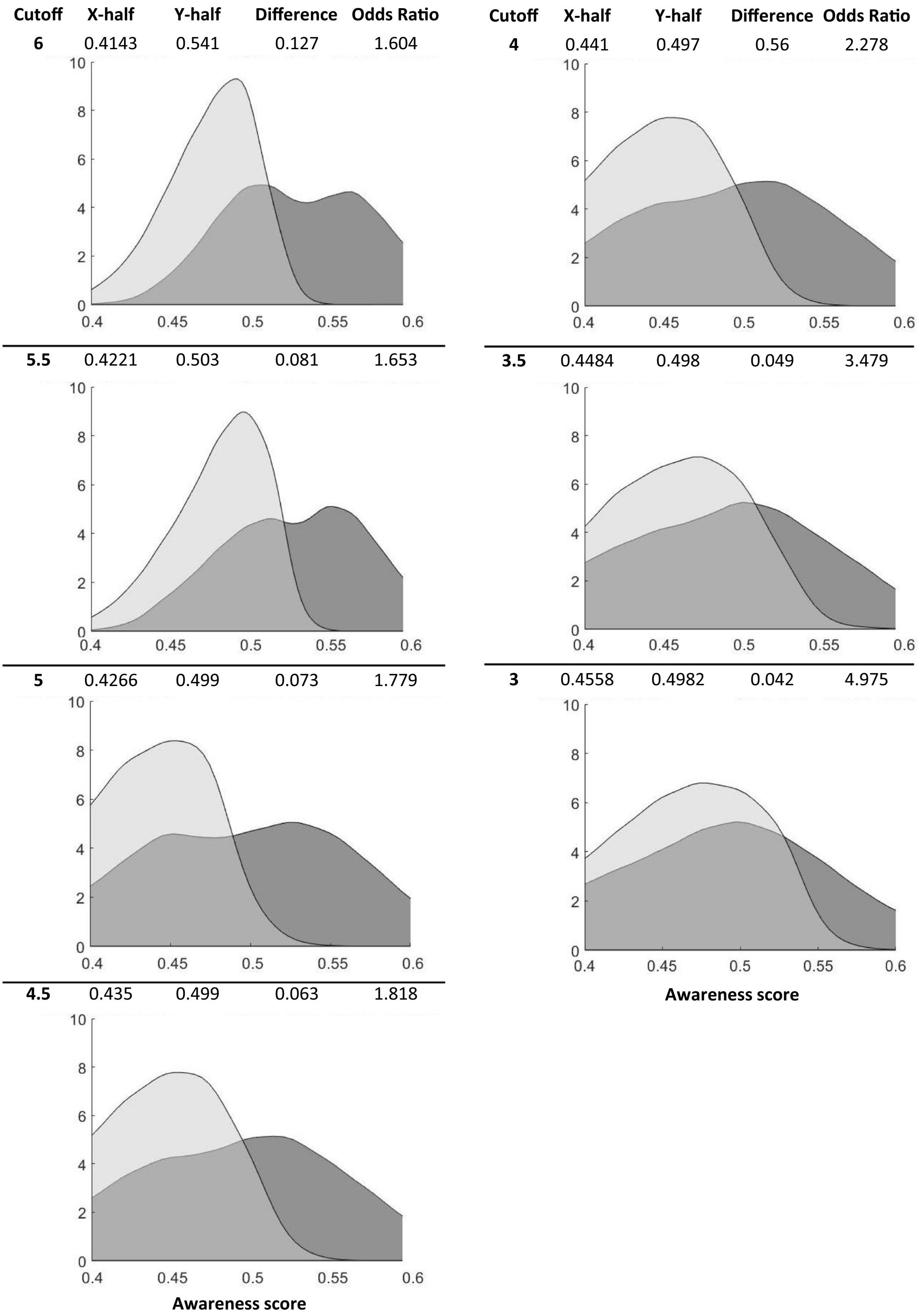

Fig. 3 Density plots on the group boundary optimization procedure on awareness scores for X and Y halves in *B*_Unaware iterations. The figure presents the mean awareness score on X and Y halves for different cutoffs on the Bayes factor. This shows that with more stringent cutoffs, the scores on X-halves are more extreme and separated from Y-halves, as noted by the difference scores. For cutoffs = 6 and 5.5, regression to the mean occurs, as Y halves are above chance level. More liberal cutoffs produce less regression to the mean, as shown by the lesser difference between X and Y halves, and an increase in the *B*_Unaware/*B*_Insensitive odds ratio

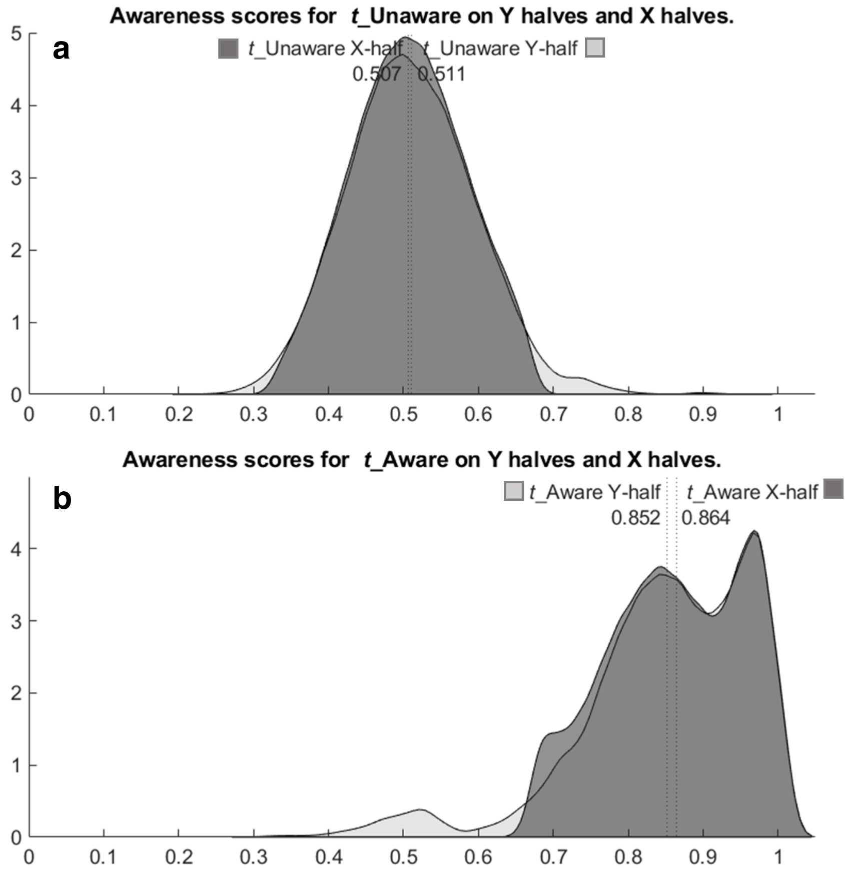

Fig. 4 Density plots on the resampling procedure on awareness scores for X and Y halves in *t*_Unaware (**a**), *t*_Aware (**b**) iterations. Iterations resulting in a nonsignificant awareness categorization on the X-halves (**a**) produced cumulative awareness scores on the Y-halves significantly above chance level, showing regression-to-the-mean effects.

Their scores on a concurrent awareness measure regressed towards chance level, strengthening the argument that their true state is performance at chance level. This was tested by firstly categorizing participants as *B*_Aware, *B*_Unaware, or *B*_Insensitive. We then performed *t* tests on the aggregated awareness scores, for each group, finding that whereas *B*_Aware and *B*_Insensitive groups scored above chance level, *B*_Unaware participants had scores below chance level. To address regression to the mean, using a split-half resampling procedure with multiple iterations we found awareness scores as they should be according to the original classification—that is, high for those classified as *B*_Aware and not above chance for those classified as *B*_Unaware. Crucially, using this Bayesian awareness categorization we detected preferential attentional responses towards reward CS in *B*_Unaware participants (as reported in Leganes-Fonteneau et al., 2019; Leganes-Fonteneau et al., 2018). The present results increase the validity of those findings and the evidence for the existence of implicit reward learning.

The fact that *B*_Unaware participants showed awareness scores (just) below chance level both on aggregate means and on the resampling procedure is surprising. Similar results have been observed in the past in subliminal perception (Kemp-Wheeler & Hill, 1988) and implicit learning (Ziori & Dienes, 2012) paradigms, with d' scores below 0. Negative d' could arise because some participants had incorrect conscious knowledge. In this case, any incorrect conscious knowledge is unlikely to explain the flanker effect, where highly rewarded stimuli were more distracting than unrewarded stimuli for *B*_Unaware participants, whereas results for *B*_Aware were insensitive.

It is important to consider that *B*_Insensitive participants did show above-chance performance. Participants whose Bayes factors did not reveal sensitive evidence for or against the null hypothesis showed above-chance level scores on their aggregated means and on the resampling procedure. These participants would be deemed, according to most traditional categorization procedures, as *t*_Unaware, as their scores would fail to show significant awareness above chance level. Hence, the *t*_Unaware group comprises *B*_Unaware and *B*_Insensitive participants, the latter being the ones driving regression-to-the-mean effects observed in this group. We propose, therefore, that *B*_Insensitive participants are excluded as a group when examining learning effects. Regression to the mean need not in general be driven by *B*_Insensitive participants, but it can also in principle arise from the more extreme group selection incurred with *B*_Unaware participants using a high threshold, as we also showed.

The group boundary optimization procedure is a useful tool because it allows to flexibly apply different cutoffs to the BACT, providing an indication of what is the highest cutoff that can be used in the categorization without incurring regression-to-the-mean effects. After the analysis, one can decide whether to select a more common and arbitrary cutoff (i.e., 3-1/3) for the categorization procedure, or whether cutoff selection can be informed as a function of the data in hand. However, when the threshold is more stringent, more participants would be deemed as *B*_Insensitive. It is therefore necessary to balance out how strict one wants the cutoff to be, and what percentage of the population will be lost in that categorization procedure.

This procedure shows that the occurrence of regression to the mean depends on the criterion applied. Perhaps counterintuitively, we observed that a too extreme criterion on the selection of *B*_Unaware iterations yields a strongly biased subsample of X-halves, which in turn generates discrepancies on the retest procedure. That is because, despite strengthening the evidence for there being an effect or no effect on the X-half categorization; it also results in a more extreme subgroup of participants being selected. The more extreme the group selection criterion, the more biased the measurement error in the group, and the further the observed results (X-halves) will be from the retest scores (Y-halves), producing regression-to-the-mean effects.[1] It seems that how this balances out in terms of overall regression to the mean is hard to predict and depends on, for example, the level of noise in the measurements originally taken. However, the upshot is that regression to the mean cannot be taken as guaranteeing there will always be a problem with subgroup selection. Conversely, one cannot presume it is not a problem without due argument and examination of the behavior of the BACT under different thresholds.

The analysis reported here validates the use of Bayes factors as an awareness categorization tool. The Aware/Unaware odds-ratio analysis showed a strong association between being classified as aware versus unaware on one half, and the same classification on the second half, whereas the Unaware/Insensitive odds-ratio was higher for more lenient significance cutoffs. We believe that this new methodology has strong implications for the study of unconscious and implicit processes. These results have been obtained in the context of implicit conditioning, and this tool allows to categorize participants according to their levels of stimulus–outcome contingency awareness. However, we consider that this technique should be used in other implicit learning paradigms as well as in subliminal perception tasks. We hope this analysis will prime researchers to utilize this methodology in any experiment in which post hoc awareness categorizations are used. Furthermore, including a short reliability analysis in experimental reports would provide evidence for the sensitivity of an awareness categorization tool.

This categorization procedure is not only novel due to the use of Bayes factors, but also for the inclusion of Type II measures of metacognitive knowledge to extrapolate a suitable prior for Type I analyses within the data set. Therefore, to the existing recommendations for awareness measurements in implicit paradigms (i.e., Lovibond & Shanks, 2002; Mertens et al., in press) we can add that measures of awareness should be based on a Type I dichotomous measure (i.e., “Will you get a reward? Yes-No”), followed by a Likert scale measure of confidence. This Likert scale (Dienes, 2007) can be transformed to a dichotomous variable to obtain Type II d' scores. In case measures of confidence are not available, it is possible to circumvent the use of Type II d' scores by specifying a suitable model of H1 for the Bayes factor, obtaining a $\mathrm{Log}_{d1'}$ from a different sample of aware participants based on previous research.

Our use of Bayes factors (with common or individualized models of H1) contrasts with approaches which model the variability across subjects in order to gain more information about each subject (e.g., Bernardo et al., 2011; Haaf & Rouder, 2019; Williams, Martin, & Rast, 2019). The variance of the learning measure (i.e., modified flanker task in this case), could also be investigated by these methods, and specifically the extent to which this variability reflects different individuals actually performing below, at or above chance (Haaf & Rouder, 2019). An advantage of our method is that the models of H1 (priors) are informed for each subject. There are good reasons for each individual why the parameter values for the models of H1 should have certain approximate values. Thus, on the hypothesis that the person has conscious knowledge of a certain sort, there are constraints on the range of values that conscious knowledge could be. Thus, parameters (scale factors for priors) are not arbitrary, but are dictated by the scientific problem. Future research could usefully compare and contrast these different approaches.

Other procedures, such as equivalence testing (Lakens, McLatchie, Isager, Scheel, & Dienes, 2020), could be explored as an alternative to the use of Bayes factors. However, equivalence testing relies on the determination of minimal interesting effect sizes (Dienes, 2020), which are problematic in the context of determination of unawareness. Moreover, Bayes factors have a crucial advantage over equivalence testing, as they allow quantifying evidence for the null hypothesis.

Regression to the mean is a special case of problems that can arise from selective inference (e.g., Leeb & Potscher, 2006; Meir & Drton, 2017). Other problems may arise—for

[1] Using two different data sets obtained with a similar experimental paradigm (manuscript in preparation, data not presented, n=86 and n=40) the threshold at which the BACT did not generate regression to the mean was different in each case (e.g., 3-1/3 and 6-1/6 respectively).

example, selecting on one criterion may produce a truncated postselection distribution for further analyses, which may be impossible to estimate, or in general may violate assumptions of further tests. In our simulations, we selected people scoring above or below a criterion on half the trials, and measured again on the other half (for different 50/50 splits). The distribution of proportion correct scores before selection is shown in the Appendix. It is positively skewed and probably bimodal as in fact, skewness = 0.58, *SE* = 0.34. We select according to a Bayes factor criterion, which means the selection is not a strict hard cutoff, but may approximate a hard cutoff, producing approximate/noisy truncation. The result for unaware participants, the group we are most interested in, can be seen in Fig. 2a. In fact, cutting off the positive tail results in a distribution with a skew no higher in magnitude, just negative −0.54, *SE* = .51 (for the X-halves, post selection). So, in terms of skew, assumptions are no more violated as a result of selection; and in fact, the distribution no longer appears bimodal, so selection has rendered conclusions using *t* tests or Bayes factors assuming the normal approximation more appropriate. A similar conclusion holds for log *d*′ as the dependent variable; in unselected data, there is a large positive skew and possible polymodality; and a skew of equal magnitude, but negative with a unimodal distribution for the postselected aware group in the X-halves. However, there is no guarantee that other data sets will behave in this way. Postselection distributions should be inspected to determine the suitability of the data for further analyses. The methodological advances provided by the use of Bayes factors should also be applied to experiments previously published. Considering the number of criticisms and inconsistent findings in the implicit literature, it would be highly beneficial for the field to reexamine some of the most influential papers using this methodology. In order to facilitate the implementation of this methodology, we have made the MATLAB scripts necessary to run this analysis available online.[2]

In summary, Bayes factors, in the context of awareness categorization, present a twofold advantage over customary measures of learning. First, they allow determining the sensitivity of awareness scores individually in a reliable way, basing the categorization on statistically informative methods instead of on a lack of significant results. Second, Bayes factors allow excluding insensitive data points and in doing so drive a reduction in measurement error. This way it is possible to avoid regression-to-the-mean effects observed using traditional statistics that can misrepresent the occurrence of implicit processes.

We present here evidence for the existence of regression to the mean effects in traditional awareness categorization procedures, but propose and validate a novel methodology using Bayes factors that can prevent the occurrence of regression to the mean effects. The ability of Bayes factors to tease apart sensitive evidence for the null hypothesis (unawareness) from insensitive evidence seems to drive the reliability of this awareness categorization tool. This methodology has the potential to improve the quality and accuracy of research on implicit and unconscious processes as well as other fields where post hoc group selection is necessary.

Appendix

The original Bayesian awareness categorization used in Leganes-Fonteneau et al. (2018) and (2019) was based on the use of Type I and Type II *d*′ scores to obtain sensitive evidence for the lack of awareness. Using a novel conditioning paradigm, we did not have a prior study to inform expectations of effect size for Type I awareness, we therefore designed a methodology in which constraints defined by aspects of each participant's performance could be used as constraints for other aspects (see Dienes, 2015, 2019, for this general approach). In other cases, where prior studies exist or a norming study is done, Type I (or raw accuracy) scores from a prior study can be used to constrain Type II performance in order to model H1. The procedure used to compute Bayes factors for Type I and Type II *d*′ scores is detailed here.

Type I refers to the ability to discriminate states of the world (e.g., octagons are followed by reward); Type II refers to the ability to discriminate the accuracy of one's knowledge (being more confident in accurate responses). A Bayes factor requires a model of the range of effects predicted (the model of H1). The model of H1 can be a uniform [0, max], indicating that the effect can be anything from 0 to a maximum. A Type II *d*′ typically is not higher than the Type I *d*′; thus, the analysis utilizes participants' own Type I knowledge in the model of H1 for Type II *d*′; specifically, the Bayes factor for Type II *d*′ used the uniform [0, Type I *d*′ for that participant]. Then, the mean Type I score (M1) of participants with Type II knowledge can be taken to indicate the amount of Type I knowledge a participant has if they have metacognitively conscious knowledge. Presuming that unconscious Type I knowledge would not be more than Type I knowledge that is metacognitively conscious, a model of H1 can be constructed for Type I knowledge—namely the uniform [0, M1]. This model of H1 was used to determine the existence of Type I knowledge.

In detail, using expectancy data, the number of individual Hits (H, answering Yes on a high-reward trial), Correct Rejections (CR, answering No on a low-reward trial), False Alarms (FA, answering Yes on a low-reward trial), and Misses (M, answering No on a high-reward trial) were computed to obtain an odds ratio (*OR*). In this way $\log_{d1'}$ scores (logistic *d*′) were obtained for each participant (Eq. 1) and their

[2] https://osf.io/mqgw4/?view_only=94f9f196a95a4e71b168f65988b66ed2

corresponding $SE_{d1'}$ (Eq. 2).

$$log_{d1'} = In(\mathrm{OR})*\frac{\sqrt{3}}{\pi}, \quad (1)$$

$$SE_{d1'} = \sqrt{\frac{1}{H}+\frac{1}{CR}+\frac{1}{FA}+\frac{1}{M}}*\frac{\sqrt{3}}{\pi}. \quad (2)$$

For Type II scores, providing information about the metacognitive knowledge of participants about CS-outcome contingencies, an equivalent analysis was performed using accuracy and confidence scores. The Likert scale was transformed to a dichotomous variable. A score of 2 (more or less guessing) or less was considered as low confidence and a score equal or higher than 3 (fairly sure) as high confidence. Hits (correct expectancy with high confidence), Correct Rejections (incorrect expectancy with low confidence), False Alarms (incorrect expectancy with high confidence), and Misses (correct expectancy with low confidence), were used to obtain $\log_{d2'}$ scores and corresponding $SE_{d2'}$.

For each participant, a Bayes factor was computed on their $\log_{d2'}$ and $SE_{d2'}$, modeling H1 with a Uniform going from 0 (chance level) to their own $\mathrm{Log}_{d1'}$ as a model of H1. Participants with $B_{d2'} < 1/3$ were categorized as Metacognitively Unaware, whereas those with a $B_{d2'} > 3$ were considered Metacognitively Aware, and the rest ($1/3 < B_{d2'} < 3$) as Insensitive. The mean $\mathrm{Log}_{d1'}$ of Metacognitively Aware participants (M1) was then used as a maximum for a uniform Bayes factor to model H1 testing the sensitivity of each participant's $\mathrm{Log}_{d1'}$. Participants with $B_{d1'} < 1/3$ were categorized as *B*_Unaware of the contingencies, whereas those with a $B_{d1'} > 3$ were considered *B*_Aware, and the rest as *B*_Insensitive Figs. 5 and 6.

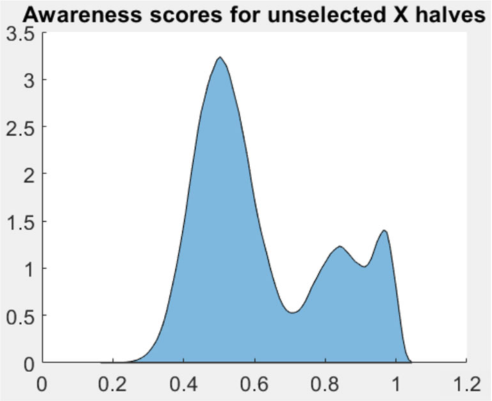

Fig. 5 Distribution of awareness scores for all X-halves before the categorization is performed

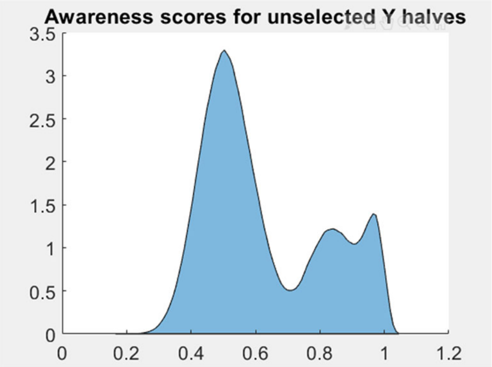

Fig. 6 Distribution of awareness scores for all Y-halves before the categorization is performed. Figures 5 and 6 look identical but they are not, note thedifference in density at 0.95

Acknowledgments The authors wish to acknowledge Prof. David Shanks from UCL for his suggestions on the data analysis and his comments on the manuscript. We are also extremely grateful to very thoughtful reviews from Donny Williams and Julia Haaf.

References

Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., & West, M. (2011). Bayesian variable selection for random intercept modeling of Gaussian and non-Gaussian data. *Bayesian Statistics*, *9*, 165–185.

Chun, M. M., & Jiang, Y. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology*, *36*(1), 28–71. https://doi.org/10.1006/cogp.1998.0681

Davenport, S., & Nichols, T. E. (2020). Selective peak inference: Unbiased estimation of raw and standardized effect size at local maxima. *NeuroImage, 209,* 116375. https://doi.org/10.1016/j.neuroimage.2019.116375

Dienes, Z. (2007). Subjective measures of unconscious knowledge. *Progress in Brain Research*. Elsevier. https://doi.org/10.1016/S0079-6123(07)68005-4

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*, 781. https://doi.org/10.3389/fpsyg.2014.00781

Dienes, Z., & Overgaard, M. (2015). How Bayesian statistics are needed to determine whether mental states are unconscious. Behavioural methods in consciousness research, 2015, 199–220

Dienes, Z. (2019). How do i know what my theory predicts? *Advances in Methods and Practices in Psychological Science*, *2*(4), 364–377. https://doi.org/10.1177/2515245919876960

Dienes, Z. (2020). *Obtaining evidence for no effect*. PsyArXiv Preprint. https://doi.org/10.31234/osf.io/yc7s5

Jeffreys, H. (1939). The Theory of Probability, 1st Edn. Oxford, England: Oxford University Press.

Haaf, J. M., & Rouder, J. N. (2019). Some do and some don't? Accounting for variability of individual difference structures. *Psychonomic Bulletin & Review*, *26*(3), 772–789. https://doi.org/10.3758/s13423-018-1522-x

Hofmann, W., De Houwer, J., Perugini, M., Baeyens, F., & Crombez, G. (2010). Evaluative conditioning in humans: A meta-analysis. *Psychological Bulletin*, *136*(3), 390–421. https://doi.org/10.1037/a0018916

Kemp-Wheeler, S. M., & Hill, A. B. (1988). Semantic priming without awareness: Some methodological considerations and replications. *The Quarterly Journal of Experimental Psychology Section A*, *40*(4), 671–692. https://doi.org/10.1080/14640748808402293

Lakens, D., McLatchie, N., Isager, P. M., Scheel, A. M., & Dienes, Z. (2020). Improving inferences about null effects with Bayes factors and equivalence tests. *The Journals of Gerontology: Series B*, *75*(1), 45–57. https://doi.org/10.1093/geronb/gby065

Leeb, H., & Potscher, B. M. (2006). Can one estimate the conditional distribution of postmodel-selection estimators? *The Annals of Statistics, 34*(5), 2554–2591.

Leganes-Fonteneau, M., Nikolaou, K., Scott, R., & Duka, T. (2019). Knowledge about the predictive value of reward conditioned stimuli modulates their interference with cognitive processes. *Learning & Memory (Cold Spring Harbor, N.Y.)*, *26*(3), 66–76. https://doi.org/10.1101/lm.048272.118

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. https://doi.org/10.1016/j.bbr.2017.12.015

León, O. G., & Suero, M. (2000). Regression toward the mean associated with extreme groups and the evaluation of improvement. *Psicothema, 12*(1), 145–149

Lovibond, P. F., & Shanks, D. R. (2002). The role of awareness in Pavlovian conditioning: Empirical evidence and theoretical implications. *Journal of Experimental Psychology: Animal Behavior Processes*, *28*(1), 3–26. https://doi.org/10.1037/0097-7403.28.1.3

Meir, A., & Drton, M. (2017). *Tractable post-selection maximum likelihood inference for thelasso.* ArXiv Preprint. arXiv:1705.09417

Mertens, G., & Engelhard, I. M. (2020, January 1). A systematic review and meta-analysis of the evidence for unaware fear conditioning. *Neuroscience and Biobehavioral Reviews*. Elsevier. https://doi.org/10.1016/j.neubiorev.2019.11.012

Mertens, G, Basci, A., & Engelhard, I.M. (in press). *A critical review and meta-analysis of the evidence for unaware fear conditioning*. https://doi.org/10.31234/osf.io/qz5st

Nikolaou, K., Field, M., & Duka, T. (2013). Alcohol-related cues reduce cognitive control in social drinkers. *Behavioural Pharmacology*, *24*(1), 29–36. https://doi.org/10.1097/FBP.0b013e32835cf458

Reber, A. S., & (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General*, *118*(3), 219–235. https://doi.org/10.1037/0096-3445.118.3.219

Rebuschat, P. (2013). Measuring implicit and explicit knowledge in second language research. *Language Learning*, *63*(3), 595–626. https://doi.org/10.1111/lang.12010

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. https://doi.org/10.1016/j.concog.2016.06.012

Scott, R. B., & Dienes, Z. (2010). Knowledge applied to new domains: The unconscious succeeds where the conscious fails. *Consciousness and Cognition*, *19*(1), 391–398. https://doi.org/10.1016/j.concog.2009.11.009

Shanks, D. R. (2017). Regressive research: The pitfalls of post hoc data selection in the study of unconscious mental processes. https://doi.org/10.3758/s13423-016-1170-y

Vadillo, M. A., Konstantinidis, E., & Shanks, D. R. (2016). Underpowered samples, false negatives, and unconscious learning. *Psychonomic Bulletin & Review*, *23*(1), 87–102. https://doi.org/10.3758/s13423-015-0892-6

Williams, D. R., Martin, S. R., & Rast, P. (2019). *Putting the individual into reliability: Bayesian Testing of homogeneous within-person variance in hierarchical models.* https://doi.org/10.31234/osf.io/

Yokoyama, T., Padmala, S., & Pessoa, L. (2015). Reward learning and negative emotion during rapid attentional competition. *Frontiers in Psychology*, *6*, 269. https://doi.org/10.3389/fpsyg.2015.00269

Ziori, E., & Dienes, Z. (2012). The time course of implicit and explicit concept learning. *Consciousness and Cognition*, *21*(1), 204–216. https://doi.org/10.1016/j.concog.2011.12.008