Correlation Analysis to Investigate Unconscious Mental Processes: A Critical Appraisal and Mini-Tutorial

Simone Malejka[1], Miguel A. Vadillo[2], Zoltán Dienes[3], and David R. Shanks[1]

[1]University College London, United Kingdom

[2]Universidad Autónoma de Madrid, Spain

[3]University of Sussex, United Kingdom

Author Note:

Simone Malejka https://orcid.org/0000-0002-7012-627X

Miguel A. Vadillo https://orcid.org/0000-0001-8421-816X

Zoltán Dienes https://orcid.org/0000-0001-7454-3161

David R. Shanks https://orcid.org/0000-0002-4600-6323

Correspondence concerning this article should be addressed to Simone Malejka, who is now at the Department of Psychology, University of Cologne, Albertus-Magnus-Platz, 50923 Cologne, Germany. Email: simone.malejka@uni-koeln.de

Abstract

As a method to investigate the scope of unconscious mental processes, researchers frequently obtain concurrent measures of task performance and stimulus awareness across participants. Even though both measures might be significantly greater than zero, the correlation between them might not, encouraging the inference that an unconscious process drives task performance. We highlight the pitfalls of this null-correlation approach and provide a mini-tutorial on ways to avoid them. As reference, we use a recent study by Salvador, Berkovitch, Vinckier, Cohen, Naccache, Dehaene, and Gaillard (2018) reporting a non-significant correlation between the extent to which memory was suppressed by a Think/No-Think cue and an index of cue awareness. In the *Null Hypothesis Significance Testing* (NHST) framework, it is inappropriate to interpret failure to reject the null hypothesis (i.e., correlation = 0) as evidence for the null. Furthermore, psychological measures are often unreliable, which can dramatically attenuate the size of observed correlations. A Bayesian approach can circumvent both problems and compare the extent to which the data provide evidence for the null versus the alternative hypothesis (i.e., correlation > 0), while considering the usually low reliabilities of the variables. Applied to Salvador et al.'s data, this approach indicates no to moderate support for the claimed unconscious nature of participants' memory-suppression performance—depending on the model of the alternative hypothesis. Hence, more reliable data are needed. When analyzing correlational data, we recommend researchers to employ the Bayesian methods developed here (and made freely available as R scripts), rather than standard NHST methods, to take account of unreliability.

Keywords: unconscious cognition, memory suppression, correlation attenuation, reliability, measurement error, Bayes factor

Efforts to understand the scope and importance of unconscious mental processes form a prominent part of current research in psychology and cognitive neuroscience, and show no sign of abating despite decades of controversy (e.g., Eriksen, 1960; Hassin, 2013; Hedger, Gray, Garner, & Adams, 2016; Holender, 1986; LeDoux, Michel, & Lau, 2020; Shanks & St. John, 1994). At their heart, many of the disagreements stem from alternative viewpoints about the inferences that can validly be drawn from particular experimental methods and data-analysis techniques. Here we describe some reasons to be extremely cautious about what at first glance seems to be an intuitive and valid type of evidence (namely, interpreting non-significant correlations), but which on deeper reflection should be treated with considerable skepticism.

In the following, we outline the standard problem in research on unconscious mental processes, and summarize the traditional approaches to solve it. We then highlight the pitfalls of the prominent null-correlation approach with reference to an empirical study published in this journal. As the main part of this article, we offer alternative and arguably better methods to analyze the data, and present a mini-tutorial on how to use them. Our results for the example data support our main claim: Researchers should routinely check and report the reliability of their measures as unreliable data do not allow strong conclusions about the existence, or non-existence, of unconscious mental processes. Fortunately, the methods presented here can flag these non-diagnostic cases by showing researchers when better data should be collected.

Problem Outline

Many studies of unconscious mental processes collect bivariate data across a sample of participants, with the data comprising a performance measurement (e.g., accuracy, reaction times) and an awareness measurement (e.g., recognition, discriminability, visibility) from

each person. What might a researcher infer from such a dataset? One possibility is that at the aggregate group level, awareness is not greater than some baseline or chance level. For example, in a subliminal priming task in which performance on an indirect test is influenced by a brief masked visual stimulus, participants might lack any ability to detect or discriminate the stimulus in a direct test (*indirect-without-direct effect* data pattern; Klauer, Draine, & Greenwald, 1998). Often this would take the form of an estimate close to zero for the discrimination index d' from signal-detection theory (SDT; Green & Swets, 1966), such that the above-chance performance in the indirect test can be interpreted as arising from unconscious processing.[1] However, data conforming to such a pattern are rare (e.g., Dehaene et al., 1998; Finkbeiner, 2011), and many studies in the field are underpowered due to the small number of trials typically included in the awareness test (see Vadillo, Konstantinidis, & Shanks, 2016; Vadillo, Linssen, Orgaz, Parsons, & Shanks, 2020).

More frequently, patterns are reported in which awareness is numerically—and often significantly—greater than zero. What can the researcher do in such cases to determine whether unconscious processes have played a part in task performance? Two commonly used approaches exist. One approach is to remove from the analysis *post hoc* all participants who score at or above zero on the awareness measure, and to ask whether the remaining participants (for all of whom awareness is at or below zero) score above chance on the task-performance measure. However, this approach has been strongly criticized on statistical grounds: As Shanks (2017) showed, the phenomenon of regression to the mean virtually guarantees that some participants classified as "unaware" will achieve above-chance task performance. The reason for this is that if there is random measurement error in the awareness index, some of the participants retained in the analysis will truly

[1] To prove the absence of an effect, researchers have traditionally relied on non-significant null hypothesis tests. Alternative methods include confidence intervals around the effect size that lie between equivalence bounds (Lakens, 2017) and Bayes factors in favor of the null hypothesis (Lakens, McLatchie, Isager, Scheel, & Dienes, 2018; Rouder et al., 2009).

have zero awareness, while others will actually have above-zero awareness. The latter participants obtained a score at or below zero only due to chance, and they would show an above-chance score on a second, independent measurement. For these participants, measurement error will make it appear, wrongly, that they lack awareness. If awareness correlates with task performance at the unobserved level (perhaps because the two are based on the same latent construct), then the "unaware" subgroup will perform above chance on the task-performance measure.[2]

A second approach to determine whether unconscious processes have influenced task performance is to correlate task performance and awareness. If such an analysis shows that the correlation coefficient is close to zero, then it seems legitimate to infer that awareness does not explain any of the variance in performance, which must therefore rest on some other process—presumably an unconscious one (see Greenwald, Klinger, & Schuh, 1995, for a detailed analysis of the assumptions underlying this inference). Many studies across a range of domains have employed this line of reasoning (for recent examples, see Table 1). In the present article, we explore some of the pitfalls inherent in the approach. We illustrate these by reference to a recent study by Salvador, Berkovitch, Vinckier, Cohen, Naccache, Dehaene, and Gaillard (2018), but it is important to emphasize that our arguments are general and apply equally to other instances of the approach's use. To preview, we show that Salvador et al.'s empirical evidence neither supports the inference that the observed correlation is zero nor that it is larger than zero. The small number of trials causes the indirect measure to be too unreliable to draw any conclusion—apart from the conclusion that new data must be collected.

---

[2] For a Bayesian solution to this problem, see Rouder, Morey, Speckman, and Pratte (2007).

Table 1

Selection of Recent Published Articles Using Different Tasks that Employ Correlation or Regression Analysis to Test the Null Hypothesis of No Relationship Between Indirect and Direct Performance Measures

Study	Paradigm	Indirect Measure	Direct Measure	Statistic
Batterink, Reber, Neville, and Paller (2015)	Statistical learning	Difference in reaction times	Recognition accuracy	Correlation
Berkovitch and Dehaene (2019)	Syntactic priming	Difference in reaction times	Discriminability (d')	Correlation
Chiu and Aron (2014)	Response inhibition	Difference in reaction times	Discriminability (d')	Correlation
Colagiuri and Livesey (2016)	Contextual cueing	Difference in reaction times	Recognition accuracy	Correlation
Dickinson and Brown (2007)	Evaluative conditioning	Difference in valence ratings	Contingency measure	Correlation
Geyer, Shi, and Müller (2010)	Contextual cueing	Difference in reaction times	Recognition accuracy	Correlation
Hedger, Garner, and Adams (2019)	Visual probe paradigm	Difference in reaction times	Discriminability (d')	Correlation
Jensen, Kirsch, Odmalm, Kaptchuk, and Ingvar (2015)	Classical conditioning	Difference in pain ratings	Recognition accuracy	Correlation
Kalra, Gabrieli, and Finn (2019)	Artificial grammar learning	Discriminability (d')	Various	Correlation
	Probabilistic classification	Accuracy measure	Various	Correlation
	Serial response	Reaction time score	Various	Correlation
	Category learning	Accuracy measure	Various	Correlation
Paciorek and Williams (2015)	Language learning	Difference in recognition accuracies	Difference in confidence ratings	Correlation
Salvador et al. (2018)	Memory suppression	Difference in memory accuracies	Discriminability (d')	Slope
Sanchez, Gobel, and Reber (2010)	Sequence learning	Difference in performances	Difference in confidence ratings	Correlation
Sklar, Levy, Goldstein, Mandel, Maril, and Hassin (2012)	Mathematical-equation priming	Difference in reaction times	Discriminability (d')	Slope
	Multiple-word priming	Difference in reaction times	Discriminability (d')	Slope

Example Dataset

In two experiments, Salvador et al. (2018) first taught participants a list of word pairs such as *candle – champagne* and *wood – knife*. Next, they trained them to link one geometrical shape (e.g., a diamond) with recalling a hint word's (e.g., *candle*) associate (e.g., *champagne*) and another shape (e.g., a square) with suppressing recall of a hint word's (e.g., *wood*) associate (e.g., *knife*). In this version of the Think/No-Think procedure, the aim was to establish the geometrical shapes as cues of either recalling the associate or suppressing its recall (Anderson & Green, 2001). Critically, on some trials, the shape cue was presented "subliminally":[3] It was flashed for just 16 ms and followed by a mask. In the final test phase, participants were given the hint words and asked to recall all associates without any shape cues present. Salvador et al. reported that participants' ability to recall the correct associate was impaired for words previously paired with the No-Think shape compared to ones paired with the Think shape, revealing a memory suppression effect. Crucially, this effect extended to words preceded by both supraliminal and "subliminal" shape presentations. The key performance measure was the difference in the probability of correct recall of the associate given the hint word formerly paired with a Think shape versus a No-Think shape presented for only 16 ms (termed δ in the following), which was significantly larger than zero in one-tailed one-sample t-tests, mean = .06, $t(43) = 2.03$, $p = .024$, $d = 0.31$, in Experiment 1, and mean = .17, $t(29) = 3.55$, $p < .001$, $d = 0.65$, in Experiment 2.[4]

---

[3] We place "subliminal" in quotes because the issue here is whether or not unconscious perception of the geometric shapes was established.

[4] Note that Salvador et al. (2018) subtracted the percentage correctly recalled in the Think condition from that in the No-Think condition. We instead calculated probabilities and subtracted the probability of correct recall in the No-Think condition from the probability of correct recall in the Think condition, with the latter leading to reversed signs for the slopes and correlations. The conventional significance level of $\alpha = .05$ was used for all frequentist statistical tests.

To establish to what extent participants were consciously aware of the briefly flashed shapes, Salvador et al. (2018) conducted a visibility test at the end of each experiment in which these shapes were again presented for 16 ms and immediately masked, but now participants had to indicate whether each shape was a square or a diamond. The key awareness measure was SDT's discrimination index d' calculated from the hit and false alarm rates. Participants' ability to discriminate between the geometrical shapes was reported as low but significantly above zero in one-tailed one-sample t-tests, mean = 0.35, $t(43) = 5.01$, $p < .001$, $d = 0.75$, in Experiment 1, and mean = 0.21, $t(29) = 2.23$, $p = .017$, $d = 0.41$, in Experiment 2.

For present purposes, the exact nature of the tasks employed by Salvador et al. (2018), and indeed their domain of investigation (memory suppression), are not critical. What matters is that they collected a performance measure (which happened to be a memory suppression score) and an awareness measure (forced-choice discrimination) from each of a sample of participants, and determined that the slope when regressing the performance measure on the awareness measure did not differ significantly from zero, slope = −0.05, $t(42) = -0.77$, $p = .445$, in Experiment 1, and slope = 0.01, $t(28) = 0.07$, $p = .945$, in Experiment 2. These results were interpreted by them as evidence for memory suppression in the absence of awareness. The scatterplot and regression lines are depicted in the top-left panel of Figure 1. Although not reported by Salvador et al., the Pearson correlations between performance and awareness were $r = -.12$ and $r = .01$ for Experiments 1 and 2, respectively. For expositional purposes, we switch from the slope to the correlation. The correlation is identical to the standardized slope and removes the unit of analysis.

How valid is the inference of null correlations? We will argue that the evidence reported by Salvador et al. (2018) provides no more than anecdotal support for their claim about unconscious processes when considered in terms of correlations. Our key target is the

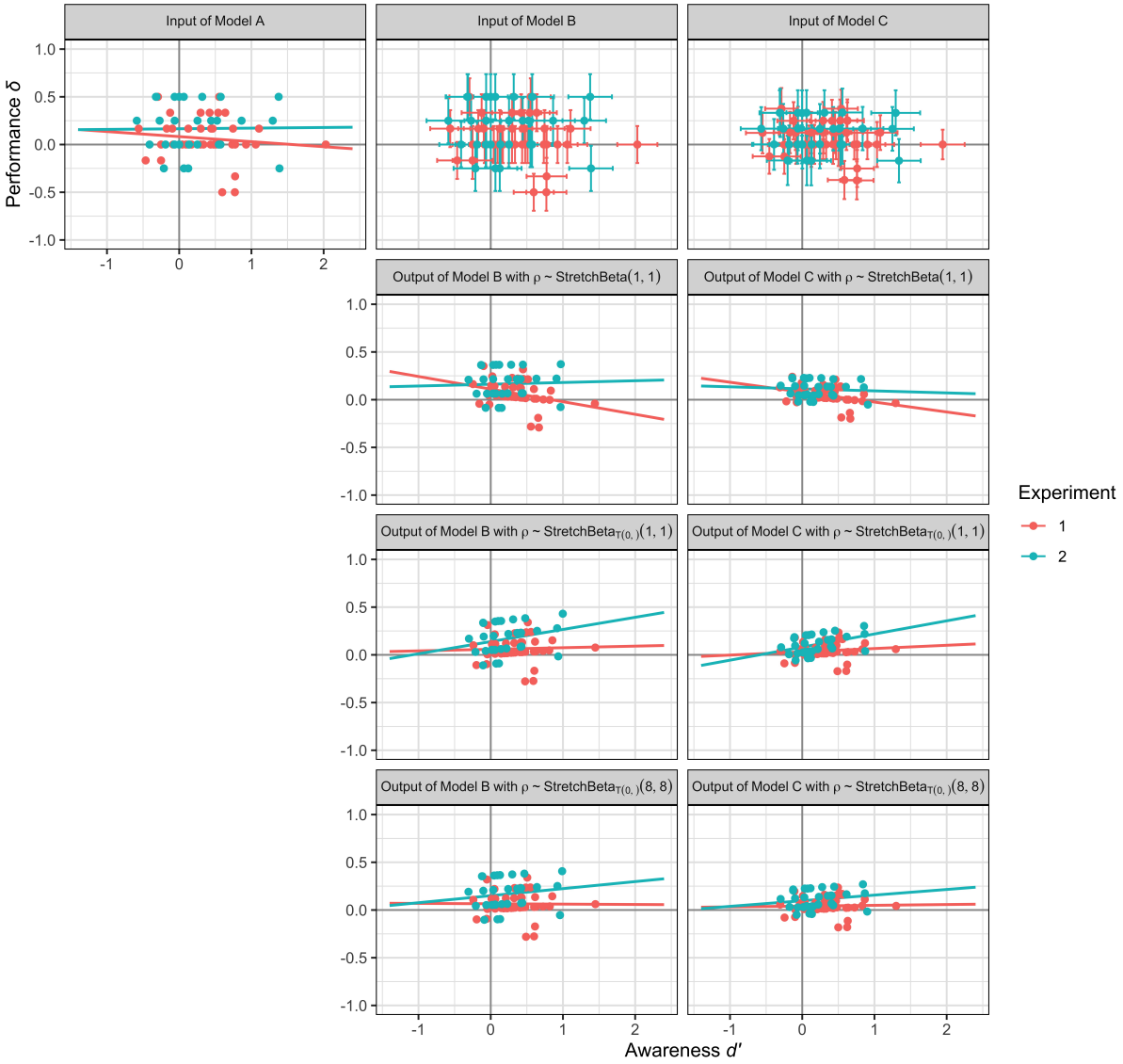
nature of the analyses they conducted on their data. In particular, we raise concerns with their analysis, and we offer a mini-tutorial on alternative and arguably better methods for analyzing correlational data in studies on unconscious mental processes. Because Salvador et al.'s data suffer from low reliabilities, we also provide advice on how to increase the accuracy of the measures in order to obtain evidence for or against a null correlation. Our intention is that this mini-tutorial will be a useful resource for future research.

A first and rather elementary observation is that failing to reject the null hypothesis is not the same as obtaining positive evidence for the null (Dienes, 2014; Gallistel, 2009; Rouder, Speckman, Sun, Morey & Iverson, 2009; Wagenmakers, 2007). Within the *Null Hypothesis Significance Testing* (NHST) framework, which Salvador et al. (2018) adopted, it is inappropriate to interpret failure to reject the null hypothesis (i.e., correlation = 0) as evidence for the null hypothesis. Simply stated, it is an inferential error to conclude that participants' memory performance was independent of their awareness of the geometrical shape's identity on masked trials (at least as measured by d'). Hence, an approach is needed which compares the extent to which the data provide evidence for the null hypothesis versus a plausible alternative hypothesis (e.g., correlation > 0).

A slightly deeper observation is to note that the correlation coefficients that Salvador et al. (2018) obtained are estimates of the true correlation. Whether or not these coefficients are significantly different from zero, it is essential to ask how precise their estimates of these parameters are. It is easy to calculate the 95% confidence interval (CI) on these estimates: $[-.40, .19]$ and $[-.35, .37]$ for Experiments 1 and 2, respectively. These CIs are compatible with quite substantial positive correlations between performance and awareness (greater, for example, than the decidedly non-trivial correlation between IQ and income; Strenze, 2007). Hence, Salvador et al. failed to rule out the possibility that performance and awareness were substantially correlated. Of course, their small sample sizes of $N = 44$ and

Figure 1

Individual Scores of Task Performance (δ) and Cue Awareness (d') for Salvador Et Al.'s (2018) Experiments 1 and 2 as Used in Models A, B, and C Reported in the Main Text



Note. First row: Model A uses the δ and d' scores as calculated in the original experiments; Model B uses the same data and additionally the standard error of measurement σ_ϵ (visualized as error bars); Model C uses δ and d' scores estimated from two Bayesian data models and the standard deviations of the individual posterior distributions σ_{ϵ_j} (visualized as error bars). Second row: Inferred δ and d' scores from Models B and C when estimating the correlation coefficient ρ. Third row: Inferred δ and d' scores from Models B and C when testing $\mathcal{H}_0\colon \rho = 0$ against $\mathcal{H}_+\colon \rho \sim \text{StretchBeta}_{\text{T}(0,)}(1, 1)$. Fourth row: Inferred δ and d' scores from Models B and C when testing $\mathcal{H}_0\colon \rho = 0$ against $\mathcal{H}_+\colon \rho \sim \text{StretchBeta}_{\text{T}(0,)}(8, 8)$. For the stretched beta distribution, see the main text. Superimposed lines represent the regression lines.

$N = 30$ in Experiments 1 and 2, respectively, make this conclusion almost inevitable.

But analyses such as these are only partly helpful in providing a fuller perspective on Salvador et al.'s (2018) data. Instead, one would also want to account for the fact that the correlation coefficient can underestimate the true strength of the association when observations suffer from measurement error (i.e., ignoring that data are imprecise or unreliable) or when parameter estimates such as SDT's d' index are treated as observations (i.e., ignoring estimation uncertainty arising because latent processes cannot be observed). In particular, ignoring uncertainty in the predictor variable can bias the regression slope towards zero—the phenomenon of *regression attenuation* (or *regression dilution*; Spearman, 1904). Hence, any correlation must be inferred in a way that disattenuates the weakening effect of uncertainty in the variables. Salvador et al. ignored measurement error, despite the fact that their indirect measure δ only included twelve and eight trials overall in Experiments 1 and 2, respectively.

Correlation coefficients can be disattenuated in the NHST framework as well as the Bayesian inferential framework. In the following, we present a Bayesian approach because it can quantify the relative evidence for both null and alternative hypotheses, allows us to include the distribution and uncertainty of data in a generative model, and is traditionally used to discretize the statistical continuum into a ternary decision space (accept, reject, and inconclusive). We now turn to the general idea behind Bayesian hypothesis testing of correlation coefficients. Thereafter, we start the reanalysis of Salvador et al.'s (2018) data with a simple Bayesian correlation model, which will then be extended twofold to account for observations contaminated by measurement error and for uncertainty in estimating the to-be-correlated variables.

Bayesian Correlation Analysis

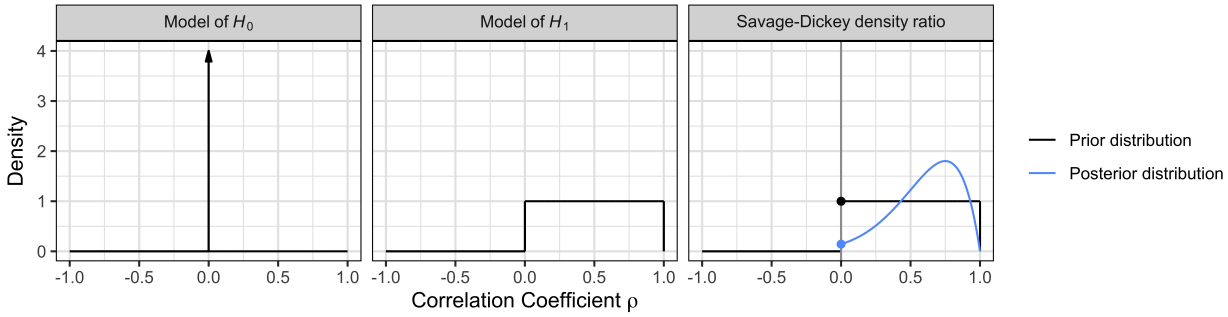
A principled way of testing hypotheses on a parameter in a Bayesian framework is to reformulate them as a question of Bayesian model comparison and to decide how convincing the observed data are with respect to each model. The competing hypotheses in unconscious-cognition research are essentially two models with a single parameter—the correlation coefficient ρ. In Salvador et al.'s (2018) case, one model predicts a null correlation between memory suppression and awareness ($\mathcal{H}_0 \colon \rho = 0$) and the other model predicts a positive correlation between them ($\mathcal{H}_1 \colon \rho > 0$).

Before seeing the data, each of the competing models makes a prediction about the value of parameter ρ and assigns each possible value a certain probability. This is quantified by a probability distribution over the range of ρ, defining the model for each hypothesis. Because $\mathcal{H}_0$ is a point null hypothesis, it can be modeled with all its probability mass at $\rho = 0$ (see left panel of Figure 2). According to $\mathcal{H}_1$, ρ can take on any value larger than 0; in a simple (but scientifically implausible) model all these values are equally likely, leading to a flat prior distribution from 0 to 1 (see middle panel of Figure 2). Other, potentially more plausible models of $\mathcal{H}_1$ are discussed later.

After seeing the data, the aggregate probability of the data can be determined by calculating how probable the data are for each parameter value (the likelihood) times the probability of that parameter value under the model (the result being the marginal likelihood). The evidence for one model rather the other is then given by how well predicted the data are by one model rather than the other. When we compare two hypotheses, we are not interested in their absolute evidence, but in their relative evidence. In order to compare the predictive performance of two hypotheses, we specifically form the ratio of their marginal likelihoods (Jeffreys, 1961). The resulting statistical ratio is called the *Bayes factor* (BF). Alternatively, the BF can be understood as the change from prior

Figure 2

Prior and Posterior Distributions in Bayesian Correlation Models to Obtain a Bayes Factor



Note. Left panel: Point null prior; the arrow indicates infinite density at zero. Middle panel: Uniform prior over the positive range of ρ. Right panel: Prior and fictional posterior distribution to calculate a Bayes factor using the Savage–Dickey method; gray vertical line at zero marks the point of interest; the Bayes factor of 7 in favor of the alternative hypothesis is illustrated by the ratio of the black and the blue dot.

model odds to posterior model odds (Kass & Raftery, 1995). The BF tells us how much more likely the data are under each hypothesis. But if the hypotheses are considered equally likely a priori, the BF can also be interpreted as how much more likely one hypothesis is over the other after having seen the data.

In order to obtain a BF, we can calculate what the posterior distributions would be for the models of the null and the alternative hypothesis when updated by the data (treating this updating as a calculational aid; the function of the model is to represent theoretical predictions, which may in fact remain unchanged). In the current case, the fact that the model distribution of ρ under the null hypothesis is nested in that of ρ under the alternative hypothesis allows us to easily compute the BF by using the Savage–Dickey density ratio (Dickey, 1971; Dickey & Lientz, 1970). Here we only need the prior distribution of ρ under the alternative hypothesis (such as a flat distribution from 0 to 1) and the posterior distribution of ρ under the alternative hypothesis, which can be obtained using a Bayesian correlation model and the observed data. We can then divide the height of the posterior distribution for ρ by the height of the prior distribution for ρ at the point of

interest (i.e., the null correlation; see right panel of Figure 2).

In order to decide which hypothesis receives more support, the absolute value of the BF, B, can be interpreted: $B_{01} = 3$ means that the data are three times more likely under the null hypothesis than under the alternative hypothesis, whereas $B_{01} = 1/3$ means that the data are three times more likely under the alternative hypothesis than under the null hypothesis, as the evidence in favor of the null hypothesis and the evidence in favor of the alternative hypothesis are inversely related ($B_{10} = 1/B_{01}$). Hence, $B_{01} = B_{10} = 1$ provides evidence favoring neither hypothesis. When comparing a null hypothesis and a directional alternative hypothesis such as in Salvador et al.'s (2018) case, B_{01} and B_{10} can be denoted as B_{0+} and B_{+0}, respectively. Although the interpretation of how likely one hypothesis is compared to another is an inherent characteristic of the BF and represented on a continuum, some benchmark BFs have been assigned shorthand labels. According to Jeffreys (1961), a BF between 1 and 3 provides "anecdotal" evidence for a hypothesis, a BF between 3 and 10 provides "substantial" evidence (though "moderate" evidence would be a better term; Lee & Wagenmakers, 2013), and a BF between 10 and 30 provides "strong" evidence.

Measurement Error in Observed Variables (Measurement Uncertainty)

Although BFs naturally provide gradual evidence for each competing hypothesis, they still do not account for correlation attenuation (Charles, 2005; Spearman, 1904). All psychological variables are measured with error. In classical test theory, the observed score on a test is the sum of its true score and an error term (Novick, 1966).[5] When the error

[5] Correspondingly, the variance of observed scores across participants is equal to the sum of the variance of true score and the error variance. To complete the foundational structure for the classical linear test-theory model, three main assumptions are required: (1) the mean of the error scores in the population is zero, (2) true scores and error are uncorrelated, and (3) the errors of different tests are

term is greater than zero, the correlation between two observed variables is always lower than the correlation between the unobserved true scores, and ignoring the error will lead to an underestimation of the true correlation. One easy solution is to adjust the observed correlation using Spearman's disattenuation formula. This adjustment is related to errors-in-variables regression models, which are standard linear regression models accounting for measurement (observational) error in the predictor variable (Carroll, Ruppert, Stefanski, & Crainiceau, 2006). If both predictor and criterion variable are measured with error, both methods result in the same disattenuation (Behseta, Berdyyeva, Olson, & Kass, 2009).

The Spearman correction for attenuation removes all measurement error from a correlation coefficient by increasing the reliabilities of each variable to 1 (Nunnally, 1970). It is calculated as $r'_{xy} = r_{xy}/\sqrt{r_{xx}r_{yy}}$ where r_{xy} is the observed correlation coefficient, r'_{xy} is the disattenuated correlation coefficient, and r_{xx} and r_{yy} are the reliabilities of the observed variables (Schmidt & Hunter, 1999). Suitable reliability estimates are test–retest correlations, or measures of internal consistencies such Cronbach's α or split-half correlations. For Salvador et al.'s (2018) datasets, we decided to calculate odd–even split-half correlations, for which performance scores on odd-numbered trials are correlated with scores on even-numbered trials. However, split-half correlations underestimate the reliability because the size of the odd and even sets is reduced compared to the test as a whole and reliability coefficients increase with the number of observations (Nunnally, 1970).

---

uncorrelated. This modest set of assumptions underlies a wide range of psychological applications and suffices to generate all test formulas necessary for our application (i.e., the Spearman–Brown prediction formula and Spearman's disattenuation formula). For simplicity, we adhere to these assumptions in Model B below, but not in Model C. The latter allows for different error variances across individuals. Alternative measurement theories exist, which can separate participant characteristics from item characteristics, and make specific assumptions concerning the functional form of observed-score, true-score, or error distributions (e.g., item-response theory; Lord, 1980; for a comparison, see Hambleton & Jones, 1993).

As a remedy, the correlation coefficient between the two halves can be adjusted to account for test length using the Spearman–Brown prediction formula: $r^*_{xx} = 2r_{xx}/(1 + r_{xx})$ where r^*_{xx} is the adjusted split-half correlation (Brown, 1910; Spearman, 1910).

When the reliability analysis was applied to Salvador et al.'s (2018) data, a striking finding emerged: The resulting adjusted reliability coefficients for the task-performance measure δ were very low with values of .26 and $-.04$ for Experiments 1 and 2, respectively. These are far below the minimum levels expected of a psychometrically-sound measure. Furthermore, because reliability is defined as the proportion of total observed variance that is due to true-score variance (i.e., the ratio of true variance to observed variance), a negative reliability estimate can only occur when the error variance is larger than the observed variance in the data—indicating a measurement problem. The values for the memory score d' were appreciably higher at .59 and .72, respectively.

From the obtained reliabilities of the δ and d' measures, the disattenuated correlation between them can be obtained using the Spearman correction mentioned above. For Experiment 1, the disattenuated correlation is $r'_{xy} = -.30$ with a 95% CI $[-1.02, .47]$.[6] The negative sign of the disattenuated correlation in Experiment 1 is a result of the surprising negative observed correlation, which points towards a sampling problem (under the assumption that participants with higher cue awareness are not less susceptible to memory suppression), and is not symptomatic of the disattenuation formula. The CI is modified the same way as the correlation coefficient, namely by applying the disattenuation formula to the lower and upper bound (lower bound$/\sqrt{r_{xx}r_{yy}} \leq \rho_{xy} \leq$ upper bound$/\sqrt{r_{xx}r_{yy}}$; Schmidt & Hunter, 2014). Because of the negative reliability coefficient of the δ measure in Experiment 2, no disattenuated correlation can be calculated. When assuming a small but positive reliability of .10 for the δ measure, the disattenuated correlation between δ and d'

[6] All reported statistics were calculated using non-rounded values and may thus differ when calculated by hand from the rounded values reported in the text.

would be no larger than $r'_{xy} = .05$ with a 95% CI of $[-1.30, 1.38]$. Hence, it seems essential to account for the effect of measurement error in both experiments, in particular when testing for a null correlation.

For the analyses reported below, we decided to use reliability estimates calculated for both experiments jointly as these are based on more data points and allow a more accurate estimate of the true reliabilities. The Pearson correlation coefficient between the δ and d' measure in both experiments analyzed together was $-.09$ with a 95% CI of $[-.31, .15]$, and the Spearman–Brown corrected odd–even split-half correlations were .15 and .65 for the δ and the d' measure, respectively. The disattenuated observed correlation resulting from the odd–even split was $r'_{xy} = -.28$ and had a wider 95% CI of $[-1.00, .47]$. Again, the negative sign of the disattenuated correlation for both experiments analyzed jointly is a result of the negative observed correlation in Experiment 1.

Note that disattenuated correlation coefficients neither improve the quality of the measure nor provide a substitute for precise measurement. They are also not directly comparable to Pearson correlation coefficients (Muchinsky, 1996). For example, if $r_{xx}r_{yy} < r^2_{xy}$, disattenuated correlations and confidence bounds outside the range of an ordinary Pearson correlation are routinely observed (for reasons, see Charles, 2005). Furthermore, using disattenuated correlation coefficients for hypothesis testing does not solve the issue that a non-significant result within the NHST framework cannot be interpreted as evidence for the null (whereas a Bayesian framework as outlined in this article allows such an inference). Disattenuated correlations do, however, help to understand whether the observed Pearson correlation between two sets of measures is low because of a true null correlation or because of measurement error, and in the case of Salvador et al.'s (2018) data, they unmistakably speak for the latter.

Uncertainty in Parameter Estimates (Estimation Uncertainty)

Apart from uncertainty in the observed variables due to measurement error, uncertainty in variables can also be interpreted as uncertainty in their estimation (e.g., Kruschke, 2011). This interpretation may be less obvious, in particular when measurement models are not the main focus of the research question. The awareness measure d' and the task-performance measure δ are estimates of discriminability and memory performance, respectively. While the former is based on SDT, the latter is the difference in the success rates of two binomial processes (number of successful recalls of the associate word out of all trials in the Think condition vs. the No-Think condition). Hence, both measures are subject to estimation uncertainty (Matzke, Ly, Selker, Weeda, Scheibehenne, Lee, & Wagenmakers, 2017).[7] In particular, the δ score reported by Salvador et al. (2018) is based on only six and four responses per condition in Experiments 1 and 2, respectively. Such small frequencies per individual lead to binomial noise in the response rates and therefore high uncertainty in the resulting parameter estimate. If this uncertainty is ignored, variability of the estimates that reflects their uncertainty may be mistaken for variability that is due to true individual differences. It follows that the correlation between two sets of uncertain parameter estimates can be severely underestimated.

When the δ and the d' measure are understood as observations of a latent construct, it is essential to deal with measurement error. However, when they are understood as model-based estimates of latent psychological processes, researchers have to deal with estimation uncertainty. Importantly, measurement error and estimation uncertainty are two sides of the same coin: Both address the fact that the exact value of a variable is unknown

[7] Note that we refrain from using the term *parameter uncertainty* here as it can also be used as an umbrella term for any type of uncertainty regarding a measured psychological variable (such as measurement errors or sampling errors). We instead follow Matzke et al. (2017) and use the term *estimation uncertainty* to refer to uncertainty associated with parameter estimates from a model.

because of trial noise, and if there are too few trials per participant, the variable becomes unreliable or the confidence in the parameter estimates for each participant is low. With more trials, the measure becomes more reliable or the parameter estimation becomes more precise. It holds for both scenarios that the larger the variances in the observed variables or estimates are, the smaller the correlation will become. It merely depends on whether δ and d' are treated as measures of true concepts or as measures of latent psychological processes. In both instances, uncertainty regarding the correlated variables will lead to low correlations between them.

It is important to emphasize that parameter (estimation) uncertainty is not the same as model uncertainty. In order to measure a latent cognitive process, researchers make use of formal models. When using such a model, the researcher hypothesizes that the underlying model structure is an accurate description of the world. To avoid using an incorrect model, selective-influence studies are conducted beforehand, in which experimental manipulations that should selectively influence only one model parameter but not others are used to assess the model's validity (Dzhafarov & Kujala, 2017; Schweikert, Fisher, & Sung, 2012). When there is more than one candidate model, different model-selection criteria—such as BFs—can be used to compare their performance. When researchers estimate parameters of their chosen model, and thereby assume the model is correct, these parameter estimates depend on certain aspects of the data that the model is being fitted to (e.g., reliabilities of the measures, number of data points). For example, a limited number of data points will lead to wider CIs, wider posterior distributions, and thus larger uncertainty in the estimates. For the analyses reported below, we decided to model the δ measure as the difference between two binomial success rates and the d' measure following a Bayesian SDT model, which are implicitly assumed to be valid models of the parameters of interest and explained in detail in the Appendix.

Reanalysis of Salvador Et Al.'s (2018) Experiments 1 and 2

Method

In order to test the hypothesis of a null correlation between the task-performance measure δ and the awareness measure d' in Salvador et al.'s (2018) data, we first need to obtain a posterior distribution for the correlation coefficient ρ and then calculate the BF for the relative evidence in favor of the null hypothesis versus the alternative hypothesis (B_{0+}). In the following, we discuss three different Bayesian models to obtain the posterior distribution for ρ (*Bayesian posterior estimation* in Models A–C), before we explain in more detail how the BFs for correlations can be calculated (*Bayesian hypothesis testing* using the Savage–Dickey density method). Model A is ready to use for any paradigm that correlates two variables under the assumption that they were measured without error, and Model B is ready to use as long as reliability estimates for both variables are available. Model C, however, requires a pre-processing step. First, posterior distributions for the to-be-correlated variables need to be obtained through two separate Bayesian estimations. These data models are tailored to the implicit and explicit task used by Salvador et al., respectively, and need to be adapted if different tasks were used. Second, information from the obtained posterior distributions serves as the new data in a ready-to-use correlation model.

Model A: Standard correlation. The first model is used to obtain a posterior distribution for the standard Pearson correlation coefficient (Pearson, 1895) and is displayed graphically in the left panel of Figure 3. The observed variables δ and d' for each participant i (vectorized as $\boldsymbol{x}_i$ in the model) are sampled from a bivariate normal distribution, which takes the means μ_δ and $\mu_{d'}$ of both variables and their variances σ^2_δ and $\sigma^2_{d'}$ as parameters. The variances are combined with the correlation coefficient ρ, leading to covariance matrix Σ.

As suggested by Lee and Wagenmakers (2013) in Chapter 5.1, the means are assigned normal priors, the variances are assigned inverse gamma priors, and the correlation coefficient itself is given by a uniform distribution from -1 to 1. The priors are scientifically informed by anticipating the overall scale (numerical unit) of the data (Dienes & McLatchie, 2018; Etz, Haaf, Rouder, & Vandekerckhove, 2018; Lee & Vanpaemel, 2018).[8] Because δ can take on values from -1 to 1, μ_δ is assigned a normal prior with a mean of 0 and a variance of .25 truncated from below at -1 and above at 1, and σ^2_δ is assigned an inverse gamma prior with shape and scale parameters of 0.5.[9] Although d' can take on the entire range of real numbers, values below -3 and above 3 are considered very low and very high, respectively. Hence, $\mu_{d'}$ is assigned a normal prior with a mean of 0 and a variance of 2, and $\sigma^2_{d'}$ is assigned an inverse gamma prior with shape and scale parameters of 0.1.[10] These decisions would need to be revisited if the model were applied to a different research paradigm.[11]

---

[8] Model A leads to the same results as the correlation model that is implemented in the statistical software package JASP (JASP Team, 2020; Wagenmakers, Love, Marsman, Jamil, Ly, Verhagen, & Morey, 2017) when the group means and variances are assigned priors with a very high spread, such as $\mu_j \sim \text{Normal}(0, 1000)$ and $\sigma^2_j \sim \text{InvGamma}(.001, .001)$ with $j = \{1, 2\}$ for the two to-be-correlated variables. However, these distributions assign prior evidence to values outside the range of the variables used in the hypothesis tested by Salvador et al. (2018).

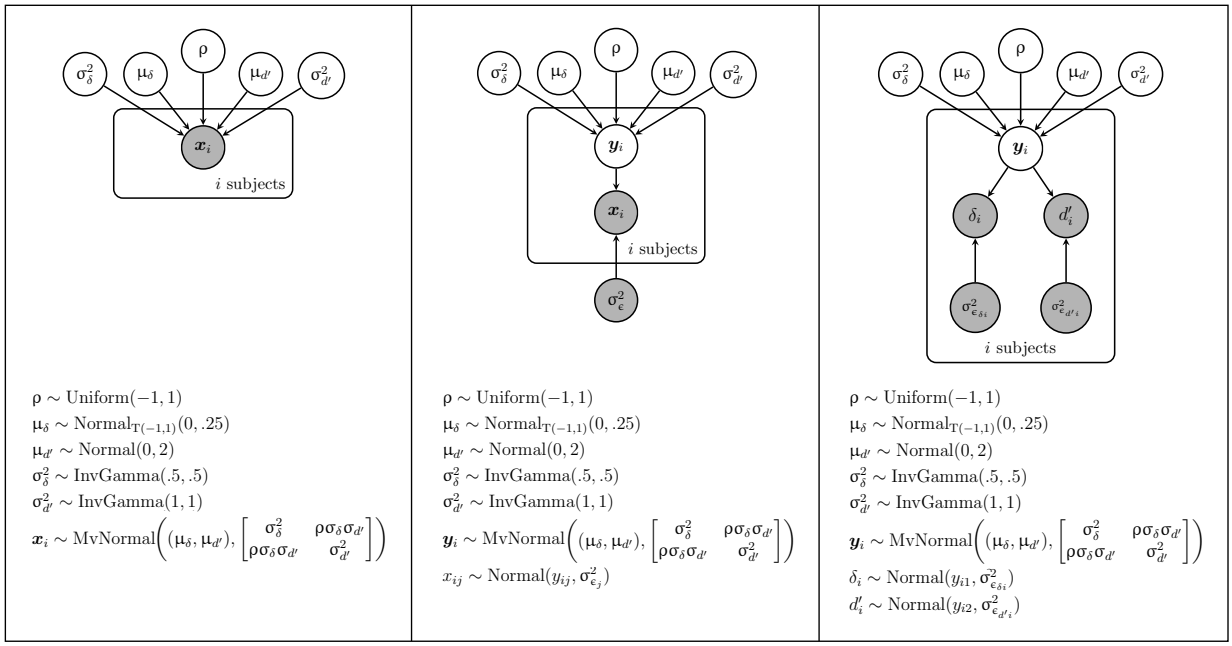
[9] Although the δ measure is based on binomial processes and bounded between -1 and 1, the pairs of δ or d' values are modeled as draws from a bivariate normal distribution. This decision is easily justified. First, with a sample size of $N \geq 30$ in each of Salvador et al.'s (2018) experiments, the sampling distribution of μ_δ will approach normality according to the central limit theorem as applied to binomials. Second, all observed δ values were within the interval $[-.50, .50]$ and thus far away from the boundaries, such that a transformation to the real line was not necessary.

[10] Note that normal distributions are commonly parametrized using the mean and the variance. However, the MCMC sampler Stan (Stan Development Team, 2018) uses the standard deviation, whereas the MCMC sampler JAGS (Plummer, 2013) uses the precision (i.e., the inverse of the variance). JAGS also uses the inverse of the covariance matrix for multivariate normal distributions.

[11] Note that JAGS has difficulties sampling when variance priors are allowed to start at exactly zero. Hence, it is often advisable to raise the lower bound slightly, while trying to retain the overall shape of the prior.

Figure 3

Graphical Depictions of the Three Bayesian Models for the Pearson Correlation Coefficient



Note. Left panel: Bayesian model for the standard Pearson correlation as implemented in Model A. Middle panel: Bayesian correlation model accounting for measurement error as implemented in Model B. Right panel: Bayesian correlation model accounting for estimation uncertainty as implemented in Model C. Shaded nodes represent observed variables, white nodes represent latent parameters, plates represent replications over subject index i, bold parameters indicate vectors, and arrows indicate dependencies between nodes (e.g., the true correlation coefficient, the means, and the variances all influence the observed data $\boldsymbol{x}$ in Model A directly). The error variances σ^2_ϵ must be provided by the user.

The reader may have noticed that the alternative hypothesis is directional, which assumes a parameter prior for the correlation coefficient that is a uniform distribution from 0 to 1. However, we focus here on estimating the correlation coefficient. For the purpose of testing the model of $\mathcal{H}_0$ against a model of $\mathcal{H}_+$, we will later revise this assumption. In either case, because Model A infers the standard Pearson correlation, it assumes that both variables are measured with perfect accuracy (i.e., without measurement error). This assumption, of course, is likely to be false in most psychological research, leading to correlation attenuation.

Model B: Correlation when accounting for measurement uncertainty. The second model, displayed in the middle panel of Figure 3, extends the first model by adjusting the correlation coefficient for attenuation as suggested by Behseta et al. (2009). To do so, the observed variables δ and d' for each participant i are modeled as draws from two separate normal distributions, $j = \{1, 2\}$, with the mean equal to the unobserved true δ or d' of that person (vectorized as $\boldsymbol{y}_i$ in the model) and variance equal to the error variance of the respective variable. Because the individual error variances are unknown, it is necessary to adapt Behseta et al.'s model in such a way that it takes an estimate of the error variance for the entire sample of participants as input (similar to the model proposed by Lee & Wagenmakers, 2013, Chapter 5.2). As an estimate for the error variance, we use $\sigma_\epsilon^2 = (1 - r_{xx})\sigma_x^2$ where r_{xx} is the reliability of the measure and σ_x^2 is the variance of the measure (Nunnally, 1970). As reliability estimates, we used the Spearman–Brown corrected odd–even split-half correlations calculated for both of Salvador et al.'s (2018) experiments together, which avoids the negative reliability estimate of the δ measure observed in Experiment 2 and bases all reliability estimates on a larger sample size. The error variances for δ were .04 and .06, and the error variances for d' were 0.07 and 0.09 for Experiments 1 and 2, respectively (see error bars in the top row of Figure 1).[12]

In contrast to Model A, the true values of δ and d', not the observed values, are modeled as draws from a bivariate normal distribution. The joint distribution acts as prior to adjust extreme individual observed values caused by measurement error by making them more moderate. This makes Model B hierarchical. The observed variables will be shrunk towards their respective group mean and the degree of shrinkage is determined by their corresponding error variance. The correlation coefficient is then computed for the shrunken

[12] Note that although the data used for parameter estimation in a Bayesian framework cannot be used to inform the priors, the error variances used to inform the priors here are based on the odd–even split-half reliabilities, which in turn are based on data not used in the estimation of the correlation coefficient.

variables and thereby automatically adjusted for the error in measurement. As in Model A, the priors for both group means are still modeled as normal distributions, the priors for their variances as inverse gamma distributions, and the correlation coefficient as a uniform prior from -1 to 1.

Because Model B is hierarchical and the sample sizes are small, it is pivotal to consider the priors of the variances carefully. For rich data, the posteriors will be dominated by the data, such that the choice of prior is less critical. However, for sparse data, the posteriors will be very sensitive to prior information (see Gelman, Carlin, Stern, Dunson, Vehtari, & Rubin, 2014, Chapter 5). In the latter case, an improperly scaled prior on a variance can impose a dependency between the correlation and the variance (see Tokuda, Goodrich, Van Mechelen, Gelman, & Tuerlinckx, 2020). The same dependency holds true if priors are not put on the marginal variances and the correlation coefficient, but instead on the true covariance matrix Σ. For example, Behseta et al. (2009) put an inverse Wishart prior on Σ (see also Rouder, Kumar, & Haaf, 2019). Although putting a prior on Σ is computationally faster, we accept the costs of slower computation for having the flexibly to assign priors directly to the variances, and more importantly, to the correlation coefficient (Barnard, McCulloch, & Meng, 2000; Matzke et al., 2017).

Model C: Correlation when accounting for estimation uncertainty. Instead of accounting for uncertainty in measurement, Model C accounts for uncertainty in parameter estimation. In this model, the δ and d' values are not treated as observed variables measuring a psychological construct, but as parameters estimated from two separate Bayesian models. Following Matzke et al. (2017), the posterior distributions for the model parameters are obtained for each participant prior to estimating the correlation. Thereafter, the correlation is inferred using the Bayesian model proposed by Behseta at al. (2009)—but this time the observed δ and d' values are replaced by point estimates (i.e., the

individual means of the posterior distributions) and the error variances are replaced by the uncertainty of the point estimates (i.e., the individual variances of the posterior distributions). The correlation part of Model C is displayed graphically in the right panel of Figure 3. It is identical to the correlation part of Model B, except that the error variances are now different for each participant i. This allows inferring a posterior distribution for the correlation coefficient that explicitly addresses the uncertainty associated with the individual parameter estimates.

Note that the Bayesian correlation model can be used to analyze data in any paradigm that correlates two variables. The Bayesian data models for the correlated parameters, however, are task-specific and need to be adapted if other measures are to be used. The data models together with the correlation model constitute a generative modeling approach: If there is an association between two latent parameters and those parameters generate data through other processes, all of that should be modeled and Bayesian inference can be applied to estimate the parameters and infer their correlation (Lee & Vanpaemel, 2018). The Bayesian data models for δ or d' in Salvador et al.'s (2018) experimental task are described in detail in the Appendix, followed by a brief explanation of why we did not implement a Bayesian hierarchical model that estimates participant-level data and the correlation simultaneously.

Bayesian hypothesis testing: Bayes factors. In order to compare two competing hypotheses, it is necessary to evaluate their relative evidence. One way to do this is by calculating the BF using the Savage–Dickey density ratio rule (Dickey, 1971; Dickey & Lientz, 1970), that is, the height of the posterior for ρ at point zero is divided by the height of the prior for ρ at the same point. The height of the posterior for ρ is obtained from one of the three correlation models described above (after the prior of ρ given by the theoretical model of $\mathcal{H}_1$ has been updated by the data), whereas the height of the prior is given by the

theoretical model of $\mathcal{H}_1$ (before it has been updated by the data). If the model of $\mathcal{H}_1$ (the prior) and the data (the likelihood) clash, the density will be substantially lower for the posterior as compared to the prior.

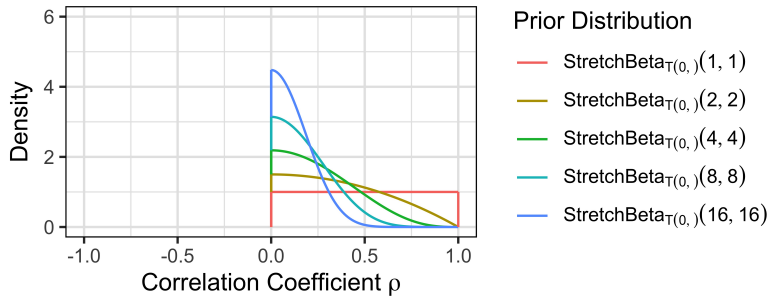
The required BF in the case of Salvador et al.'s (2018) hypothesis test is given by $B_{0+} = p(D|\mathcal{H}_0)/p(D|\mathcal{H}_+)$. This BF is one-sided because we are comparing a null hypothesis about parameter ρ (i.e., $\mathcal{H}_0 : \rho = 0$) against a directional alternative hypothesis (i.e., $\mathcal{H}_+ : \rho > 0$). Different models of $\mathcal{H}_+$ can be thought of, which are formalized through different prior distributions over the range of possible values for ρ—in our case through the family of stretched beta distributions truncated from below at zero (stretched half-beta priors) depicted in Figure 4.[13] Because standard beta distributions only supply a positive probability density to every value in the interval $[0, 1]$, but correlation coefficients range over the interval $[-1, 1]$, the beta distribution can be stretched to cover the required range. To do so, a beta random variable X defined on the interval $[0, 1]$ is shifted and rescaled to obtain the beta random variable X' on the interval $[-1, 1]$ that still integrates to 1 through the following transformation: $X' = -1 + 2X$. In order to test hypotheses such as Salvador et al.'s directional alternative hypothesis, the stretched beta distribution can then be truncated from below at zero. Although it may seem unnecessary to first stretch and then truncate the distribution to the same range as before, only this allows the user to choose between different default priors with mode set at zero, which ensures that smaller correlations close to the critical null correlation are more probable than larger ones.

Note that a prior with more mass on smaller values (and equally a flat prior bounded from above at a value smaller than 1) in the estimation would pull the parameter estimate

[13] Note that the depicted distributions are probability density functions. They represent a selection of possible alternative hypotheses. Each hypothesis states how likely different values of the correlation coefficient are. They do not represent distributional assumptions. For example, we do not postulate that the sampling distribution of the correlation coefficient or the true distribution of correctional coefficients in the population of implicit-cognition studies follows a stretched beta distribution.

Figure 4

Different Stretched Half-Beta Prior Distributions Over the Correlation Coefficient



Note. The different distributions illustrate different models of the alternative hypothesis. The distributions are parameterized by two shape parameters, which are set to be equal and called α. The width of the distribution κ is equal to the inverse of α. The more mass is given to small values of ρ, the more the estimate will be pulled towards smaller values of ρ.

towards smaller values of ρ. Only a non-informative prior allows accurate posterior estimation by being as vague as possible. Because estimation and hypothesis testing serve different purposes, they can and often should use different parameter priors. Hence, in order to obtain an unbiased posterior estimate of ρ but a one-sided BF B_{0+}, we retain the flat prior from -1 to 1 in the models during estimation and allow the user to choose a stretched half-beta prior for hypothesis testing as their model of $\mathcal{H}_+$. To spare the user from making the necessary changes in the model code, our code follows Matzke et al. (2017) and computes the one-sided BF B_{0+} from the two-sided BF B_{01}. This involves fitting a stretched beta distribution to the posterior data to construct a density estimator of the posterior distribution, applying the Savage–Dickey rule to obtain B_{01}, and correcting B_{01} using the proportion of samples from the posterior distribution that are consistent with the order restriction $\rho_0 < \rho_+$ under the selected $\mathcal{H}_+$ (for details, see Morey & Wagenmakers, 2014). The resulting BF is equivalent to changing the prior in the estimation to the selected stretched half-beta prior, which expects only positive values, and applying the Savage–Dickey rule on the new posterior of ρ.

Selecting an alternative hypothesis: Prior distributions. When formalizing the alternative hypothesis for a specific research question about a correlation coefficient, one important consideration is what the theory being tested predicts with regard to which values of ρ are more likely than others (Dienes, 2019). When choosing from the family of default priors, existing knowledge of the size of the correlation can be used. If all positive correlations are equally plausible a priori, a non-informative flat prior from 0 to 1 can be used that gives equal weight to all values between 0 and 1, such that the model of the alternative hypothesis would be $\mathcal{H}_+ \colon \rho \sim \text{Uniform}(0, 1)$, or equivalently, $\mathcal{H}_+ \colon \rho \sim \text{StretchBeta}_{\text{T}(0,)}(1, 1)$. However, if small positive correlations are more likely, more weight should be given to small relative to large correlations. This can be accomplished by an informative unimodal prior that sets the mode at zero, such that the models of the alternative hypothesis can be constructed as $\mathcal{H}_+ \colon \rho \sim \text{StretchBeta}_{\text{T}(0,)}(\alpha, \alpha)$ with two identical shape parameters α. The width of the distribution κ is inversely related to its shape, such that $\kappa = 1/\alpha$. Hence, $\kappa = 1$ generates StretchBeta$(1, 1)$, $\kappa = 1/2$ generates StretchBeta$(2, 2)$, and so forth. If $\kappa > 1$, correlations closer to 1 are more plausible than those closer to 0, whereas if $\kappa < 1$, correlations closer to 0 are more plausible.

For expositional purposes, we decided to reanalyze Salvador et al.'s (2018) data using two different models of $\mathcal{H}_+$. First, we will assume an uninformed model of the alternative hypothesis, $\mathcal{H}_+ \colon \rho \sim \text{StretchBeta}_{\text{T}(0,)}(1, 1)$. Second, we will assume an informed model of the alternative hypothesis, $\mathcal{H}_+ \colon \rho \sim \text{StretchBeta}_{\text{T}(0,)}(8, 8)$. For a detailed account of how the informed model was selected, we refer to the Appendix.

Results

All computations were performed in the statistical programming language R (R Core Team, 2018) in combination with the MCMC sampler JAGS (Plummer, 2013) adapting

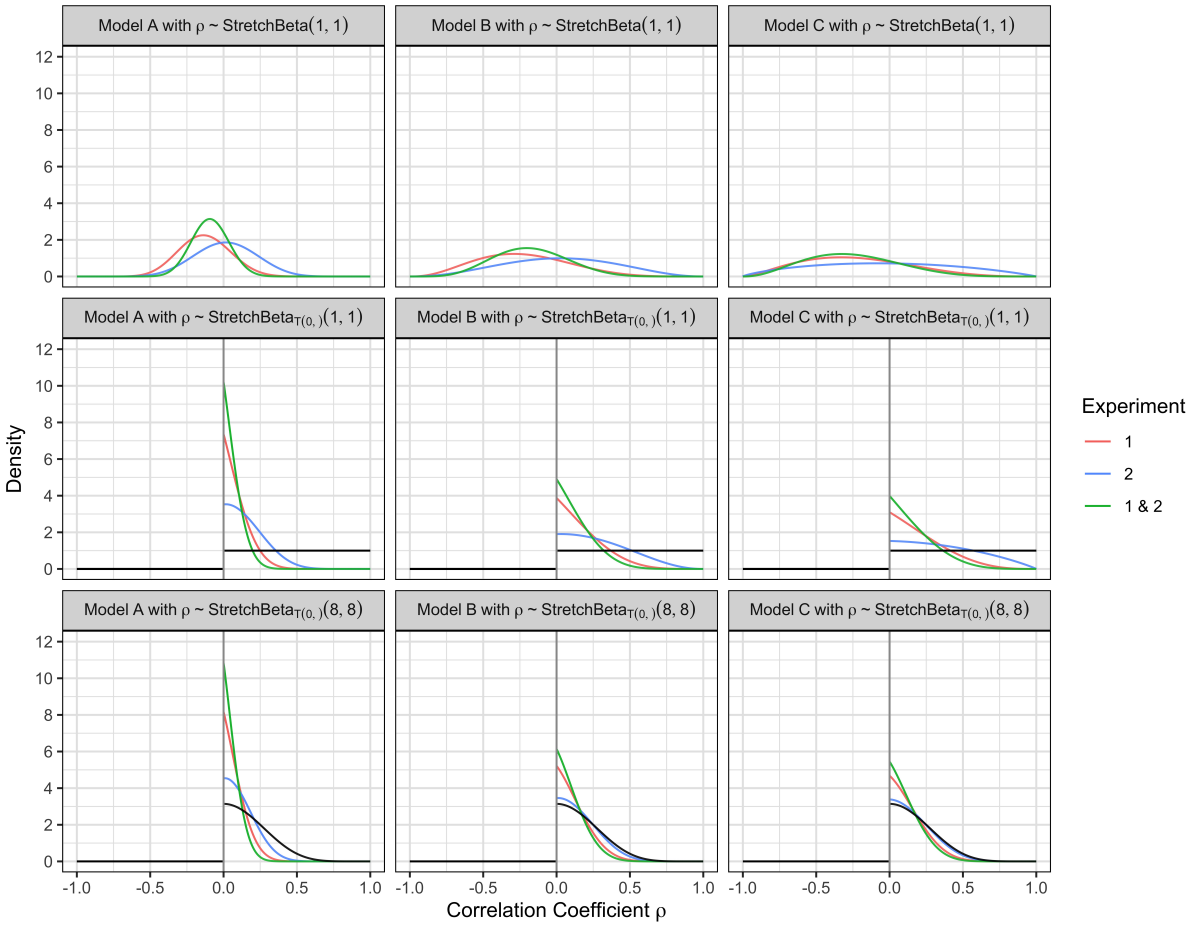
code provided by Lee and Wagenmakers (2013) and Matzke et al. (2017), and cross-checked using the MCMC sampler Stan (Stan Development Team, 2018). All code is available through the Open Science Framework at `https://osf.io/pq7ug/`. Details of the model-fitting routines can be found in the Appendix. We report the results for the experiments by Salvador et al. (2018) analyzed separately as well as jointly for a larger dataset. The second row of Figure 1 shows the individual values that are used to compute the correlation coefficient after being shrunk by Models B and C.

The posterior distributions of the correlation coefficient in each dataset and each model are visualized in the top row of Figure 5. Different point estimates and 95% credibility intervals are reported in Table 2. The point estimates correspond to the mean, median, and mode (sometimes called MAP for maximum a posteriori probability) of the posterior distribution based on flat priors from -1 to 1; the credibility intervals correspond to equal-tailed intervals (ETIs) including the median, and highest density intervals (HDIs) including the mode and all points with higher probability density than points outside the interval. BFs testing the null hypothesis against the alternative hypothesis of a positive correlation based on two different stretched half-beta priors are reported in Table 3. To highlight the importance of specifying the alternative hypothesis using relevant prior information, we present BFs based on two different prior distributions over the correlation coefficient. The first prior formalizes the uninformed model of the alternative hypothesis, $\mathcal{H}_+ \colon \rho \sim \text{StretchBeta}_{\text{T}(0,)}(1, 1)$, and the second prior formalizes the informed model of the alternative hypothesis using the disattenuated upper CI bound obtained from the "conscious" condition as the maximum possible correlation, $\mathcal{H}_+ \colon \rho \sim \text{StretchBeta}_{\text{T}(0,)}(8, 8)$ (see the Appendix for details). Accordingly, the middle and bottom rows of Figure 5 show the posterior distributions based on these two priors.

Because the choice of the model of $\mathcal{H}_+$ (i.e., the prior over the positive range of ρ) can

Figure 5

Posterior Distributions of the Correlation Coefficient in Salvador Et Al.'s (2018) Experiments 1, Experiment 2, and Both Combined



Note. Top row: Models A to C with a flat prior from −1 to 1 on ρ for parameter estimation. Middle row: Models A to C with a flat prior from 0 to 1 (black distribution) on ρ for default hypothesis testing. Bottom row: Models A to C with a stretched half-beta prior of width $\kappa = 1/8$ (black distribution) on ρ for informed hypothesis testing. Gray vertical lines mark the point of interest where the height of the posterior distribution is compared to the height of the prior distribution in order to obtain the Bayes factor.

influence the conclusion drawn from a BF, we report a robustness region (RR) with each BF (Dienes, 2019). This interval provides information on how robust the selected prior is in comparison to other priors from the same distributional family (McLatchie, 2018), and is based on the idea of a multiverse analysis (Steegen, Tuerlinckx, Gelman, & Vanpaemel,

Table 2

Reliability Estimates, Correlation Coefficients, and Bayesian Point Estimates of the Relationship Between Recall Performance and Cue Awareness in Salvador Et Al.'s (2018) Experiments 1, 2, and Both Combined

		Experiment 1	**Experiment 2**	**Experiments 1 & 2**
NHST	Reliability of δ	.26	−.04	.15
	Reliability of d'	.59	.72	.65
	Correlation [95% CI]	−.12 [−.40, .19]	.01 [−.35, .37]	−.09 [−.31, .15]
	Disattenuated correlation [95% CI]	−.30 [−1.02, .47]	n/a	−.28 [−1.00, .47]
Model A	Posterior mean	−.13	.01	−.09
	Posterior median [95% ETI]	−.13 [−.45, .22]	.02 [−.39, .41]	−.09 [−.33, .16]
	Posterior mode [95% HDI]	−.14 [−.46, .21]	.02 [−.39, .41]	−.09 [−.33, .16]
Model B	Posterior mean	−.23	.02	−.17
	Posterior median [95% ETI]	−.25 [−.76, .40]	.02 [−.66, .70]	−.18 [−.68, .32]
	Posterior mode [95% HDI]	−.29 [−.79, .36]	.03 [−.66, .70]	−.20 [−.65, .30]
Model C	Posterior mean	−.23	−.04	−.26
	Posterior median [95% ETI]	−.26 [−.81, .49]	−.04 [−.88, .80]	−.28 [−.80, .37]
	Posterior mode [95% HDI]	−.38 [−.85, .44]	−.07 [−.89, .80]	−.32 [−.83, .33]

Note. NHST = null-hypothesis significance testing; reliability = odd–even split-half correlation corrected for length using the Spearman–Brown formula; CI = frequentist confidence interval, ETI = Bayesian equal-tailed interval, HDI = Bayesian highest density interval. Due to the negative reliability estimate for δ in Experiment 2, no disattenuated correlation coefficient can be calculated. Bayesian point estimates and credibility intervals are based on a flat prior from −1 to 1 on ρ.

2016). In our case, the family of stretched half-beta distributions allows for different widths of the prior. The robustness region is notated as $RR[\kappa_{\text{min}}, \kappa_{\text{max}}]$ where κ_{min} is the minimum beta width and κ_{max} is the maximum beta width that lead to the same conclusion. For example, $B_{0+} = 4.0$, $RR[0.50, 1.30]$ shows that data are four times more likely under $\mathcal{H}_0$ than under $\mathcal{H}_+$ and thus provide moderate evidence for $\mathcal{H}_0$. The RR shows that the conclusion of moderate evidence holds for priors with a beta width between 0.50 and 1.30. For beta widths smaller than 0.50, the data would provide anecdotal evidence for the null, whereas for beta widths larger than 1.30, they would provide strong evidence for the null. Importantly, we are not interested in whether the BFs across different priors agree in their

Table 3

Bayes Factors [And Robustness Regions] Testing the Hypothesis of a Null Correlation Against Different Hypothesis of a Positive Correlation for Salvador Et Al.'s (2018) Experiments 1, 2, and Both Combined

			Experiment 1	**Experiment 2**	**Experiments 1 & 2**
Model A	$\text{StretchBeta}_{\text{T}(0,)}(1,1)$	$\kappa = 1$	7.4 [0.18, 1.62]	3.5 [0.75, 4.64]	10.2 [0.97, 4.68]
	$\text{StretchBeta}_{\text{T}(0,)}(8,8)$	$\kappa = 1/8$	2.6 [0.00, 0.17]	1.4 [0.00, 0.74]	3.5 [0.10, 0.96]
Model B	$\text{StretchBeta}_{\text{T}(0,)}(1,1)$	$\kappa = 1$	3.9 [0.61, 4.24]	1.9 [0.00, 2.55]	4.9 [0.37, 2.91]
	$\text{StretchBeta}_{\text{T}(0,)}(8,8)$	$\kappa = 1/8$	1.7 [0.00, 0.60]	1.1 [0.00, 2.55]	1.9 [0.00, 0.36]
Model C	$\text{StretchBeta}_{\text{T}(0,)}(1,1)$	$\kappa = 1$	3.1 [0.91, 6.02]	1.5 [0.00, 5.35]	4.0 [0.55, 4.26]
	$\text{StretchBeta}_{\text{T}(0,)}(8,8)$	$\kappa = 1/8$	1.5 [0.00, 0.90]	1.1 [0.00, 5.35]	1.7 [0.00, 0.54]

Note. The different stretched half-beta priors on ρ implementing different models of $\mathcal{H}_+$. A stretched half-beta distribution with shape parameters of 1, $\text{StretchBeta}_{\text{T}(0,)}(1,1)$, is equivalent to a uniform distribution from 0 to 1, $\text{Uniform}(0,1)$. Parameter κ represents the width of each distribution.

evidentiary conclusion (evidence in favor, evidence against, or no evidence), but rather whether an evidentiary conclusion is robust in the sense that the range of beta widths spans much of the range of scientifically plausible beta widths. Given that correlations above .60 are scientifically implausible given the conscious trials of Salvador et al.'s (2018) data (see the Appendix), any width above 1 can be ruled out because such widths give more plausibility to values closer to 1 than those closer to 0. More specifically, any width even above 0.50 should be ruled out because such a width still assigns 21% of its mass to correlations above .60 (cf. Figure 4).

Model A: Standard correlation. The posterior distributions of the Pearson correlation coefficient using Model A were normal in shape with their highest density close to zero (see the top row of Figure 5). In the case of symmetric unimodal distributions, the distribution's mean, median, and mode are identical, and it does not matter which is chosen as the Bayesian point estimate of the parameter of interest. The means of the posterior distribution for ρ were $-.13$ for Experiment 1, .01 for Experiment 2, and $-.09$ for Experiments 1 and 2 jointly, and thus almost identical to the Pearson correlation

coefficients (see Table 2). The 95% credibility intervals reported in Table 2 were wide, and included correlations of small and medium size according to Cohen (1988) in the negative and positive direction.

The BFs for the null hypothesis that the correlation coefficient is 0 against the alternative hypothesis that it is somewhere between 0 and 1 with equal probability were 7.4, 3.5, and 10.2 for Experiments 1, 2 and both together, respectively (see Table 3 and middle row of Figure 5). This means that the data were about seven, four, and ten times more likely under the null hypothesis than under the uninformed alternative hypothesis. These results can be interpreted as moderate support for the null hypothesis according to Jeffrey (1961) in the case of Experiments 1 and 2, and as strong support in case of both experiments analyzed jointly. However, the latter interpretation is based on a beta width that is extremely close to the lower bound of the RR, which in turn excludes scientifically plausible widths and shows that the evidentiary conclusion is not robust. The BFs for the null hypothesis against the alternative hypothesis that the correlation coefficient is somewhere between 0 and .60 with higher probabilities for small values were 2.6, 1.4, and 3.5, respectively (see Table 3 and the bottom row of Figure 5). For both experiments together, the BF can be interpreted as providing moderate support for the null hypothesis when compared to the informed alternative hypothesis. For Experiments 1 and 2, the BF provided anecdotal evidence for the null hypothesis. However, the BF for Experiment 2 was very close to 1, suggesting no support for either the null or the informed alternative hypothesis. Furthermore, the minima of the RRs for the conclusion that the evidence was not good enough in Experiments 1 and 2 were 0, which means that this conclusion holds for any reasonable small width. Taken together, the RRs for "evidence in favor of $\mathcal{H}_0$" and the RRs for "not good enough evidence" all included scientifically possible beta widths. So neither conclusion is robust. Although the overall evidence could favor unconscious memory suppression (given some width assumption), the underlying model assumes that task

performance δ and awareness measure d' were measured without error. To assess the consequences of that assumption, we must look at Model B.

Model B: Correlation when accounting for measurement uncertainty. As soon as measurement error was considered by using Spearman–Brown corrected odd–even split-half correlations as reliability estimates to obtain an error variance estimate for each variable (see error bars in the top row of Figure 1), the individual measurements were shrunk towards their respective group mean with the degree of shrinkage determined by their respective group error variance (see second column of Figure 1). As a result, the posterior means for ρ became more extreme with values of $-.23$ for Experiment 1, .02 for Experiment 2, and $-.17$ for both experiments together (see Table 2 for the medians and modes). The posterior distributions were flatter (see the top row of Figure 5) and the credibility intervals were much wider (see Table 2) than in Model A.

The evidence in favor of the null hypothesis provided by the BFs diminished noticeably, as the BFs reduced to 3.9, 1.9, and 4.9 for Experiments 1, 2, and both combined, respectively, when using the uninformed model of $\mathcal{H}_+$ (see Table 3 and the middle row of Figure 5). Although this still implied moderate support for a null correlation in case of Experiment 1 and both experiments together, we should put less trust in the null hypothesis. When using the informed model of $\mathcal{H}_+$, the BFs were 1.7, 1.1, and 1.9, respectively (see Table 3 and the bottom row of Figure 5). This means that all datasets provided no support for either the null hypothesis or the alternative hypothesis. Regarding the robustness of the conclusion that the evidence was not good enough, the maxima of the RRs were larger or not far off 0.50 (i.e., the maximum scientifically plausible width) and all RRs included any smaller width (i.e., most of the scientifically plausible widths). In sum, shrinking the δ and d' estimates towards the group mean in order to infer the disattenuated correlation coefficient changed the interpretation of the data: The data did not lend strong

support to the null hypothesis when the model of $\mathcal{H}_+$ was uninformed and did not lend any support to either hypothesis when the model of $\mathcal{H}_+$ was informed.

Model C: Correlation when accounting for estimation uncertainty. When the δ and d' measures were not treated as observations, but as parameter estimates, and their respective posterior distributions were used to account for the uncertainty in them, the support for a null correlation was even further reduced. In Step 1, parameter estimates and their individual error variances were obtained from the Bayesian data models described above (see data points and error bars in the first row of Figure 1). In Step 2, the obtained parameter estimates were used as data and shrunk towards their respective group mean in the Bayesian correlation model with the degree of shrinkage determined by the posterior error variances (see the second row of Figure 1). The degree of shrinkage was comparable to the shrinkage in Model B (cf. first and second row of Figure 1). However, the posterior means for ρ were more extreme than in Model B with values of $-.23$ for Experiment 1, $-.04$ for Experiment 2, and $-.26$ for both experiments together. Although Experiment 2 now also showed a negative correlation, this can be explained by the much flatter posterior distributions (see the top row of Figure 5) with posterior medians and modes moving far away from 0 (see Table 2). The credibility intervals were also wider, spanning almost the entire range of possible values for a correlation coefficient in Experiment 2 and a considerable range of possible values in the other datasets (see Table 2).

The BFs for comparing the null model against the uninformed model of $\mathcal{H}_+$ were calculated as 3.1, 1.5, and 4.0 (see Table 3 and the middle row of Figure 5), while the BFs for comparing the null model against the informed model of $\mathcal{H}_+$ were calculated as 1.5, 1.1, and 1.7 (see Table 3 and the bottom row of Figure 5) for Experiments 1, 2, and both combined, respectively. This means that support for the null hypothesis was even further diminished. It was also safe to conclude now that the experiments analyzed jointly showed

no support for a null correlation, because the RR for the beta width spanned the scientifically plausible range of 0 to 0.50. Hence, shrinking the δ and d' estimates towards the group mean to account for estimation uncertainty when inferring the correlation coefficient also changed the overall interpretation of the data away from support of the null hypothesis, but towards the interpretation that no conclusion regarding the existence of unconscious memory suppression can be drawn.

Discussion

We have presented three models to test for a null correlation between two variables in a Bayesian framework. While one model assumes that both variables are measured without error, the other two models account for measurement or estimation uncertainty in both correlated variables post hoc. A common case scenario, in which the correlation between two uncertain variables constitutes the main research question, is the issue of whether there are unconscious influences on behavior. As an example, we used the memory-suppression measure δ and the cue-awareness measure d' from two datasets reported by Salvador et al. (2018). Posterior distributions of the correlation coefficient ρ were obtained from fitting the Bayesian models to the data and used to calculate BFs for the hypothesis that there is a null correlation between the true δ and d' against the alternative hypothesis of a positive correlation. Because different models of the alternative hypothesis can be imagined, we compared a non-informative flat prior on the correlation coefficient (uninformed BF) and a prior using additional information provided by the research paradigm (informed BFs).

The uninformed BFs for Model A, which do not take any uncertainty into account, show some support for a null correlation, whereas the informed BFs of Model A and all BFs of Models B and C, which do take uncertainty into account, are less supportive or inconclusive (i.e., the question cannot be answered with the data at hand). Importantly, for

Models B and C, the evidentiary interpretations provided by these BFs were robust to changes in beta widths (apart for the uninformed BF of Model C in Experiment 1). Taken together, Salvador et al.'s (2018) data do not provide information to answer their research question. Hence, the data do not provide convincing evidence for the conclusion that memory suppression is conscious or unconscious—and this is independent of the prior assigned to ρ.

The absence of convincing evidence on which hypothesis is supported by Salvador et al.'s (2018) studies was due to non-diagnostic data. This leads to the question of what would make the data diagnostic. Better measures? More participants? If the paradigm cannot be changed, the reliability of the measures and the certainty in the individual parameter estimates cannot be improved. However, increasing the sample size will at least improve the certainty in the group-level parameter estimates. Also, a large sample size helps the models deal with unreliable measures as it increases the power of the statistical test (Cohen, 1988). Therefore, before recommending the models as general tools for analyzing correlational data, we need to show that they are able to provide conclusive results when there are enough data points—even when the data are unreliable.

Correlation Recovery and Bayes Factor Sensitivity: Simulation Study

The effect of unreliable performance measures and small sample sizes on the sensitivity of the BF can be studied systematically using simulated data. The following simulation study had two main goals. First, we wanted to know whether the new models can recover the true correlation, even when there is a substantial amount of measurement error present in the data (*recovery analysis*). Second, we wanted to check how sensitive the BFs are for different samples sizes and different levels of reliability (*sensitivity analysis*). For these purposes, we simulated four types of datasets: one with a null true correlation and the

others with a small, medium, or large positive true correlation according to Cohen's (1988) terminology. We also varied the sample size and added different proportions of noise to the true data (to model variation in reliability), before comparing the inferred correlations and the BFs for $\mathcal{H}_0$ versus $\mathcal{H}_+$ provided by the models.

Model A ignores any reliability issues completely and takes the generated noisy δ and d' values at face value. Models B and C take the uncertainty in the δ and d' values into account when inferring the correlation coefficient. While Model B uses estimates of the error variance calculated from reliability estimates, Model C uses the individual posterior variances of the δ and d' parameters obtained in separate Bayesian data models fitted a priori. Importantly, the correlation parts of Models B and C are very similar: Model B assumes one error variance per variable, whereas Model C assumes one per variable and per participant. Hence, the two models can be tested as one Bayesian correlation model accounting for uncertainty (labeled as Model BC in the following for simplicity). This avoids the need to simulate trial-level data, which would be necessary to obtain parameter estimates of δ and d' from raw frequencies, and simplifies the data generation to pairs of δ and d' values with known population correlation and known reliability (measurement error).

At this point, it is important to foreshadow how any evidence measure (e.g., a BF) of a reasonable method (e.g., Model BC) will behave. According to Morey (2015), such an evidence measure has four desired properties. First, without having observed any data, the evidence measure should favor neither of the competing hypothesis. Second, for a null effect, the evidence measure should be an increasing function of sample size and no other effect size can exceed it; in other words, however much evidence a null observed effect provides for the null hypothesis, no other observed effect size can provide more. Third, for a fixed non-null effect, the evidence measure should become arbitrarily large when the sample size increases. Fourth, the closer the true effect size is to the null effect size, the more the

evidence measure should look like the null. These properties can be tested visually by plotting the evidence obtained in our simulation study against the selected sample sizes in bivariate space.

Data Generation

We generated pairs of δ and d' values for sample sizes $N = \{10, 30, 100, 300, 1000\}$. The pairs were drawn from a bivariate normal distribution with means and variances equal to the means and variances of the δ and d' measures calculated across all participants in Salvador et al.'s (2018) Experiments 1 and 2, given by $\mu = \{.11, 0.29\}$ and $\sigma^2 = \{.05, 0.24\}$, taking them as the true population-level information. The population correlation was varied in four steps, $\rho = \{0, .10, .30, .50\}$. We then added noise to the sampled δ and d' values. The noise came from a normal distribution with $\mu_\epsilon = 0$ and $\sigma^2_\epsilon = \sigma^2/\rho_{xx} - \sigma^2$ where ρ_{xx} is the reliability of the measure that was varied in five steps for δ, $\rho_{xx} = \{.10, .30, .50, .70, .90\}$, and was fixed at .65 for d'. Each parameter combination was repeated 200 times, leading to 20,000 datasets in total.

For simplicity, we used a uniform prior between 0 and 1 to model $\mathcal{H}_+$, that is, a stretched half-beta distribution of width $\kappa = 1$. This model is more vague than scientifically plausible, but allows the demonstration below to be uncluttered. It also presents a worst case scenario in determining the behavior of BFs. Instead of determining their behavior when the true $\mathcal{H}_+$ is sampled from the model of $\mathcal{H}_+$ used for the BF (see Rouder, 2014), the true $\mathcal{H}_+$ is treated as a specific value of ρ. Note that optimal behavior is therefore not always to return evidence for $\mathcal{H}_+$ when $\mathcal{H}_+$ is true. If the $\mathcal{H}_+$ value of ρ is similar to its $\mathcal{H}_0$ value, but the model of $\mathcal{H}_+$ allows much larger values, such a model should be penalized when the standard error of measurement is not small (Morey, 2015).

Results

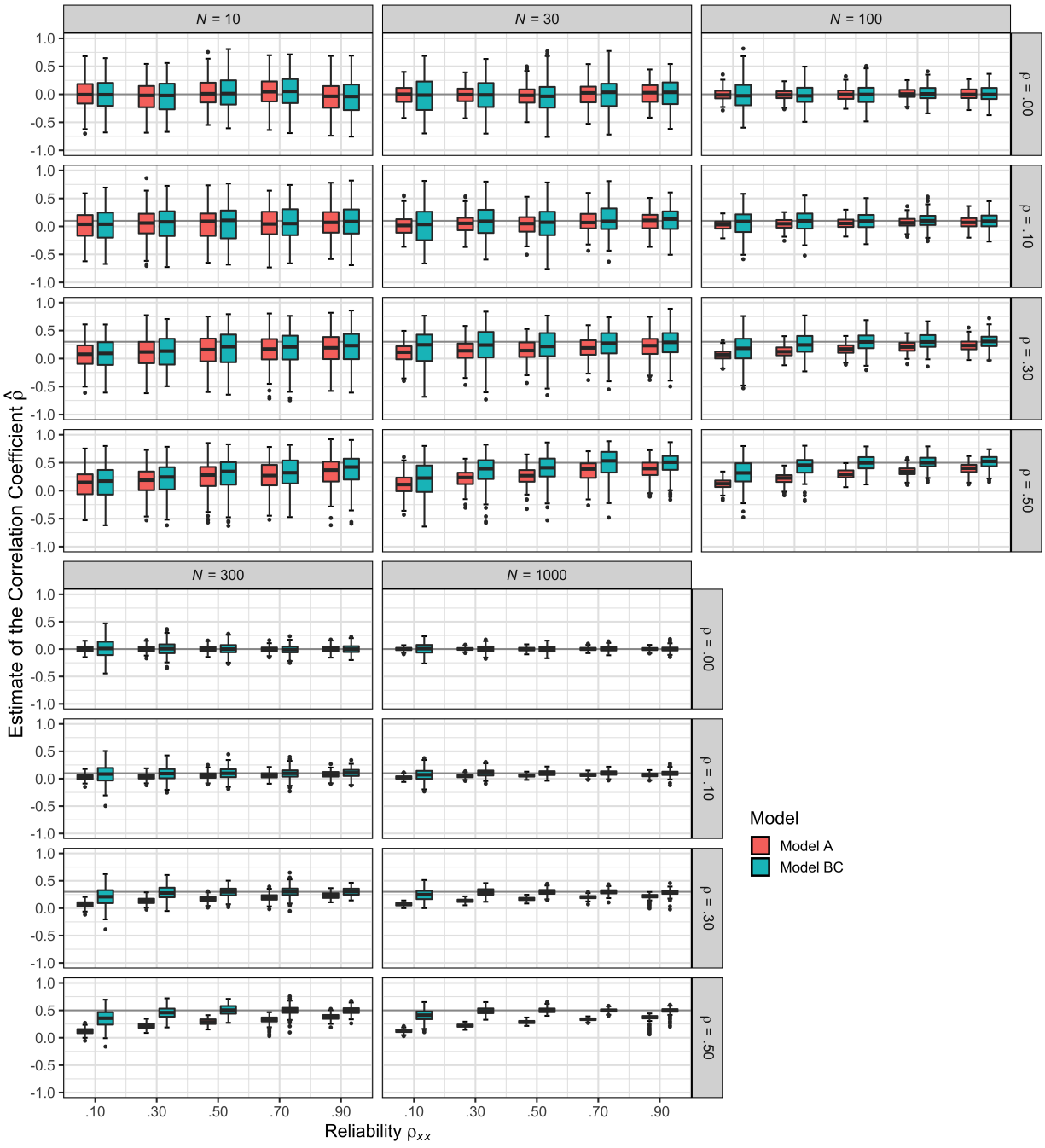
Models A and BC were fitted to each of the 20,000 datasets. First, we were interested in whether the models can recover the true population correlation in the presence of measurement error. Figure 6 shows the results of the parameter recovery analysis. As expected, Model A performed well and recovered the true correlation on average when there was only very little to no measurement error. However, with decreasing reliability, Model A struggled to recover the true parameter, leading to severely attenuated correlation coefficients when the true correlation was medium to large. Model BC performed better in all cases and was able to recover the true correlation more often—even when the sample size was small ($N = 30$). However, both models struggled with a very small sample size ($N = 10$), where they underestimated the true correlation and produced large interquartile ranges.

Second, we tested if the BFs of both models behave sensibly and which sample size is required for diagnostic hypothesis testing in the presence of different levels of reliability. Figure 7 shows the results of the sensitivity analysis by plotting the median of the one-sided BFs for obtaining evidence in favor of a null correlation versus evidence in favor of a positive correlation (B_{0+}) as a function of sample size N and reliability ρ_{xx}. The plot allows us to check which combination of sample size and reliability offers threshold-level BFs (e.g., $B \geq 10$) and provides enough evidence to draw a conclusion (e.g., "strong evidence").

In the case of a null true correlation, Model A led to median BFs between 1.6 and 25.7 in favor of the correct null hypothesis, which grew with sample size. As expected, this result was independent of the reliability of the data. Model BC also showed BFs that increased with sample size, but they were dependent on reliability. With decreasing reliability, the BFs became more conservative, with median BFs between 1.2 and 18.9. For a small true correlation and sample sizes of up to 100, Model A returned median BFs between 1.6 and

Figure 6

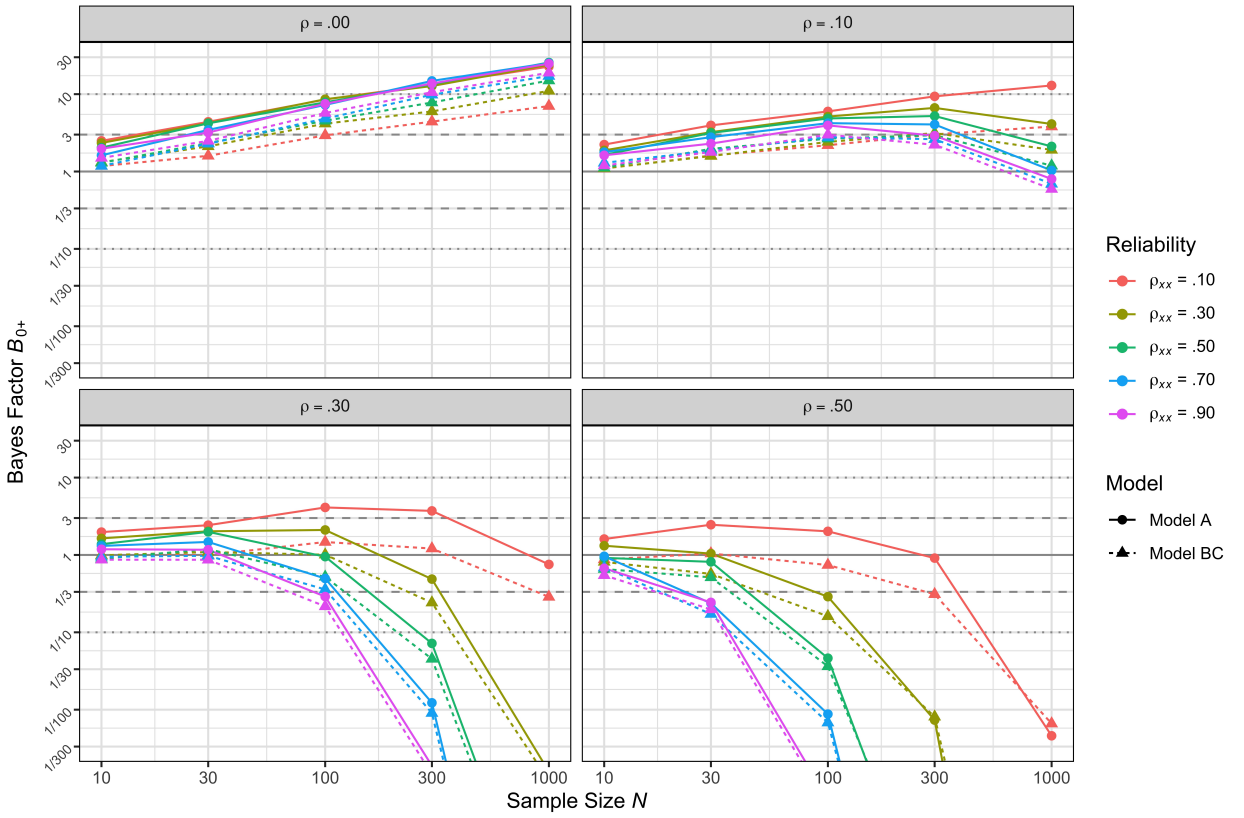
Boxplots of Posterior Mean Estimates of the Correlation Coefficients Across 20,000 Simulated Datasets Recovered by Models A and BC



Note. The boxplots are depicted as a function of sample size N and reliability ρ_{xx} for a null, small, medium, and large true correlation. Gray horizontal lines show the true correlation coefficient. Note that the boxplots of Models A and BC are offset for the given reliability levels to facilitate readability.

Figure 7

Median One-Sided Bayes Factors for Testing the Hypothesis of a Null Correlation Against a Positive Correlation Across 20,000 Simulated Datasets Using Models A and BC



Note. The one-sided BFs are indicated by B_{0+} and depicted as a function of sample size N and reliability ρ_{xx} for a null, small, medium, and large true correlation. Gray horizontal lines show the conventional boundaries for Bayes factors of 1/10, 1/3, 1, 3, and 10 for interpretation as strong evidence in favor of either $\mathcal{H}_+$, moderate evidence in favor of $\mathcal{H}_+$, no evidence in favor of $\mathcal{H}_+$ or $\mathcal{H}_0$, moderate evidence in favor of $\mathcal{H}_0$, and strong evidence in of $\mathcal{H}_0$, respectively. Note that the x-axis and the y-axis are logarithmic, and that the y-axis is clipped for values smaller than 1/300 to facilitate readability.

6.0 in favor of the incorrect null hypothesis, which grew with sample size. Model BC also yielded BFs in favor of the incorrect null hypothesis; but with median values between 1.1 and 3.0, they were lower than those of Model A. For a large sample size ($N = 300$) and all but very low reliability ($\rho_{xx} \geq .30$), the BFs returned by Model BC reached a plateau, before yielding less evidence in favor of the incorrect null hypothesis or even evidence in

favor of the correct alternative hypothesis for a very large sample size ($N = 1,000$) and fairly high reliabilities ($\rho_{xx} \geq .70$). Model A showed the reversal in evidence later (i.e., for larger samples) and with higher BFs in the wrong direction at every parameter combination.

For a medium true correlation, a trade-off between sample size and reliability started to emerge. Using smaller sample sizes ($N \leq 30$), the models mostly returned median BFs between 1/3 and 3. However, as soon as the sample sizes were larger ($N \geq 100$) and the reliability was acceptable ($\rho_{xx} \geq .50$), the models led to BFs in favor of the correct alternative hypothesis—with Model BC showing slightly stronger support. For a large true correlation, relatively small sample sizes ($N \geq 30$) and acceptable reliabilities ($\rho_{xx} \geq .50$) were sufficient: Both models were capable of providing BFs in the expected direction. Moreover, the BFs already became diagnostic with $B_{0+} \leq 1/10$ for a practical sample size ($N = 100$).

Discussion

The results of the simulation study show that Model BC outperforms Model A robustly with regard to posterior estimation and with regard to Bayesian hypothesis testing. While Model A underestimates the true correlation in the presence of measurement error, Model BC provides much better point estimation that is more robust for small sample sizes and in the presence of measurement error. Model BC also consistently provides slightly stronger evidence in favor of the correct alternative hypothesis than Model A, more conservative evidence in favor of the incorrect null hypothesis, and reflects measurement error in the size of the BF. However, both models have their limits for recovering the true correlation when the sample size is small or the reliability is low, and they both show evidence for the incorrect null hypothesis when the true correlation is small.

Importantly, the last observation is not a flaw, but desired behavior. The results of the

sensitivity analysis show that the BFs of both models fulfill all four desired properties of a reasonable evidence measures: The different curves in the bivariate plot of evidence against sample size, shown in Figure 7, behave exactly as recommended by Morey (2015). First, with as little as ten data points, the models consistently return BFs close to 1, and this is almost independent of the reliability (all panels of Figure 7). Second, for a null true correlation, the BFs in favor of the correct null hypothesis increase with N and are the highest BFs in that direction overall (top-left panel of Figure 7). Third, for a medium to large true correlation, the BFs become arbitrarily large when N increases, which means we become more and more certain that the null hypothesis is incorrect (bottom panels of Figure 7). Fourth, the closer the true correlation is to a null correlation, the more the BFs look like the null: For small sample sizes, the curves for a small correlation first approach the curves for a null correlation (cf. top panels of Figure 7); but with increasing sample sizes, the curves for a small correlation start to diverge downward and cross the line of $B = 1$ to provide evidence for the correct alternative hypothesis (top-right panel of Figure 7).

Although Model BC consistently outperforms Model A within the error of the simulation, the fact that small BFs in favor of a null correlation and low reliabilities could still mean that a correlation is present is less than ideal—the situation that characterizes Salvador et al.'s (2018) data. In general, a small true correlation ($\rho = .10$) is very likely to be observed under a point null hypothesis. This is not a consequence of the models or the BF. How quickly the curves in the top-right panel of Figure 7 diverge downward through the bivariate space depends on the data and the available prior information about ρ. When the true correlation is small, a conventional sample size or a non-informative prior that assigns equal a priori evidence to all values between 0 and 1 (i.e., assumes an overly vague model of $\mathcal{H}_+$) will lead to a wide posterior distribution—therefore a small BF. Hence, three straightforward solutions exist to ensure the evidence curve for a small, or any kind of,

correlation will diverge towards support of the correct alternative hypothesis: (1) researchers can collect data from a very large sample, (2) researchers can strive for reliable measures, and (3) researchers can use prior knowledge about the extent of the association between their measures to inform the estimation of the correlation coefficient.

Taken together, the simulation study shows that a Bayesian model that considers the reliability of correlated variables such as Model BC is on average able to recover their true latent correlation—even when the data are quite unreliable (i.e., contain a considerable proportion of measurement error). Importantly, this is also the case when the true correlation is small. Hence, for researchers interested in inferring the correlation coefficient, Model BC is a powerful tool. Although it holds in general that BFs can be non-diagnostic for realistic scenarios of small samples sizes, low reliabilities, and no prior information on the size of the correlation, this is desirable behavior as BFs always reveal the uncertainty that comes from a poorly designed experiment. While the BFs of Models A and BC are sensitive to sample size and prior information, only the BFs of Model BC are also sensitive to reliability. Hence, Model BC should be preferred when researchers are interested in demonstrating a null correlation between measures of performance and awareness in the presence of measurement error. Luckily, all BFs can offer diagnostic evidence when trade-offs between sample size, reliability, and prior information on the extent of the correlation are considered. Inevitably, inferences are stronger when the sample size increases. However, reducing measurement error (for instance by increasing the number of trials) seems to be an equally if not more important consideration as no model can be a substitute for precise measurement. Finally, whenever there is prior information available, it should be formalized and included in the inferential process.

General Discussion

The null-correlation approach of testing the influence of unconscious mental processes on task performance (i.e., interpreting a non-significant correlation between task performance and cue awareness as evidence of an unconscious effect on task performance) assumes that (a) the absence of a significant correlation is evidence for a true null correlation in the population, and (b) the variables of interest can be taken at face value. However, in the NHST framework, the absence of a significant correlation cannot be interpreted as evidence for a true null effect; in contrast, a Bayesian approach allows that kind of inference. Furthermore, the variables of interest are almost always measured with error or are subject to estimation uncertainty. Using the datasets provided by Salvador et al. (2018), which apparently supported a null correlation between recall performance and awareness of a suppress-recall cue, we calculated Bayes factors to test for a null correlation without taking the unreliability of their measures into account (Model A), with taking measurement error into account (Model B), and with taking estimation uncertainty into account (Model C).

As estimates of the variables' reliabilities, odd–even split-half correlations can be used. These were moderate for the awareness measures and very low for the task-performance measure. The latter in itself is an important finding, sufficient to raise considerable doubts over any interpretation of the data, and more than justifies inferring the correlation using one of the proposed models to obtain a BF for testing a null correlation. While the standard model (Model A) provided anecdotal to moderate evidence for a null correlation, the models accounting for uncertainty in the correlated variables (Models B & C) provided only moderate evidence when using an uninformed model of the alternative hypothesis (by assigning a flat prior to parameter ρ), but provided no evidence for a null or for a positive correlation when using an informed model of the alternative hypothesis (by assigning a

default prior to ρ that puts more weight on small positive values). It follows that there is not enough evidence in the data to make a strong claim in favor, or indeed against, a null correlation. In other words, the data are too insensitive to draw a valid conclusion regarding the competing hypotheses (Dienes, 2014). Hence, the question whether unconscious memory suppression exists cannot be answered fully until an extremely large number of participants is tested, or the reliability of the involved measures measure is improved considerably.

We have established that the data quality of Salvador et al.'s (2018) experiments is insufficient to decide between a null and a positive correlation. Besides the rather small sample sizes, the low reliability of task performance is stark and as such raises the question of what causes it. The δ measure is the difference between two binomial processes (recall of the word in the Think condition minus recall in the No-Think condition). Each binomial process alone is already subject to binomial noise (for a similar argument in recognition memory, see Bröder & Malejka, 2017). With as few as six and four trials per condition in Experiments 1 and 2, respectively, a minor response change in one condition—such as a single incorrect response due to a slip of the finger or due to guessing behavior—would inevitably lead to a major change in the δ parameter and thus influence the correlation between δ and d'. Hence, it is important to base any measure on a moderate number of trials per condition; otherwise estimating parameters from limited data will increase uncertainty in those estimates.

One of us (Z.D.) would not use correlation analysis (standardized variables and slopes) in order to assess the evidence for no relation between task performance and awareness provided by Salvador et al. (2018), and would instead use regression analysis (unstandardized variables and coefficients). When working with raw measures to obtain BFs for a hypothesis test, it is more intuitive to select the prior that formalizes the

alternative hypothesis, as the distribution is in the natural metric of the variable of interest. Nonetheless, the current demonstration shows the appeal of BFs. They behave reasonably, they flag data that are non-diagnostic, they allow users to include prior knowledge about ρ, and without them, we would not be aware of the issues at stake. Furthermore, simulation studies can be tailored to the specific paradigm under investigation. Such customized explorations allow researchers to test the effects of different samples sizes, reliabilities, priors, and expected true correlations on BFs, which can greatly aid the interpretation of low evidence. In particular, comparing the impact of different priors can be helpful. Here, robustness regions allow us to assess how robust the evidential conclusion provided by a BF is to changes in the model of the alternative hypothesis. If the prior representing the model of the alternative hypothesis is close to the bounds of the robustness region, the conclusion will change when a different prior is used because the data do not have the resolution to adjust between the different substantive positions formalized by those priors.

Of course, however weak the evidence from Salvador et al.'s (2018) experiments, and however problematic the more general correlation method ignoring unreliability that they and other studies employ, it must be acknowledged that claims about unconscious processes are supported by many other forms of evidence—much of it far more compelling. For instance, studies have been reported that manipulate visibility parametrically in order to demonstrate above-chance performance in the complete absence of awareness (e.g., Biderman et al., 2020; Lin & Murray, 2015), and—even more strikingly—experimental manipulations have been found that affect task performance and stimulus awareness in opposite directions (so-called *double dissociations*; e.g., Biafora & Schmidt, 2020; Schmidt & Vorberg, 2006).

The Reliability Paradox

The challenges of dealing with unreliable data have recently been highlighted by Hedge, Powell, and Sumner (2018). In what they termed the reliability paradox, robust experimental tasks that produce large, replicable effects at the group level tend to yield low reliability because of low variance across individuals. In classical test theory, reliability is defined as the ratio of true-score variance to total observed variance, where the latter is the sum of true-score variance (between-subjects or overall variability due to individual differences) and error variance (within-subjects or trial-to-trial variability due to random error). On the one hand, experimentalists are interested in effects that are easy to replicate, and they consider error variance to be a fixed characteristic of their measurement tool. Hence, they aim to remove the true-score variance, such that the resulting size of the observed effect will be large (De Schryver, Hughes, Rosseel, & De Houwer, 2016). On the other hand, psychometricians are interested in individual differences and need variables with good psychometric properties that can be measured with little to no error. Hence, they only aim at removing random variance, which in turn will lead to small effect sizes (LeBel & Paunonen, 2011).

It follows that robust experimental paradigms that produce large overall effects are not necessarily good paradigms for correlational studies: Although the error variance tends to remain the same with changes in the dispersion of true scores, the proportion of random error will increase relative to the individual variation. This destroys any correlation with other variables due to smaller reliability coefficient, and in turn undermines the true correlation. In other words, when the participants' performance scores are similar (low between-subjects variability), even a small numerical change from the first to a second measurement will necessarily lead to a large reduction in test–retest reliability (high within-subjects variability) as the numerical change is confined within the range of observed values.

How can implicit-cognition researchers break this vicious cycle? To reconcile both recommendations, they must be aware that low reliability can lead to large effects in terms of Cohen's d, but it can also hide a true effect behind measurement error (for discussion of how this relates to power, see Parsons, 2018). For some paradigms, hitting the sweet spot between effect size and individual differences may be a good solution. For example, an increase in sample size will increase the power to detect an effect, even when individual differences are present, as long as the reliability is not too low. For other paradigms, the number of trials can be increased. Adding trials will decrease measurement error (and estimation uncertainty), while leaving between-subjects variability untouched (Rouder & Haaf, 2018).

Hierarchical Data Models

Another note of caution concerns the data models used to estimate the to-be-correlated parameters. We did not implement hierarchical versions of the Bayesian data models for δ and d'. Such models are commonly used when observable behaviors are affected by individual differences. The rationale is that different participants have different parameters, but they are assumed to stem from the same group-level distribution. For example, δ might depend on different individual memory accuracies and d' on different individual sensitivities, such that some participants will naturally show higher or lower performances. A hierarchical data model allows to inform each participant's estimates by other participants' data, which will pull participants with extreme parameters towards the group mean. Generally, the resulting individual parameter estimates will be more certain (narrower posterior distributions). Although accounting for individual differences is generally a good idea, there are two important drawbacks here.

The first problem of hierarchical data models concerns the degree of shrinkage. If

parameter estimates are based on a very small number of trials, as is the case with Salvador et al.'s (2018) data, stable individual differences cannot emerge. Using a hierarchical model then leads to over-shrinkage: The estimates of the data models will occupy only a very small region of the parameter space, the posteriors will be highly influenced by the priors, and the shrunken data that would not be informative in the correlation model. Hence, for a small number of trials, we would advise against using hierarchical data models.

The second problem of hierarchical data models concerns the importance of individual differences. Accounting for individual differences in a data model means reducing the dispersion of true scores, while the error variance remains the same but is inflated relative to individual variation. This, in turn, will lead to lower reliability of the variable and destroys any correlations with other variables—exactly what the reliability paradox states. Hence, some individual differences might be necessary if we cannot increase the measures' reliabilities.

Recently, it has been shown that Bayesian hierarchical models designed to account for variation across trials, variation across individuals, and covariation between individuals and tasks may have their limits for recovering the true correlation as they provide very wide credibility intervals (see Rouder et al., 2019). Discussing how important it is to keep or account for individual differences is beyond the scope of this work. But the usefulness of our approach over a standard correlation model is supported by the results of the simulation studies. The Bayesian correlation models that include measurement or estimation uncertainty (i.e., account for correlation attenuation) recover the true correlation better than the Bayesian correlation model that ignores such types of uncertainty. They also provide inconclusive results ($1/3 < B < 3$) when there are insufficient data points, but diagnostic results ($B \leq 1/3$, $B \geq 3$) when there were enough data points—even when the data are unreliable due to moderate proportions of measurement error. Hence, without

these models, researchers would be ignoring the dramatic effect of error (trial) variance on the observed correlation. We therefore encourage researchers to use such models, but to aim for a moderate number of participants, increase the number of trials, and retain individual differences if the number of trials cannot be increased simultaneously (for an interesting experimental design varying trial-level difficulty, see Siegelman, Bogaerts, & Frost, 2017).

Regression Models

The reader may ask why we focused on the correlation (standardized slope) of the regression plot and not the intercept, in particular because Salvador et al. (2018) report both and because regressing task performance onto an awareness measure and then testing for a positive intercept (i.e., zero awareness but above-chance task performance) circumvents the problem that we cannot prove a null effect in the NHST framework (see Greenwald et al., 1995). The reason is that the intercept approach still suffers from being unable to take account of unreliable measures. If the slope is close to zero and aggregate task performance is significantly greater than chance, it is inevitable that the intercept will be positive (or negative as in Salvador et al.'s Experiment 2). If the slope is zero because of regression attenuation, the regression approach can be adjusted to account for measurement error in the awareness measure (the predictor variable) in the errors-in-variables framework (Klauer, Draine, & Greenwald, 1998; Klauer, Greenwald, & Draine, 1998). However, the effectiveness of the method as well as its underlying assumptions have been criticized (Miller, 2000), questioning its usefulness. Furthermore, Bayesian regression analysis is more heavily affected by parameter priors than correlation analysis, as the variables are scale-dependent, which can severely change the posteriors and thus the resulting BFs.

Conclusion

In attempting to derive conclusions about unconscious influences on behavior from non-significant correlations between performance on explicit and implicit tests, Salvador et al. (2018) and others are in effect rediscovering an analytic technique that was popular some years ago in the 1970s and 1980s. The technique was later heavily challenged on grounds not dissimilar to those described here, and it subsequently fell out of favor (rightly in our view, because the solutions such as the one presented here were not available then). Shortly after the discovery of the distinction between implicit and explicit memory (see Roediger & McDermott, 1993), researchers began to ask whether they are dissociable. Given that in many implicit and explicit memory tests, performance is binary (e.g., a word fragment is either completed or not completed with a study word), tests of *stochastic independence* were undertaken (Tulving & Schacter, 1990). Many of these tests revealed no significant association between performance on implicit and explicit tests, such that many researchers interpreted their results as evidence for functional and neural separation between conscious and unconscious memory processes—just as Salvador et al. have done. However critical appraisal of the technique made it clear that it yielded many false-negative results (i.e., wrongly indicating an absence of association; e.g., Hintzman, 1980; Poldrack, 1996). Indeed Poldrack (1996, p. 437) even pointed out, just as we have done here, the problem of low power leading to incorrect acceptance of the null hypothesis: "The conclusion that performance on two tests is stochastically independent—which is often the theoretically interesting outcome—has usually rested upon the failure to reject the hypothesis that the tests are independent (i.e., acceptance of the null hypothesis). Thus, it is vitally important to establish that tests for dependence have enough power to find dependence when it exists given the number of subjects and items used in the study." Reflecting on the historical pedigree of the technique they employed might have led

Salvador et al. to be rather more cautious about its utility.

On a more positive note, researchers have taken the issue of unreliable data and correlation attenuation into account in the past (e.g., Hofmann, Gawronski, Gschwendner, Le, & Schmitt, 2005; Klauer, Draine, & Greenwald, 1998; Klauer, Greenwald, & Draine, 1998; Parsons, 2018; Siegelman et al., 2017). However, the proposed inferential methods may have lacked applicability in the current context or an off-the-shelf tool that could be easily applied, both of which can hinder widespread adoption.

In summary, the present article shows that alternatives to the NHST null-correlation approach can readily be applied to data such as those collected by Salvador et al. (2018). Importantly, these alternatives need to recognize the near-inevitability of trial error (measurement error and estimation uncertainty). As Fisher (1938, p. 17) famously noted, "To consult the statistician after an experiment is finished is often merely to ask him to conduct a *post mortem* examination." We contend that our post-mortem suggests that the data Salvador et al. collected fail to yield clear evidence for or against unconscious processing, because they were inadequate to test their experimental hypothesis. However, our analysis goes considerably beyond this in providing tools for guiding researchers to ensure that they collect more adequate data in the future. Most importantly, these tools are freely available for use in future research attempting to explore the nature and extent of unconscious mental processes, or indeed in any domain where inferences depend on the magnitude of correlations.

R Scripts

The R scripts of the analyses reported in this article are available publically through the Open Science Framework at `https://osf.io/pq7ug/`.

Data Statement

The authors affirm that they have satisfied the journal's data archiving policy. Two published and archived datasets were reanalyzed, and no new data were collected. Hence, the request to archive raw data is void.

Data Reference

The datasets reanalyzed in this article were taken from the following publication: Salvador, A., Berkovitch, L., Vinckier, F., Cohen, L., Naccache, L., Dehaene, S., & Gaillard, R. (2018). Unconscious memory suppression. *Cognition, 180*, 191–199. `https://doi.org/10.1016/j.cognition.2018.06.023`. The supplementary materials provided under the URL include the aggregate data of the cue-awareness measure d' and the trial-level data of the task-performance measure δ. The trial-level data of the d' measure were obtained directly from Salvador et al.

Acknowledgments

The authors would like to thank Alexandre Salvador and Raphaël Gaillard for sharing the trial-level data of their experiments.

Funding Agency

This research was supported by a grant to D.S. (PI), Z.D., and M.V. from the Economic and Social Research Council (ESRC; grant number ES/P009522/1), a grant to M.V. from the Programa de Atracción de Talent Investigador, Comunidad de Madrid (grant number 2016-T1/SOC-1395), and a grant to M.V. from the Agencia Estatal de Investigación and the Fondo Europeo de Desarrollo Regional (AEI, FEDER; grant number PSI2017-85159-P).

References

Anderson, M. C., & Green, C. (2011). Suppressing unwanted memories by executive control. *Nature, 410*(6826), 366–369. https://doi.org/10.1038/35066572

Barnard, J., McCulloch, R., & Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviation and correlations, with applications to shrinkage. *Statistica Sinica, 10*(4), 1281–1311.

Batterink, L. J., Reber, P. J., Neville, H. J., & Paller, K. A. (2015). Implicit and explicit contributions to statistical learning. *Journal of Memory & Language, 83*, 62–78. https://doi.org/10.1016/j.jml.2015.04.004

Behseta, S., Berdyyeva, T., Olson, C. R., & Kass, R. E. (2009). Bayesian correction for attenuation of correlation in multi-trial spike count data. *Journal of Neurophysiology, 101*(4), 2186–2193. https://doi.org/10.1152/jn.90727.2008

Berkovitch, L., & Dehaene, S. (2019). Subliminal syntactic priming. *Cognitive Psychology, 109*, 26–46. https://doi.org/10.1016/j.cogpsych.2018.12.001

Biafora, M., & Schmidt, T. (2020). Induced dissociations: Opposite time courses of priming and masking induced by custom-made mask-contrast functions. *Attention, Perception, & Psychophysics, 82*(3), 1333–1354. https://doi.org/10.3758/s13414-019-01822-4

Biderman, D., Shir, Y., & Mudrik, L. (2020). B or 13? Unconscious top-down contextual effects at the categorical but not the lexical level. *Psychological Science, 31*(6), 663–677. https://doi.org/10.1177/0956797620915887

Bröder, A., & Malejka, S. (2017). On a problematic procedure to manipulate response bias in recognition experiments: The case of "implied" base rates. *Memory, 25*(6), 736–743. https://doi.org/10.1080/09658211.2016.1214735

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics, 7*(4), 434–455.

https://doi.org/10.1080/10618600.1998.10474787

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3*(3), 296–322. https://doi.org/10.1111/j.2044-8295.1910.tb00207.x

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: A modern perspective*. New York, NY: Taylor & Francis.

Charles, E. P. (2005). The correction for attenuation due to measurement error: Clarifying concepts and creating confidence sets. *Psychological Methods, 10*(2), 206–226. https://doi.org/10.1037/1082-989x.10.2.206

Chiu, Y.-C., & Aron, A. R. (2014). Unconsciously triggered response inhibition requires an executive setting. *Journal of Experimental Psychology: General, 143*(1), 56–61. https://doi.org/10.1037/a0031497

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

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*(6), 1996–2009. https://doi.org/10.3758/s13423-016-1063-0

DeCarlo, L. T. (1998). Signal detection theory and generalized linear models. *Psychological Methods, 3*(2), 186–205. https://doi.org/10.1037/1082-989x.3.2.186

Dehaene, S., Naccache, L., Le Clec'H, G., Koechlin, E., Mueller, M., Dehaene-Lambertz, G., ... Le Bihan, D. (1998). Imaging unconscious semantic priming. *Nature, 395*(6702), 597–600. https://doi.org/10.1038/26967

De Schryver, M., Hughes, S., Rosseel, Y., & De Houwer, J. (2016). Unreliable yet still replicable: A comment on LeBel and Paunonen (2011). *Frontiers in Psychology, 6*, 2039. https://doi.org/10.3389/fpsyg.2015.02039

Dickey, J. M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics, 42*(1), 204–223. https://doi.org/

10.1214/aoms/1177693507

Dickey, J. M., & Lientz, B. P. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *The Annals of Mathematical Statistics, 41*(1), 214–226. https://doi.org/10.1214/aoms/1177696919

Dickinson, A., & Brown, K. J. (2007). Flavor-evaluative conditioning is unaffected by the contingency knowledge during training with color–flavor compounds. *Learning & Behavior, 35*(1), 36–42. https://doi.org/10.3758/bf03196072

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. (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., & McLatchie, N. (2018). Four reasons to prefer Bayesian analyses over significance testing. *Psychonomic Bulletin & Review, 25*(1), 207–218. https://doi.org/10.3758/s13423-017-1266-z

Dzhafarov, E. N., & Kujala, J. (2017). Probability, random variables, and selectivity. In W. H. Batchelder, H. Colonius, E. N. Dzhafarov, & J. Myung (Eds.), *New handbook of mathematical psychology: Foundations and methodology.* (Vol. 1, pp. 85–150). Cambridge, UK: Cambridge University Press.

Eriksen, C. W. (1960). Discrimination and learning without awareness: A methodological survey and evaluation. *Psychological Review, 67*(5), 279–300. https://doi.org/10.1037/h0041622

Etz, A., Haaf, J. M., Rouder, J. N., & Vandekerckhove, J. (2018). Bayesian inference and testing any hypothesis you can specify. *Advances in Methods and Practices in Psychological Science, 1*(2), 281–295. https://doi.org/10.1177/2515245918773087

Finkbeiner, M. (2011). Subliminal priming with nearly percept performance in the prime-classification task. *Attention, Perception, & Psychophysics, 73*(4), 1255–1265.

http://doi.org/10.3758/s13414-011-0088-8

Fisher, R. A. (1938). Presidential address to the First Indian Statistical Congress. *Sankhya, 4*, 14–17.

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review, 116*(2), 439–453. https://doi.org/10.1037/a0015251

Gelman, A., Carlin, J. B., Stern, H. A., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis*. Boca Raton, FL: CRC Press.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

Geyer, T., Shi, Z., & Müller, H. J. (2010). Contextual cueing in multiconjunction visual search is dependent on color- and configuration-based intertrial contingencies. *Journal of Experimental Psychology: Human Perception and Performance, 36*(3), 515–532. https://doi.org/10.1037/a0017448

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York, NY: Wiley.

Greenwald, A. G., Klinger, M. R., & Schuh, E. S. (1995). Activation by marginally perceptible ("subliminal") stimuli: Dissociation of unconscious from conscious cognition. *Journal of Experimental Psychology: General, 124*(1), 22–42. https://doi.org/10.1037/0096-3445.124.1.22

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practices, 12*(3), 38–47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

Hassin, R. R. (2013). Yes it can: On the functional abilities of the human unconscious. *Perspectives on Psychological Science, 8*(2), 195–207. https://doi.org/10.1177/1745691612460684

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods, 50*(3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Hedger, N., Garner, M., & Adams, W. J. (2019). Do emotional faces capture attention, and does this depend on awareness? Evidence form the visual probe paradigm. *Journal of Experimental Psychology: Human Perception and Performance, 45*(6), 790–802. https://doi.org/10.1037/xhp0000640

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*(9), 934–968. https://doi.org/10.1037/bul0000054

Hintzman, D. (1980). Simpson's paradox and the analysis of memory retrieval. *Psychological Review, 87*(4), 398–410. https://doi.org/10.1037/0033-295X.87.4.398

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the implicit association test and explicit self-report measures. *Personality and Social Psychology Bulletin, 31*(10), 1369–1385. https://doi.org/10.1177/0146167205275613

Holender, D. (1986). Semantic activation without conscious identification in dichotic listening, parafoveal vision, and visual masking: A survey and appraisal. *Behavioral and Brain Sciences, 9*(1), 1–23. https://doi.org/10.1017/s0140525x00021269

JASP Team (2020). *JASP Version 0.9.2* [Computer software]. Retrieved from: https://jasp-stats.org/

Jensen, K., Kirsch, I., Odmalm, S., Kaptchuk, T.J., & Ingvar, M. (2015). Classical conditioning of analgesic and hyperalgesic pain responses without conscious awareness. *Proceedings of the National Academy of Science, 112*(25), 7863–7867. https://doi.org/10.1073/pnas.1504567112

Jeffreys, H. (1961). *Theory of probability.* Oxford, UK: Oxford University Press.

Kalra, P. B., Gabrieli, J. D. E., & Finn, A. S. (2019). Evidence of stable individual differences in implicit learning. *Cognition, 190*, 199–211. https://doi.org/10.1016/j.cognition.2019.05.007

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*(430), 773–795. https://doi.org/10.2307/2291091

Klauer, K. C., Draine, S. C., & Greenwald, A. G. (1998). An unbiased errors-in-variables approach to detect unconscious cognition. *British Journal of Mathematical and Statistical Psychology, 51*(2), 253–267. https://doi.org/10.1111/j.2044-8317.1998.tb00680.x

Klauer, K. C., Greenwald, A. G., & Draine, S. C. (1998). Correcting for measurement error in detecting unconscious cognition: Comment on Draine and Greenwald (1998). *Journal of Experimental Psychology: General, 127*(3), 318–319. https://doi.org/10.1037/0096-3445.127.3.318

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science, 6*(3), 299–312. https://doi.org/10.1177/1745691611406925

Lakens, D. (2017). Equivalence tests: A practical primer for t-tests, correlations, and meta-analyses. *Social Psychological and Personality Science, 8*(4), 355–362. https://doi.org/10.1177/1948550617697177

Lakens, D., McLatchie, N., Isager, P. M., Scheel, A. M., & Dienes, Z. (2018). Improving inferences about null effects with Bayes factors and equivalence tests. *Journals of Gerontology: Series B, 75*(1), 45–57. https://doi.org/10.1093/geronb/gby065

LeBel, E. P., & Paunonen, S. V. (2011). Sexy but often unreliable: The impact of unreliability on the replicability of experimental findings with implicit measures. *Personality and Social Psychology Bulletin, 37*(4), 570–583. https://doi.org/10.1177/0146167211400619

LeDoux, J. E., Michel, M., & Lau, H. (2020). A little history goes a long way toward understanding why we study consciousness the way we do today. *Proceedings of the National Academy of Science, 117*(13), 6976–6984. https://doi.org/10.1073/pnas.1921623117

Lee, M. D., & Vanpaemel, W. (2018). Determining informative priors for cognitive models. *Psychonomic Bulletin & Review, 25*(1), 114–127. https://doi.org/10.3758/s13423-017-1238-3

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course.* Cambridge, UK: Cambridge University Press.

Lin, Z., & Murray, S. O. (2015). Automaticity of unconscious response inhibition: Comment on Chiu and Aron (2014). *Journal of Experimental Psychology: General, 144*(1), 244–254. https://doi.org/10.1037/xge0000042

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* New York, NY: Erlbaum.

Matzke, D., Ly, A., Selker, R., Weeda, W. D., Scheibehenne, B., Lee, M. D., & Wagenmakers, E.-J. (2017). Bayesian inference for correlations in the presence of measurement error and estimation uncertainty. *Collabra: Psychology, 3*(1), 25. https://doi.org/10.1525/collabra.78

McLatchie, N. (2018, January 12). Bayes: Robustness regions [Blog post]. Retrieved from: https://www.neilmclatchie.com/bayes-robustness-regions/

Miller, J. (2000). Measurement error in subliminal perception experiments: Simulation analyses of two regression methods. *Journal of Experimental Psychology: Human Perception and Performance, 26*(4), 1461–1477. https://doi.org/10.1037//0096-1523.26.4.1461

Morey, R. D. (2015, April 10). All about that "bias, bias, bias" (it's no trouble) [Blog post]. Retrieved from http://bayesfactor.blogspot.com/2015/04/all-about-that-bias-bias-bias-its-no.html

Morey, R. D., & Wagenmakers, E.-J. (2014). Simple relation between Bayesian order-restricted and point-null hypothesis tests. *Statistics and Probability Letters, 92*, 121–124. https://doi.org/10.1016/j.spl.2014.05.010

Muchinsky, P. M. (1996). The correction for attenuation. *Educational and Psychological Measurement, 56*(1), 63–75. http://doi.org/10.1177/0013164496056001004

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology, 3*(1), 1–18. https://doi.org/10.1016/0022-2496(66)90002-2

Nunnally, J. C. (1970). *Introduction to psychological measurement.* New York, NY: McGraw-Hill.

Paciorek, A., & Williams, J. N. (2015). Semantic generalization in implicit language learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 41*(4), 989–1002. https://doi.org/10.1037/xlm0000100

Parsons, S. (2018, May 24). *Visualising two approaches to explore reliability–power relationships.* PsyArXiv. https://psyarxiv.com/qh5mf/

Pearson, K. (1895). Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London, 58*(347–352), 240–242. http://doi.org/10.1098/rspl.1895.1895.0041

Plummer, M. (2017). *JAGS Version 4.3.0* [Computer software manual]. Retrieved from: http://mcmc-jags.sourceforge.net

Poldrack, R. A. (1996). On testing for stochastic dissociations. *Psychonomic Bulletin & Review, 3*(4), 434–448. https://doi.org/10.3758/bf03214547

R Core Team. (2018). *R: A language and environment for statistical computing* [Computer software manual]. Vienna, Austria. Retrieved from: https://www.R-project.org/

Roediger, H. L., III, & McDermott, K. B. (1993). Implicit memory in normal human subjects. In F. Boller & J. Grafman (Eds.), *Handbook of neuropsychology* (Vol. 8, pp. 63–131). Amsterdam: Elsevier.

Rouder, J. N. (2014). Optimal stopping: No problem for Bayesians. *Psychonomic Bulletin & Review, 21*(2), 301–308. https://doi.org/10.3758/s13423-014-0595-4

Rouder, J. N., & Haaf, J. M. (2018). Power, dominance, and constraint: A note on the appeal of different design traditions. *Advances in Methods and Practices in Psychological Science, 1*(1), 19–26. https://doi.org/10.1177/2515245917745058

Rouder, J. N., Kumar, A., & Haaf, J. M. (2019, April 4). *Why most studies of individual differences with inhibition tasks are bound to fail.* PsyArXiv. https://psyarxiv.com/3cjr5/

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*(4), 597–605. https://doi.org/10.3758/bf03196808

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*(2), 225–237. https://doi.org/10.3758/pbr.16.2.225

Salvador, A., Berkovitch, L., Vinckier, F., Cohen, L., Naccache, L., Dehaene, S., & Gaillard, R. (2018). Unconscious memory suppression. *Cognition, 180*, 191–199. https://doi.org/10.1016/j.cognition.2018.06.023

Sanchez, D. J., Gobel, E. W., & Reber P. J. (2010). Performing the unexplainable: Implicit task performance reveals individually reliable sequence learning without explicit knowledge. *Psychonomic Bulletin & Review, 17*(6), 790–796. https://doi.org/10.3758/PBR.17.6.790

Schmidt, F. L., & Hunter, J. E. (1999). Theory testing and measurement error. *Intelligence, 27*(3), 183–198. https://doi.org/10.1016/s0160-2896(99)00024-0

Schmidt, F. L., & Hunter, J. E. (2014). *Methods of meta-analysis: Correcting error and bias in research findings.* Newbury Park, CA: Sage.

Schmidt, T., & Vorberg, D. (2006). Criteria for unconscious cognition: Three types of dissociation. *Perception & Psychophysics, 68*(3), 489–504. https://doi.org/10.3758/bf03193692

Schweikert, R., Fisher, D. L., & Sung, K. (2012). Discovering cognitive architecture by selectively influencing mental processes. In *Advanced series on mathematical psychology* (Vol. 4). Singapore: World Scientific.

Shanks, D. R. (2017). Regressive research: The pitfalls of post hoc data selection in the study of unconscious mental processes. *Psychonomic Bulletin & Review, 24*(3), 752–775. https://doi.org/10.3758/s13423-016-1170-y

Shanks, D. R., & St. John, M. F. (1994). Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences, 17*(3), 367–395. https://doi.org/10.1017/s0140525x00035032

Siegelman, N., Bogaerts, L., & Frost, R. (2017). Measuring individual differences in statistical learning: Current pitfalls and possible solutions. *Behavior Research, 49*(2), 418–432. https://doi.org/10.3758/s13428-016-0719-z

Sklar, A. Y., Levy, N., Goldstein, A., Mandel, R., Maril, A., & Hassin, R. R. (2012). Reading and doing arithmetic nonconsciously. *Proceedings of the National Academy of Sciences, 109*(48), 19614–19619. https://doi.org/10.1073/pnas.1211645109

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology, 15*(1), 72–101. https://doi.org/10.2307/1412159

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*(3), 271–295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x

Stan Development Team. (2018). *Stan Version 2.18.0* [Computer software manual]. Retrieved from: http://mc-stan.org/

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science, 11*(5), 702–712. http://doi.org/10.1177/1745691616658637

Strenze, T. (2007). Intelligence and socioeconomic success: A meta-analytic review of longitudinal

research. *Intelligence, 35*(5), 401–426. https://doi.org/10.1016/j.intell.2006.09.004

Tokuda, T., Goodrich, B., Van Mechelen, I., Gelman, A., & Tuerlinckx, F. (2020). *Visualizing distributions of covariance matrices*. Unpublished manuscript. Retrieved from: http://www.stat.columbia.edu/∼gelman/research/unpublished/Visualization.pdf

Tulving, E., & Schacter, D. L. (1990). Priming and human memory systems. *Science, 247*(4940), 301–306. https://doi.org/10.1126/science.2296719

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

Vadillo, M. A., Linssen, D., Orgaz, C., Parsons, S., & Shanks, D. R. (2020). Unconscious or underpowered? Probabilistic cuing of visual attention. *Journal of Experimental Psychology: General, 149*(1), 160–181. https://doi.org/10.1037/xge0000632

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review, 14*(5), 779–804. https://doi.org/10.3758/bf03194105

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., ... Morey, R. D. (2017). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review, 25*(1), 58–76. https://doi.org/10.3758/s13423-017-1323-7

Appendix

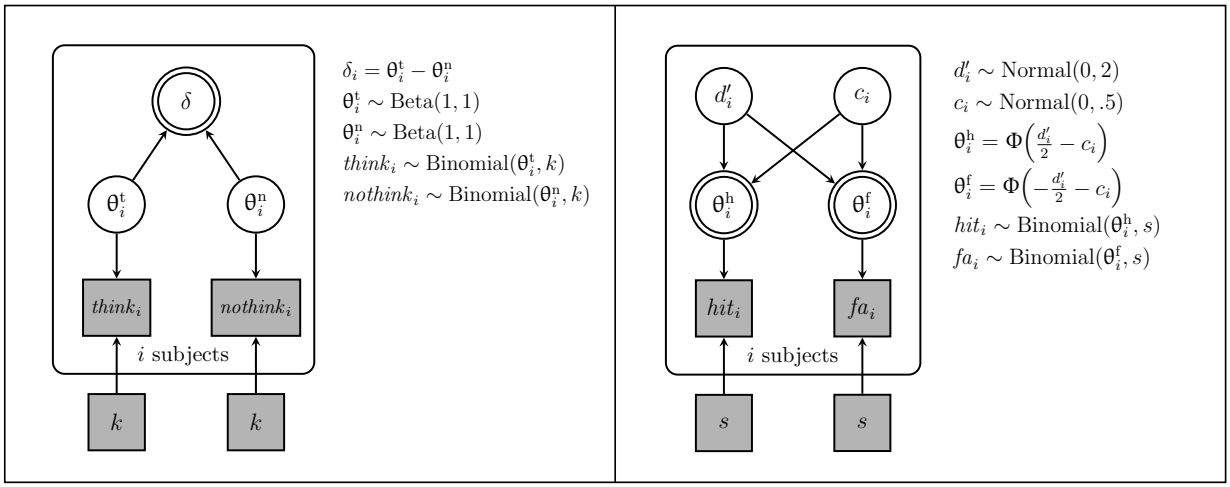
Bayesian Data Models to Obtain the Input Data for Model C

The difference δ_i in memory performance between the Think and the No-Think condition for each individual i is modeled as the difference between two proportions (see left panel of Figure A; Lee & Wagenmakers, 2013, Chapter 3.2). The proportions are based on latent variable θ^{t} that produces a certain number of successes out of k Think trials and latent variable θ^{n} that produces a certain number of successes out of k No-Think trials. Note that in both of Salvador et al.'s (2018) experiments, each individual i had to complete the same number of Think and No-Think trials. The observed numbers of successes *think* and *nothink* are modeled as draws from binomial distributions with θ^{t}, θ^{n}, and k as parameters. The priors for the success rates θ^{t} and θ^{n} are beta distributions with shape parameters of 1 that correspond to uniform distributions between 0 and 1, such that all proportions are equally likely a priori. To determine the posterior distribution of δ, parameters θ^{t} and θ^{n} are subtracted for each individual.

The estimates of d' are based on a signal-detection model that infers the discriminability measure d'_i and a response-bias measure c_i for each participant i from the frequencies of hits and false alarms of that participant (see right panel of Figure A). The hit rate and false-alarm rate are calculated through functions of the standard SDT formulae using normal distributions (Green & Swets, 1966), or more precisely, through inverse probit link functions that give the cumulative distribution function of the standard normal (DeCarlo, 1998). The observed frequencies of hits and false alarms (*hit* and *fa*, respectively) are draws from binomial distributions with parameters given by hit rate θ^{h} and the total number of diamond [square] trials s, and with false-alarm rate θ^{f} and the total number of square [diamond] trials s. Again, each individual i had to complete the

Figure A

Graphical Depictions of the Bayesian Models for Task Performance and Cue Awareness



Note. Left panel: Task-performance measure δ used in Model C. Right panel: Cue-awareness measure d' used in Model C. Rectangular nodes represent discrete (as opposed to continuous) variables, and double-bordered nodes represent deterministic (as opposed to stochastic) variables. Nodes with double outlines are functions of their parent nodes.

same number of trials showing a diamond versus a square. The priors for d' and c are normal distributions, which correspond to uniform distributions over the hit and false-alarm rates through the probit link (for details, see Lee & Wagenmakers, 2013, Chapter 11.1). Note that the response criterion c is a nuisance parameter here and not required for the analysis of the correlation between δ and d', but must be included in the model to fully describe behavior in the awareness task.

As this is a tutorial on how to use Bayesian models that account for correlation attenuation, we did not implement a full hierarchical Bayesian version of Model C where participant-level data and their correlation are estimated simultaneously. Instead, we followed Matzke et al. (2017) and implemented Model C in two steps: One has to run the data models for the variables first, before the correlation model can be applied. This modular approach has the advantage that, if the user wants to use Model C in a paradigm

different to Salvador et al.'s (2018) memory-suppression one, only the data models of the implicit and explicit tasks will need to be adapted, while all but the priors in the correlation model can be left untouched. This task might be easier for a novice to Bayesian modeling than having to change the entire model when the estimations of δ or d' were included in the correlation model. Moreover, it allows implementing combinations of the models, for example, when only one variable suffers from measurement error and the other from estimation uncertainty. For more a detailed comparison of the modular and the simultaneous approach, we refer to Matzke et al.

Example of How to Select the Model of the Alternative Hypothesis

Because Salvador et al. (2018) asked a new research question, no prior information on the size of ρ is readily available. Fortunately, Salvador et al. implemented a "conscious" condition alongside the "unconscious" condition in the Think/No-Think task. In this condition, participants were asked to recall a word's associate or suppress its recall when being presented with the unmasked geometric shapes. Hence, if unconscious memory suppression does not exist as assumed under the alternative hypothesis, a useful benchmark for the correlation between unconscious memory suppression and cue awareness is the correlation between conscious memory suppression and cue awareness. For Experiments 1 and 2 analyzed jointly, the Pearson correlation between the δ values from the "conscious" condition and the d' values was .00 with a 95% CI of $[-.23, .22]$, while the disattenuated correlation was $-.01$ with a 95% CI of $[-.59, .57]$.

The upper bound of the disattenuated correlation coefficient for the "conscious" data shows that it is safe to conclude that the correlation coefficient for the "unconscious" data cannot be higher than .57 (and is possibly much smaller than that). If correlation coefficients above .57 are implausible according to the theory of conscious memory

suppression, they should not receive any weight when testing for unconscious memory suppression. Hence, a stretched half-beta distribution with beta width $\kappa = 1/8$ is a reasonable prior to inform ρ when constructing a BF to compare $\mathcal{H}_0$ against $\mathcal{H}_+$, leading to a model of the alternative hypothesis of $\mathcal{H}_+ \colon \rho \sim \text{StretchBeta}_{\text{T}(0,)}(8, 8)$. This prior distribution has a mode of 0, assigns more weight to values close to the critical null correlation, and does not assign much weight to values over .60 (cf. Figure 4 in the main text).

Details of the Model-Fitting Routines

For each dataset and each model, three MCMC chains were started, each of which supplied 5,000 samples from the posterior distribution after the results were thinned by taking every third sample. Most MCMC chains are strongly autocorrelated (i.e., successive steps are taken near each other and are not independent). Thinning ensures a reduction in autocorrelations between two successive retained samples, such that the effective sample size N_{eff} can come close to the number of samples retained in the output files. The Gelman–Rubin convergence statistic $\hat{R}$ (Gelman & Rubin, 1992) was smaller than 1.05 for all critical parameters, indicating that all chains had converged (Brooks & Gelman, 1998). Furthermore, the reported Bayes factors were averaged over five batches with standard errors between 0 and 0.07.