Journal of Memory and Language 143 (2025) 104621

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Measurement and sampling noise undermine inferences about awareness in location probability learning: A modeling approach☆

Alicia Franco-Martínez [a], Francisco Vicente-Conesa [a,b], David R. Shanks [c], Miguel A. Vadillo [a,*]

[a] *Autonomous University of Madrid, Spain*
[b] *Universidad Internacional de la Empresa, Spain*
[c] *University College London, United Kingdom*

ARTICLE INFO

Keywords:
Unconscious processes
Awareness measures
Location probability learning
Measurement noise
Sampling noise
Cognitive modeling
Memory-guided visual search

ABSTRACT

Occasionally, experimental psychologists enter into the realm of psychometrics without being fully aware of the risks involved in the study of individual differences. Here we re-assess the many studies on the interaction between memory and attention in location probability learning that suggest that people can unconsciously learn to suppress salient but irrelevant distractors frequently presented in a certain location. In the additional singleton task, one of the arguments to support this claim is that suppression in memory-guided visual search does not significantly differ between "aware" and "unaware" participants. Although rarely acknowledged, this null interaction could also result if the data are contaminated by measurement and/or sampling noise. Unfortunately, the reliability of the awareness measure cannot be assessed with standard methods, since it is a single-trial test. In the present study we offer model-based estimations of measurement and sampling noise in empirical data. Our goal is to determine how often researchers would mistakenly conclude that learning is unconscious, given data from a model based on the opposite claim (i.e., that learning is conscious) but including noise in participants' search times and awareness responses. To do so, we fitted this noisy conscious model to a dataset involving 159 participants who performed the additional singleton task. Estimated parameters from this model were used, first, to predict the observed pattern of results and, second, to simulate new responses and participants. Results suggest that, under reasonable measurement noise and sample sizes, simulated evidence from the model can paradoxically but falsely support arguments used to defend the unconscious learning hypothesis. This study serves as an illustration to experimental psychologists – particularly those investigating memory and learning – of the risks of neglecting basic psychometric requirements in individual differences research.

Introduction

Psychometricians and experimental psychologists inhabit radically different worlds (Borsboom et al., 2009; Cronbach, 1957). The former explore domains where performance varies substantially across participants, looking for ever better ways to reliably detect those individual differences. The latter, in contrast, crave robust effects that, ideally, will be observed in every single person, often with little or no variation from one participant to the next. These two species rarely venture into each other's habitat and in the exceptional cases when this happens the outcome is not always a happy one.

Although concerns about the suitability of experimental tasks for studying individual differences have been long recognized in fields such as intelligence and aging research, other areas – such as inhibition or implicit learning and memory (Rouder & Haaf, 2019; Rouder et al., 2023b) – are only now beginning to address these challenges. Unfortunately, although these tasks are known to yield large effects at the group level, they tend to be less sensitive to individual differences (Hedge et al., 2018). Consequently, attempts to use these tasks to study individual differences have sometimes resulted in contradictory findings and failed replications (e.g., Karr et al., 2018; Paap & Greenberg, 2013; Rey-Mermet et al., 2019; Ross et al., 2015; Von Bastian et al., 2020). One potential reason for this is that, when experimenters enter into the realm of psychometrics, they commit to certain assumptions without evaluating whether their data validate them. A clear example consists in submitting variables to analyses that implicitly assume they have been

☆ This article is part of a special issue entitled: 'Individual differences in memory' published in Journal of Memory and Language.

* Corresponding author at: Departamento de Psicología Básica, Facultad de Psicología, Universidad Autónoma de Madrid, 28049 Madrid, Spain.
E-mail address: miguel.vadillo@uam.es (M.A. Vadillo).

https://doi.org/10.1016/j.jml.2025.104621
Received 5 March 2024; Received in revised form 31 January 2025; Accepted 3 February 2025
Available online 14 March 2025

measured with little or no measurement noise (Vadillo et al., 2020, 2022). On the occasions when this assumption has been evaluated, the psychometric properties of the measures provided by these tasks, although sometimes acceptable (e.g., Viviani et al., 2024), often leave much to be desired (e.g., Draheim et al., 2019; Enkavi et al., 2019; Garre-Frutos et al., 2024; Hedge et al., 2018; Hernández-Gutiérrez et al., 2025; Paap & Sawi, 2016; Rothkirch et al., 2022; Vadillo et al., 2022; Yaron et al., 2024).

Measurement noise, also known as measurement error, refers to the random variability, inherent in the measurement process, which cannot be attributed to the effect intended to be measured. Note that this noise is the opposite of reliability and is not intrinsic to the task; rather, it is also a function of the sample and the administration conditions under which the measure was obtained. Other things being equal, the larger the trial sample (i.e., number of trials), the smaller the impact of this variability on the overall measurement, due to the averaging out of individual trial noise, thereby reducing the measurement noise in the data. However, assuming that having a large amount of trials will strictly lead to adequate reliability might be risky, so it is always advisable to empirically estimate the measurement noise with methods such as split-half coefficients or signal-to-noise ratios (Rouder et al., 2023a).

The study of unconscious learning and memory processes provides many examples of strong inferences drawn with insufficient consideration of psychometric requirements. A common argument to support the claim that learning and memory can operate unconsciously is to report evidence that an improvement in task performance is uncorrelated with participants' awareness of the regularities driving performance (e.g., Colagiuri & Livesey, 2016; Jiang et al., 2018; Salvador et al., 2018; Soto et al., 2011). For instance, Soto et al. found a null correlation between performance in a working memory task and awareness of the cues' visibility, which served as an argument for the inference that working memory can operate with unconscious representations.

The intuition behind these analyses is that if participants' awareness and learning are uncorrelated, they must be based on different processes or representations. Learning must be based on something that leaves no trace on awareness measures. However, an alternative and far simpler explanation for the lack of correlation is that either or both measures are contaminated by substantial amounts of measurement noise, which attenuates the observed correlation between them (Hunter & Schmidt, 2015; Spearman, 1904). If this is indeed the case they will not correlate with each other, even if at the latent level they tap onto the same cognitive processes.

One cannot discriminate between these two hypotheses (i.e., learning is unconscious vs. is conscious but there is excessive measurement noise) without having some estimation of the internal consistency of the dependent variables involved in the analysis. Sadly, reliabilities are only seldom reported in these experiments (Vadillo et al., 2022) and, in the rare occasions when they are estimated, reliabilities often turn out to be sufficiently low to cast doubt on the appropriateness of the analyses (e.g., Anderson & Kim, 2019; Arnon, 2020; Bogaerts et al., 2018; Erickson et al., 2016; Kalra et al., 2019; Kaufman et al., 2010; Siegelman & Frost, 2015; Smyth & Shanks, 2008; Vadillo et al., 2020, 2022, 2024; West et al., 2018; Yaron et al., 2024).

To make things worse, several tasks used in the study of unconscious learning and memory apply this logic to single-trial measures of awareness (such as probability cuing, e.g., Jiang et al., 2015; Vadillo et al., 2020; attentional capture, e.g., Adams & Gaspelin, 2020; and distractor suppression in the additional singleton task, Wang & Theeuwes, 2018a, 2018b, 2018c), commonly taken at the end of the experiment. Although these designs typically involve large trial samples to ensure low measurement noise for search times, no comparable level of confidence is sought for the awareness measure. And this is further exacerbated because the internal consistency of a single measurement is, by definition, undetermined, so reliability cannot be estimated with conventional methods (such as split-half reliability).

On another level, which is more extensively discussed in the literature, is the concept of sampling noise. Similar to measurement noise, sampling noise refers to the random variability, inherent in the participant sampling process, which cannot be attributed to the population effect intended to be estimated. This is why sample size is one of the main considerations in designing experiments: the larger the sample size, the smaller the sampling noise in the estimated parameter. Despite its importance, there is still no well-established routine to adequately justify sample sizes (Lakens, 2022), leading to samples that may not be sufficiently powered to detect the desired effects – in this case, an interaction between awareness and learning. A common argument to justify testing a specific number of participants is that the sample size is similar to previous ones in the same literature, but this does not guarantee that those sample sizes were sufficiently informative. For instance, in the domain of location probability learning, which is the main focus of the present study, median sample sizes are typically between 16 and 24.[1] With these samples, the minimum detectable correlation with 90 % power is .705 for contextual and probabilistic cuing tasks, and .601 for the additional singleton task.[2] Given that some amount of measurement noise is a very reasonable assumption in the measures taken in these tasks, there are good reasons to expect that the empirical correlations involving them will often be substantially smaller than these best-case values.

It is worth noting that measurement and sampling noise have related but different impacts on statistical analysis. Measurement noise induces a systematic attenuation on the effect, while sampling noise produces an unsystematic bias in its estimation. Ensuring only a large sample size without addressing the potential measurement noise from having only one trial for the awareness test is particularly risky; we could end up being very confident about an effect that is, in fact, substantially attenuated (Rouder et al., 2023b).

In the present study we illustrate the scope of these problems – measurement and sampling noise – in the domain of location probability learning by adopting a modeling-based approach. This approach will allow us to achieve two goals: First, to provide a model-based estimate of how much measurement and sampling noise there is in empirical data from these tasks; second, to find out how often researchers would mistakenly conclude that learning is unconscious, given a model based on the alternative assumption (learning is conscious) but varying the amounts of measurement and sampling noise in participants' responses. In sum, this study serves as a timely illustration for experimental psychologists on the risks of neglecting basic psychometric requirements when assessing individual differences.

Location probability learning in the additional singleton task

Numerous studies have argued that our ability to guide attention is strongly influenced by unconscious learning processes that enable us to detect and memorize statistical regularities in our environment (Chun & Jiang, 1998; Chun & Turk-Browne, 2008; Ferrante et al., 2018; Geng & Behrmann, 2005; Jiang, 2018; Gaspelin & Luck, 2018; Krishnan et al., 2022; Stilwell et al., 2019; Turk-Browne et al., 2005). While much of this research is concerned with how we learn to attend to locations where a target is likely to appear, many studies have made the complementary claim that with sufficient practice, we can unconsciously memorize and ignore locations in a scene where salient but irrelevant distractors typically appear. This phenomenon is commonly studied using the

[1] Median $N = 16$ for contextual cueing and probability cueing and median $N = 24$ for the additional singleton task (Vadillo et al., 2016, 2020; Vicente-Conesa et al., 2023).

[2] These power calculations were obtained with the function `pwr.cor.test()` from the R package `{pwr}` (Champely, 2020).

additional singleton task (e.g., Di Caro et al., 2019; Gao & Theeuwes, 2020, 2022; Lin et al., 2021; van Moorselaar & Theeuwes, 2021; Wang & Theeuwes, 2018a, 2018b, 2018c). This increasingly popular location probability learning task serves as a perfect example for the present purposes.

In a typical setting, participants are instructed to find a target with a particular shape among a series of shapes (e.g., a diamond among circles) and report the orientation of the line inside the target (see Fig. 1). Participants are warned that in some trials the display includes a salient distractor presented in a different color (e.g., a red circle among green figures, the singleton). Response times (RTs) reveal that visual search is slower when this singleton is present (e.g., Theeuwes, 1992). Presumably this happens because participants engage a strategy to detect any stimulus with a unique feature, even if this leads them to pay attention to irrelevant stimuli as well (i.e., singleton detection mode; Adams & Gaspelin, 2020; Bacon & Egeth, 1994). Crucially, if the singleton distractor appears frequently at a particular location throughout the experiment, participants are able to memorize this regularity and suppress attention to that particular location (e.g., Vicente-Conesa et al., 2023; Wang & Theeuwes, 2018a, 2018b, 2018c). This suppression effect is shown by the fact that, as learning progresses, visual search for the target is memory-guided: RTs are faster when the singleton appears at the high-probability (HP) location compared to when it appears at any other low-probability (LP) location. In essence, participants can memorize the HP location and learn not to be distracted by the singleton when it appears there. In fact, research on visual statistical learning has shaped current theorizing on memory systems (e.g., role of the hippocampus, Covington et al., 2018). Fig. 1 illustrates two possible displays of the additional singleton task for one participant.

At the end of these experiments, participants are informed that the distractor appeared more frequently at a certain location and then they are asked to recall and report this location (i.e., the awareness test). Although the suppression effect is very robust in RTs, many participants seem to be unable to correctly select the HP location (e.g., van Moorselaar & Theeuwes, 2021). When this happens, authors infer that the location probability learning was implicit. In other studies, researchers remove those participants who correctly selected the HP location – the *aware* ones[3] – from analyses and observe that the suppression effect is still significant in the remainder (e.g., Di Caro et al., 2019). Again, this is taken as evidence that learning was unconscious, in the sense that awareness is not necessary for the effect to occur.

A third strategy consists of dividing the sample between those participants who correctly selected the HP location and those who did not – the *aware* and the *unaware* ones – and comparing their suppression effects. This is usually performed through a two-way mixed ANOVA for RTs with one within-subjects factor that assesses suppression, the distractor location (HP or LP), and one between-subjects factor, accuracy in the awareness test (*aware* or *unaware* participants).[4] Obtaining a non-significant interaction between these factors has been typically interpreted as evidence that location probability learning is unconscious (e.g., Failing et al., 2019; Gao & Theeuwes, 2022; Lin et al., 2021). In other words, if the distractor suppression effect, as measured in RTs, is not related to correctly responding to the awareness test, this is taken as evidence that the suppression effect and awareness must be generated by unrelated (or weakly related) latent processes (as Rouder & Haaf, 2019, suggest for inhibition).[5] At first glance, the logic of these analyses seems completely reasonable. However, somehow implicitly, we have moved from experimental logic and entered the realm of individual differences – not because we are now interested in each individual's capacity for unconscious probabilistic learning, but because we are dividing the group based on an individual-level score (i.e., awareness accuracy). It is here where psychometric properties reign and, therefore, we need to address them to guarantee valid inferences.

Risks of blindly following this logic

An astute reader may have noticed that claiming support for a theory based on a non-significant result is scientifically unsound: As Carl Sagan (1977) said, "absence of evidence is not evidence of absence" (p. 7). Yes, a non-significant interaction can be interpreted as evidence for the claim that location probability learning is independent from awareness, but have we discounted alternative plausible scenarios in which a non-significant interaction is expected to occur? In other words, let's put ourselves in the opposite scenario and imagine that RTs and awareness are driven by a common underlying process. Would we have any reason to expect a non-significant interaction in that case too?

Imagine we are omniscient and know that there is a true interaction effect, that is, the true suppression effect in aware participants is indeed greater than in unaware ones. From our privileged point of view, the correct claim would be that "location probability learning is conscious" or, more precisely, that there is a true relationship between distractor suppression and awareness. Now imagine two research teams that want to check this interaction with empirical data. The first team has collected data and is preparing to analyse them. Unfortunately, these researchers encountered difficulty in recruitment and, although participants completed a large amount of trials, the study ends up with a small sample size. As a consequence of their lack of statistical power to detect the effect, the researchers obtain an empirically non-significant interaction. The second research team, with much better access to participants, has obtained their measures with a sufficient sample size. However, these participants had less time available and so only completed a small number of trials, increasing the measurement noise of their dependent variables. As has been well known for decades now, measurement noise tends to attenuate true effects (Hunter & Schmidt, 2015; Lord et al., 1968). Unfortunately, these researchers' measurement noise was sufficient to attenuate the actual interaction effect to the point that they also obtain an empirically non-significant interaction.

What claim might both teams risk drawing from those non-significant interactions? That the suppression effect is unrelated to awareness responses and, thus, that there is evidence supporting the claim that location probability learning is unconscious. From our omniscient perspective, we know both teams will be making an erroneous claim and can alert them about the need to first eliminate both sources of noise affecting the design (insufficient sample size: sampling noise) and measures (insufficient reliability: measurement noise).

We should not underestimate the potential impact this could be having in current theorizing. For instance, the whole idea that selection history is a "new" attentional control mechanism, different from bottom-up or top-down attention (Awh et al., 2012), is almost entirely built on the assumption that these effects are unconscious (and therefore not top-

[3] Prior to this measurement, a confidence rating is often taken on the participant's explicit awareness of the manipulation – whether they were aware that the singleton distractor appeared more probably at a certain location. Some authors classify participants as unaware not only on the basis of being unable to identify the HP location, but also based on reporting no awareness in their confidence rating.

[4] Note that this interaction is equivalent to observing a simple difference in the magnitude of the suppression effect between *aware* and *unaware* participants.

[5] Although other sources of evidence suggest that statistical learning might be unconscious (e.g., Duncan & Theeuwes, 2020; Gao & Theeuwes, 2020; Lien et al., 2024; Vicente-Conesa et al., 2024), in the vast majority of studies, poor performance in awareness tests and lack of interaction between awareness and suppression (e.g., Failing et al., 2019; Gao & Theeuwes, 2020; Wang & Theeuwes, 2018a) are the main justifications for claiming that the effect is unconscious. Our goal is only to focus attention on the risks associated with these latter types of evidence, not to defend the claim that the true nature of statistical learning is conscious.

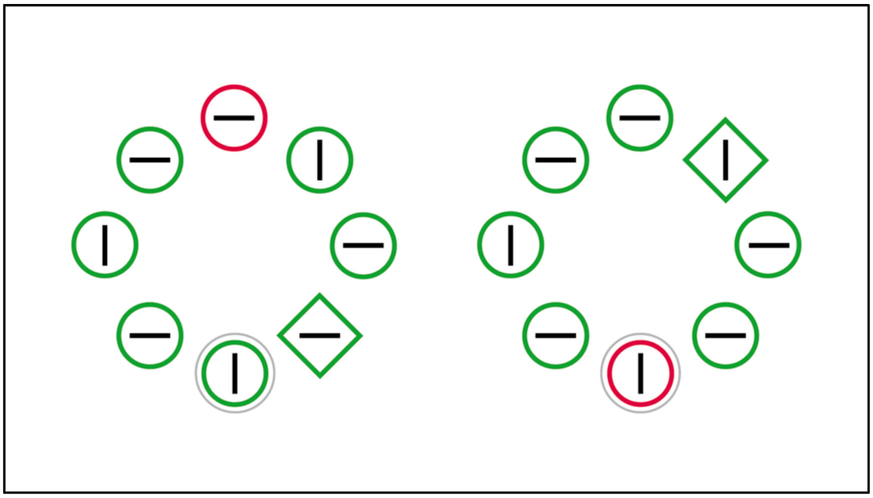

Fig. 1. *Illustration of two trials of the additional singleton task*
Note. The target is a distinct shape (diamond) and the singleton distractor is in a salient color (red). The high-probability (HP) location is indicated with a grey circle (not shown on the actual experimental trial). In the left display, the distractor appears at a low-probability (LP) location while in the right display it appears at the HP location. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

down; Giménez-Fernández et al., 2023).

Goals of the present study

Unfortunately we are not omniscient, but we can adopt the privileged position of data simulation, where the ground truth is known. To simulate data, we first need a model of how our dependent variables are distributed as a function of the independent variables manipulated in the experiment. In this article, we present a statistical model to describe both RTs and awareness responses in each participant in the additional singleton task. Given the handicaps of estimating the reliability in experimental data (impossible for the single-trial awareness measure), the model allows us to simulate data from a single-process theory (i.e., assuming that learning is conscious) while varying sampling and measurement noise.

This modeling exercise will serve as an informative illustration of how neglecting basic psychometric requirements, such as sufficient sample size and minimal measurement noise, can undermine the statistical power to detect effects of interest. Note that our goal is not to find the most accurate underlying model for real cognitive processes involved in the additional singleton task nor in location probability learning. Instead, what we will try to show is how an empirical result often found in the literature, and interpreted as evidence for a certain theory, can be generated from a contrasting theory plus some near-ubiquitous statistical artifacts. Therefore, our aim is to encourage future researchers, working on this and related tasks and research questions, to meticulously assess whether their data fulfil the crucial requirements for drawing valid conclusions about individual differences in the experimental context.

Modeling memory-guided visual search and awareness in the additional singleton task

Before entering into the details of the model, there is a general consideration we need to point out. We are interested in modeling how the experimental manipulation applied in the additional singleton task (i.e., that the distractor appears more frequently at a specific location) affects both RTs and awareness responses. In the previous literature, this influence is captured (1) as the suppression effect in RTs and (2) as accuracy in the awareness test. For the former, this influence is commonly calculated as the RT difference between LP and HP distractor conditions. Note that this specific score treats trials where the distractor appeared *close* to the HP location as equivalent to those where the distractor appears *far* from it. Similarly, for the awareness measure, authors typically divide the sample into two groups based on their awareness response: participants who correctly selected the HP location and those who failed to do so (e.g., Failing et al., 2019; Gao & Theeuwes, 2022; Lin et al., 2021). Again, note that doing this treats participants who selected a location *close* to the HP location as equivalent to those who selected a location *far* from it. In summary, both variables are dichotomized before analyzing the interaction between them.

This double dichotomization would be unproblematic if there were no evidence of a spatial gradient effect. However, many studies have found evidence for such a spatial gradient, both for RTs (Gao & Theeuwes, 2020; Failing et al., 2019; Vicente-Conesa et al., 2023; Wang et al., 2019; Wang & Theeuwes, 2018a, 2018c) and awareness responses (Vicente-Conesa et al., 2023). For instance, a typical result is that the effect of including a singleton in the search display is progressively reduced the closer it is to the HP location. Similarly, RTs to find a target progressively increase the closer it is to the HP location. Finally, in awareness tests, even when participants make a mistake, they most commonly choose locations close to the HP location. Consequently, any model for data from this task should ideally incorporate this spatial aspect, accounting for differences between distances to the HP location as an ordinal variable instead of a binary one. Below the reader will see that predictors in our model are chosen because they represent these distances and can account for the spatial gradient.

Modeling memory-guided visual search

We modeled RTs in the memory-guided visual search task with an ex-Gaussian distribution, which results from the sum of a normal and an exponential distribution. Thus, the model is $RT \sim \textit{ex-Gaussian}(\mu_{RT}, \sigma_{RT}, \tau_{RT})$, with μ_{RT} and σ_{RT} being the mean and standard deviation of the normal component of the distribution, respectively, and τ_{RT} the decay parameter of the exponential component (Luce, 1986). Previous research indicates that the ex-Gaussian distribution provides an excellent fit to RTs in a wide variety of visual search tasks (Palmer et al., 2011). Furthermore, we found that when estimating the ex-Gaussian parameters on our RTs, τ_{RT} was significantly different from zero and that the ex-Gaussian's fit is significantly better than the fit of a normal distribution for every participant.[6]

Our model assumes that, for each participant, three specific parameters, μ_{RT}, σ_{RT}, and τ_{RT}, describe their RTs on each trial. However, not every display condition will evoke the same RT distribution. For the sake of simplicity, the model predicts that changes in the properties of each display will affect the parameter μ_{RT}, but not σ_{RT} and τ_{RT}, which will remain constant for each participant[7] (for a similar approach, see Rouder & Haaf, 2019). So, what properties, manipulated in each display of the additional singleton task, can cause variations in the μ_{RT} parameter?

After evaluating alternative models with different variables and parameters (see the Supplementary Material), we finally selected a model which includes two predictors based on the two most influential features of the search display: the distance from the target to the HP location (*dT*) and the distance from the distractor to the HP location (*dD*), in each trial. In Fig. 1, two displays with different *dT* and *dD* are shown to illustrate the meaning of these parameters. In the left display, *dT* is 1 and *dD* is 4 while in the right display, *dT* is 3 and *dD* is 0. Note that, if the distractor has a distance of zero to the HP location (grey circle, $dD = 0$), this designates a HP trial (e.g., the right display).

Previous studies show how these distances affect the μ_{RT} of participants that, after some trials, have learnt (consciously or unconsciously) which is the HP location (e.g., Wang & Theeuwes, 2018a, 2018c). On trials where the *target* appears further from the HP location, they are expected to take less time to find the target, consistent with the idea of a gradient of suppression around the HP location. Note that the model suppresses unwanted interference from a distractor and inhibits wanted processing of the target via the same process. On trials where the *distractor* appears further from the HP location, participants are expected to take more time to find the target, because the distractor is not being suppressed. Separately, these influences can be mathematically expressed as the following simple linear regressions:

$$\mu_{RT} = \beta_0 - \beta_1 \cdot dT \tag{1}$$

$$\mu_{RT} = \beta_2 + \beta_1 \cdot dD \tag{2}$$

In these equations, β_0 and β_2 are the respective intercepts of each regression, the parameters for the visual search fixed cost time when *dT* or *dD* equal 0. β_0 can be conceptualized as a general visual search cost time, while β_2 is a singleton-presence cost time. The slope β_1 is the most relevant parameter in both equations. It represents the weight of *dT* and *dD*; that is, the change in μ_{RT} as the distance from the target or distractor to the HP location increases. Imagine a participant who has failed to learn that the distractor appears more frequently at a certain location. In our model, this would translate to a β_1 close to zero. Consequently, distances from target or distractor to the HP location would not affect μ_{RT}. The theoretical implication of β_1 is important in the study of location probability learning because it indicates the degree to which participants *learn* and memorize where the HP location is and create an attentional strategy to suppress it. This leads us to assign a negative sign for *dT* and a positive one for *dD*: A shorter *target* distance to the location where the singleton distractor appears with high probability and a larger *distractor* distance to this location result in slower target detection by the participant (consistent with a gradient of attentional suppression around this HP location that can affect processing of both the target and the singleton distractor); thus, a higher μ_{RT} is expected in those trials.

In the Supplementary Material, readers will find technical details explaining why fixing the slope parameter β_1 to the same value in both regressions does not result in a significant worsening in model fit. The same reasoning motivated us to fix to zero the slope of a third possible but discarded predictor, the target–distractor distance (*dTD*; van Moorselaar & Theeuwes, 2021), which was found to have little influence on μ_{RT} (also presented in the Supplementary Material). In summary, the complete model for RTs can be expressed as

$$RT \sim ex\text{-}Gaussian(\mu_{RT} = \beta_0 - \beta_1 \cdot dT + x \cdot (\beta_2 + \beta_1 \cdot dD), \sigma_{RT}, \tau_{RT}), \tag{3}$$

where x is a switch variable which takes the value 1 in trials with a distractor and the value 0 in trials without a distractor. This way, we can activate the part of the formula that involves *dD* only when the display includes a distractor. The model assumes that the distance from the target to the HP location, *dT*, is important on all trials, even when no distractor is actually presented.

How measurement noise in RTs is modeled. According to the model, the process we want to measure (i.e., location probability learning, represented by β_1) is included only in μ_{RT}. Every other influence affecting the distribution can be considered a source of uncertainty, which is here represented by the ex-Gaussian variability parameters, σ_{RT} and τ_{RT}, or more concisely by their variance (calculated as $\sigma^2_{RT} + \tau^2_{RT}$). The larger σ_{RT} and τ_{RT} are, the more the RTs will tend to deviate from the true μ_{RT} for the corresponding experimental condition. In this sense, these parameters can be understood as sources of noise that displace individual observations from the values that would be expected based on the latent process responsible for the experimental effects. The most significant effect of this measurement noise is to attenuate the size of the manipulation effect. Fig. 2 illustrates this attenuation. A participant with a low learning parameter (Fig. 2A) is expected to show small differences as *dD* varies. As this parameter increases (Fig. 2B), the ex-Gaussian curves for different distractor distances overlap less. However, if these RTs are measured with excessive measurement noise (i.e., increasingly larger values of σ^2_{RT} and τ^2_{RT}), the distributions will be flattened and overlap relatively more (Fig. 2C).

Modeling awareness

Our model is compatible with a single learning process, that is, the same underlying mechanism that determines RTs as a function of target and distractor distances to the HP location also determines the response in the awareness test. Consequently, this model assumes a conscious learning process. Specifically, the model states that when participants are explicitly asked to identify the HP location, their single awareness response, AW, is extracted from a multinomial distribution with the eight locations as categories:

$$AW \sim Multinomial\left(1,\ p_l = \frac{e^{\frac{w_l}{Q}}}{\sum_{i=1}^{L} e^{\frac{w_i}{Q}}}\right), \tag{4}$$
$$\text{where } w_l = -\beta_1 \cdot dL_l, \text{ for each } l \in \{1,\ \cdots,\ 8\},$$

where p_l is the probability of selecting location l as the HP location in the awareness test at the end of the experiment. This probability is calculated with the SoftMax rule, $\frac{e^{w_l/Q}}{\sum_{i=1}^{L} e^{w_i/Q}}$, a function that transforms a vector of eight numbers (one per location of the display) into probabilities ranging from 0 to 1, whose sum equals 1. These numbers or weights, w_l, reflect the influence of the participant's learning, β_1, at the time of the

[6] With a single-sample *t*-test, both in HP ($t_{(158)} = 36.044$, $p < .0001$, $d = 3.21$ with 95% CI [2.82, 3.59]) and LP ($t_{(158)} = 40.454$, $p < .0001$, $d = 2.86$ with 95% CI [2.51, 3.21]) conditions. We ran Likelihood Ratio Tests to compare ex-Gaussian and normal fits for every participant (α =.05) and the largest *p*-value was 0.0067. The R script for these analyses is available at osf.io/czxhr/.

[7] We tested if this assumption was reasonable by estimating the three parameters in both HP and LP distractor conditions separately and comparing each pair of parameters with a repeated-measures *t*-test (α =.05). This analysis indicates that μ_{RT} is the parameter with the largest difference across conditions ($t_{(158)} = 22.955$, p <.0001, $d = 1.82$ with 95% CI [1.57, 2.07]), much larger than σ_{RT} ($t_{(158)} = 6.592$, p <.0001, $d = 0.52$ with 95% CI [0.36, 0.69]) and τ_{RT} ($t_{(158)} = -10.922$, p <.0001, $d = -0.87$ with 95% CI [-1.05, −0.68]), although these two also reached significance. The R script for these analyses is available at osf.io/baxf8.

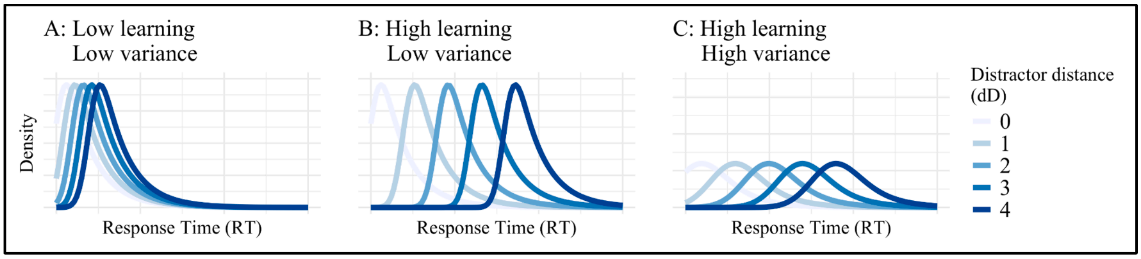

Fig. 2. *Illustration of modeled distributions for response times (RTs) in five different trial conditions each with a specific distance from distractor to high-probability location (dD), as a function of variance in RTs*
Note. These ex-Gaussian curves do not follow the complete model presented in Equation (3), but a reduced example using only one predictor, the distance from distractor to high-probability location (dD).

awareness test, as a function of the distance, dL_l, of a location l to the HP location. Consequently, w_l is the result of another simple linear regression: $-\beta_1 \cdot dL_l$. Indeed, this β_1 is the same parameter from the RT Equation (2). The parameter Q will be explained later.

Imagine a participant who has not learnt anything about where the HP location was across trials. As explained above, their parameter β_1 would be close to zero. In that case, at the time of responding to the awareness test, the distance from each location to the HP location, dL_l, would not affect their probability of selecting any location in particular and every w_l would be close to 0. In other words, they would have equal probability of selecting each location as the HP one (Fig. 3A). Conversely, for a participant demonstrating greater learning regarding the HP location, the parameter β_1 increases and the influence of dL_l becomes stronger. Specifically, locations with shorter dL_l should have greater probabilities of being selected in the awareness test, thus the negative sign in the equation for w_l (Fig. 3B). This property of the model allows researchers to use the full information from their awareness test (unlike the dichotomization typically done in previous studies), as it crucially differentiates a participant who almost guessed the HP location correctly ($dL_{AW} = 1$) from another participant who completely missed the HP location ($dL_{AW} = 4$).

How measurement noise in AW is modeled. In Fig. 3B, the influence of learning (β_1) on the multinomial distribution from which the awareness response is extracted will be most pronounced when the temperature parameter, Q, equals 1. Fig. 3C represents the same learning as in Fig. 3B, but with a higher value of Q. Note that this temperature parameter attenuates the influence of learning on the awareness response, increasing the probability of choosing any other (incorrect) location instead. Similar to RTs, the temperature parameter allows us to model the amount of measurement noise in the awareness response variable.

In the following section, we present two modeling studies. First, we fit the memory-guided visual search and awareness model to real data from the additional singleton task (Vicente-Conesa et al., 2023). We show that the model predicts most of the trends observed in the data, including a small interaction between learning and awareness that often will fail to reach statistical significance in empirical studies with small and unreliable samples. Next, we report the results of a simulation assessing how statistical power to detect this interaction between learning and awareness can be expected to change as a function of measurement and sampling noise in both dependent variables (RTs and AW).

Fitting the model to observed data

Description of the dataset

The data used to fit the model were collected by Vicente-Conesa et al. (2023). The participants were recruited in three separate experiments of $N = 80$ each, but as explained below we only modeled data from Experiments 1 and 2. It is worth noting that these sample sizes are much larger than the typical ones found in this literature (a median of 24 participants from all studies reviewed in the Supplementary Material of Vicente-Conesa et al., 2023). Furthermore, data were collected and analyzed following a preregistered protocol. The memory-guided visual search task was closely based on previous studies (Wang & Theeuwes, 2018a, 2018b). This task, as explained in the Introduction, consisted of searching for a target (with a distinct shape) in a circular display among other distractors in the same color. In some trials, the search display included a salient colored distractor (the singleton) which appeared more frequently at a specific location, enabling participants to discover and memorize this location trial-by-trial, resulting in a very robust suppression effect on search times. Although each participant completed 720 visual search trials, the number of valid trials for each participant – after applying the same exclusion criteria as in the original study – ranged from 585 to 685 with a mean of 661 trials. The memory-guided visual search and awareness model presented in Equation (3) was implemented in the R language (Version 4.3.1, R Core Team, 2023) and all scripts to reproduce our results are available at osf.io/y8v74.

Each experiment varied only in the measurement method for the awareness test at the end of the task. In Experiment 1 participants were asked to identify the HP location while in Experiment 2 they ranked the three locations that in their opinion had contained the singleton distractor most frequently. It is relatively easy to fit our model to these two dependent measures (ignoring locations ranked second and third in Experiment 2) and, accordingly, we used data from both of them. Eighty one out of the 159 participants (46 in Experiment 1, 57.5 %, and 35 in Experiment 2, 44.30 %) correctly guessed the HP location.[8] In Experiment 3, participants estimated how many times the distractor appeared at each location, and thus were not asked directly to select the HP location. Therefore, we excluded this experiment and our dataset contains RTs and AW from 159 participants (data from one participant were excluded due to a technical problem). To our knowledge, most of the studies using a comparable task did not make their data publicly available and the ones that did either did not include awareness data (Failing et al., 2019) or RTs were not interpretable as we do in this model due to a distinct design (van Moorselaar & Theeuwes, 2021).

Parameter estimation

We estimated the five parameters of the model (β_0, β_1, β_2, σ_{RT}, and τ_{RT}) separately for each of the 159 participants. We searched for pa-

[8] In Wang and Theeuwes (2018b), 47% of participants correctly guessed the HP location. In other studies, we have no access to this percentage, either because participants not initially aware of the manipulation did not report their guessing (Wang et al., 2019) or because authors did not report the total (Di Caro et al., 2019; Gao & Theeuwes, 2020). Other studies with similar (but possibly significantly distinct) tasks have reported that 31%, 20%, 32.5%, or 36.4% (Lin et al., 2021; van Moorselaar & Theeuwes, 2021, 2022, 2024, respectively) of participants guessed the HP location.

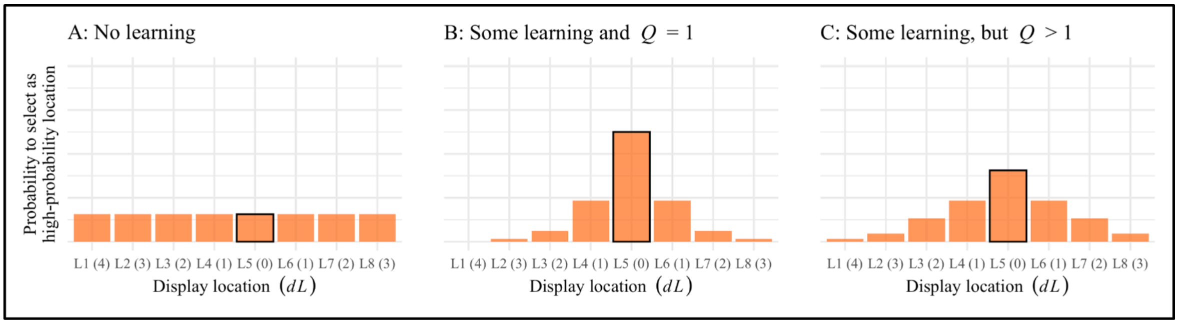

Fig. 3. *Illustration of modeled location selection distributions in the awareness test*
Note. Panel A represents no learning about the high-probability location (L5, with black border). Panels B and C represent successful location learning, in B with a temperature (Q) of 1, and in C with a Q higher than 1.

rameters that maximize the log-likelihood of the model using the `optim` function from the `{stats}` package in R (R Core Team, 2023) and the Nelder and Mead (1965) algorithm, as recommended by Balota et al. (2008). We fitted the basic ex-Gaussian distribution to all trials for each participant (without any predictors) and selected the resulting parameters μ, σ, and τ as starting values to estimate parameters β_0, σ_{RT}, and τ_{RT} in the model, respectively. The starting values for the remaining parameters, β_1 and β_2, were fixed to zero. R scripts for these estimations are available at osf.io/y8v74.

Note that the temperature parameter, Q, cannot be estimated at the participant level because there is only one observation (the single-item awareness measure) per participant.[9] To obtain a plausible estimate of Q, we first estimated the rest of the parameters with Q fixed to 1. Then, for each participant, we analytically calculated the probabilities of selecting each location of the display in the awareness test, using Equation (4) of the model. This procedure was repeated by increasing the Q parameter (common to all participants) until we found the value of Q for which the multinomial distribution of probabilities across participants best reproduced the observed distribution of awareness responses in Vicente-Conesa et al. (2023).

To assess model fit, we analytically computed the predicted RT for each trial as the mean $\mu_{RT} + \tau_{RT}$ resulting from applying dT and dD for the corresponding trial display and the estimated parameters of the corresponding participant. We obtained this prediction analytically to avoid including random estimation noise. For the awareness response in each participant, we obtained the probability of each dL_l resulting from using the SoftMax rule in the model's Equation (4).

Table 1 presents the descriptive statistics for each parameter in Vicente-Conesa et al.'s (2023) sample of 159 participants. As can be seen in Fig. 4, at the group level, the model provides an adequate fit to the mean RTs across distractor conditions, no-distractor (ND), HP, and LP (Fig. 4A), distractor distances ($dD = [0\ldots4]$) in distractor-present trials (Fig. 4C), and target distances ($dT = [0\ldots4]$) in distractor-absent trials (Fig. 4E). Note that, for all conditions, the 95 % CIs of the predicted means (dots) overlap with the 95 % CIs of the observed means (bars), suggesting that both means can be considered equal (at the $\alpha = .05$ level). It is possible that the greatest discrepancies between predicted and observed mean RTs (i.e., conditions $dD = 2$ and 3 in Fig. 4C, and $dT = 1$ and 2 in Fig. 4E) are due to a non-linear tendency underlying the distances. For the sake of simplicity, we used a linear function for both predictors (dD and dT), but other alternatives, such as an exponential function, could be worth exploring in future models. At the participant level, the model also performs adequately as illustrated in the right panels (4B, 4D, and 4F), since the dots are systematically distributed around the diagonals of the observed-predicted scatterplots (R^2 ranging from .88 to 1.00).

Table 1
Descriptive statistics for each parameter for the data from Vicente-Conesa et al. (2023).

Parameter	Mean	Standard deviation	Min	Max
β_0	682.46	138.34	430.48	1293.85
β_1	10.50	11.37	−2.00	47.18
β_2	65.61	45.59	−25.75	265.00
σ_{RT}	103.43	51.43	30.58	306.86
τ_{RT}	390.69	141.48	166.16	1199.59

Although Fig. 4 represents the trends in the data typically analyzed in this literature (e.g., corresponding to Vicente-Conesa et al., 2023, Figs. 2 and 3) and our model reproduces the observed results at that level of analysis, the fit of the model can also be assessed at a finer level of granularity. Fig. 5 illustrates the model fit of mean RTs in each of the 28 possible display conditions (i.e., 28 combinations between possible values of dT and dD). Here we can appreciate the caveats of the model predictions, especially when dT equals 0, that is, when the target appears at the HP location. Although the model predicts a pattern of increasing RTs as dD increases (common to every dT), the observed data show the opposite pattern: participants find the target *faster* as the distractor is located further from the HP location, where the target is located. The model would predict this effect if we had a term for the distance between target and distractor (dTD). As noted above, we excluded this third predictor on the basis that adding it did not significantly improve the model fit (see the Supplementary Material). Future models focusing on precise qualitative aspects of performance (van Moorselaar & Theeuwes, 2021) could include this predictor.

Finally, Fig. 6 illustrates the observed and predicted performance in the awareness test, specifically, how frequently participants selected locations at each distance from the HP location (dL_{AW}) in the observed data (bars) and the average probability of each response according to the model (dots). A temperature of 7 was the value which generates the predicted dL_{AW} probabilities most similar to the observed relative frequencies, reducing average error between them to 0.06 in proportion units.

Detecting the noisy interaction between suppression and awareness

Researchers using the additional singleton task or similar tasks tend to obtain a non-significant interaction between suppression (measured by the difference in RTs between HP and LP conditions) and awareness (whether the participant correctly selects the HP location, *aware*, or not,

[9] With only one observation in the AW equation, that fragment of the model is locally non-identifiable. This means that trying to estimate more than one parameter would imply convergence issues. For the same reason, we could not contrast an alternative model where learning is different in each equation (for instance, estimating β_1 in the AW equation as a new parameter, α_1, independent from β_1 in the RT equation).

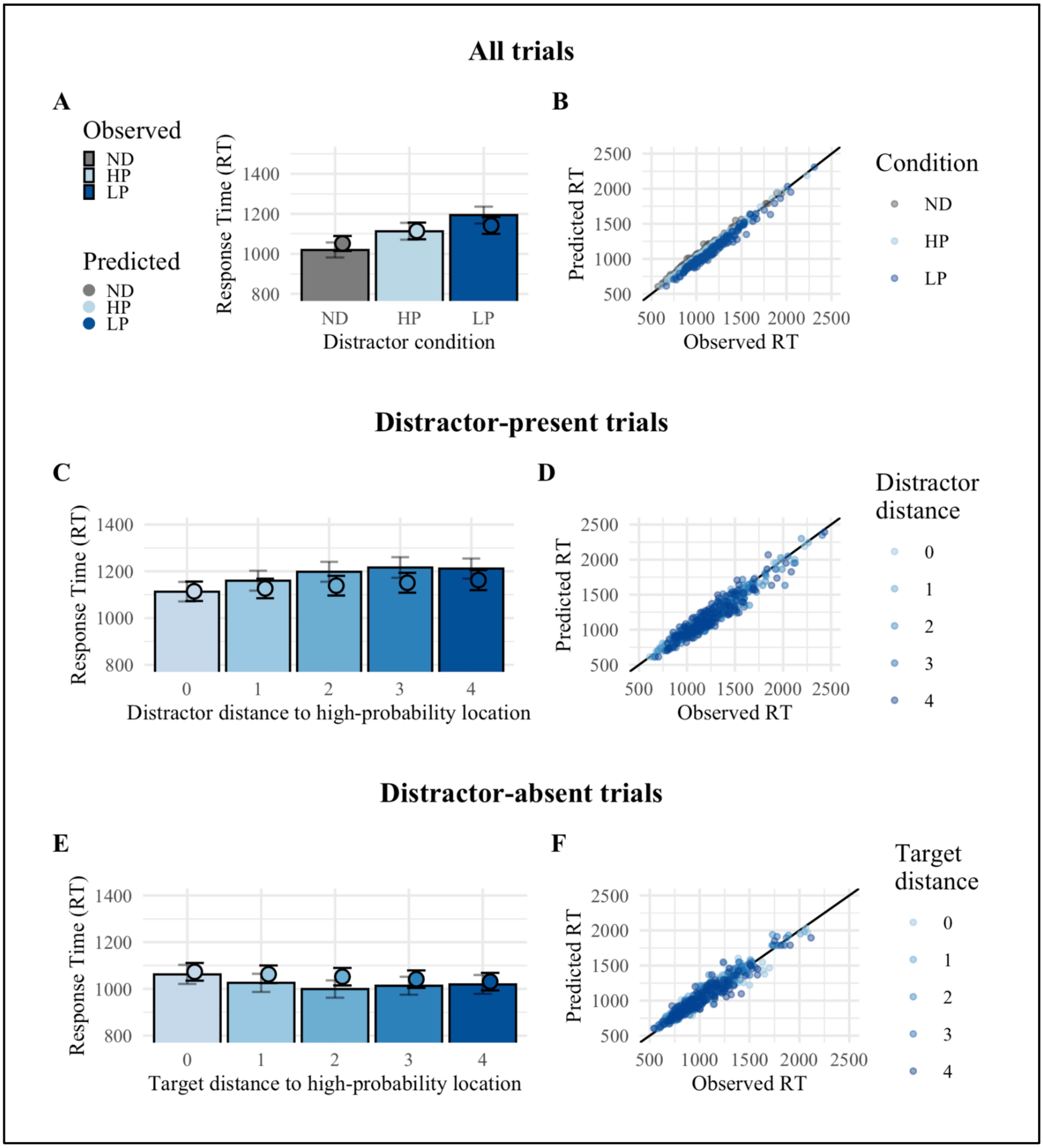

Fig. 4. *Response times (RTs) compared between observed (bars) and predicted data (dots), across distractor conditions (first row), distractor distances (second row), and target distances (third row). Figures on the right represent the fit between observed and predicted mean RTs for each participant in each condition*

Note. ND: No-distractor trial; HP: High-probability trial; LP: Low-probability trial. Panel A represents the same data as in the left panel of Fig. 2 in Vicente-Conesa et al. (2023, p. 801). Panels C and E represent the same data as in panels A and C, respectively, of Fig. 3 in Vicente-Conesa et al. (p. 802).

unaware). When analyzing Vicente-Conesa et al.'s (2023) dataset, we obtained the same non-significant interaction, $F_{(1,\,157)} = 0.45$, $p = .501$, $\eta_p^2 < .001$. This result has typically led researchers to infer that learning is unconscious, neglecting to test whether the non-significance is compatible with a true latent interaction that has been attenuated by excessive sampling and/or measurement noise.

First of all, can we empirically study the amount of distractor suppression sampling and measurement noise in Vicente-Conesa et al.'s (2023) dataset? We know that the sampling noise corresponds to 159 participants. We also know that regarding measurement noise on the first variable, the dataset has approximately 661 RTs per participant. Although our intuition might lead us to expect that the measurement noise in RTs will be low, due to having many trials, empirical reliability estimations suggest the opposite. The mean permuted split-half reliability across 1,000 replications was .415[10] and the signal-to-noise ratio (Rouder et al., 2023a) was $\gamma = 0.0705$ (1-to-14), which suggests a similar

[10] After applying Spearman-Brown's correction. The R script to compute this coefficient is available at osf.io/7cqtu.

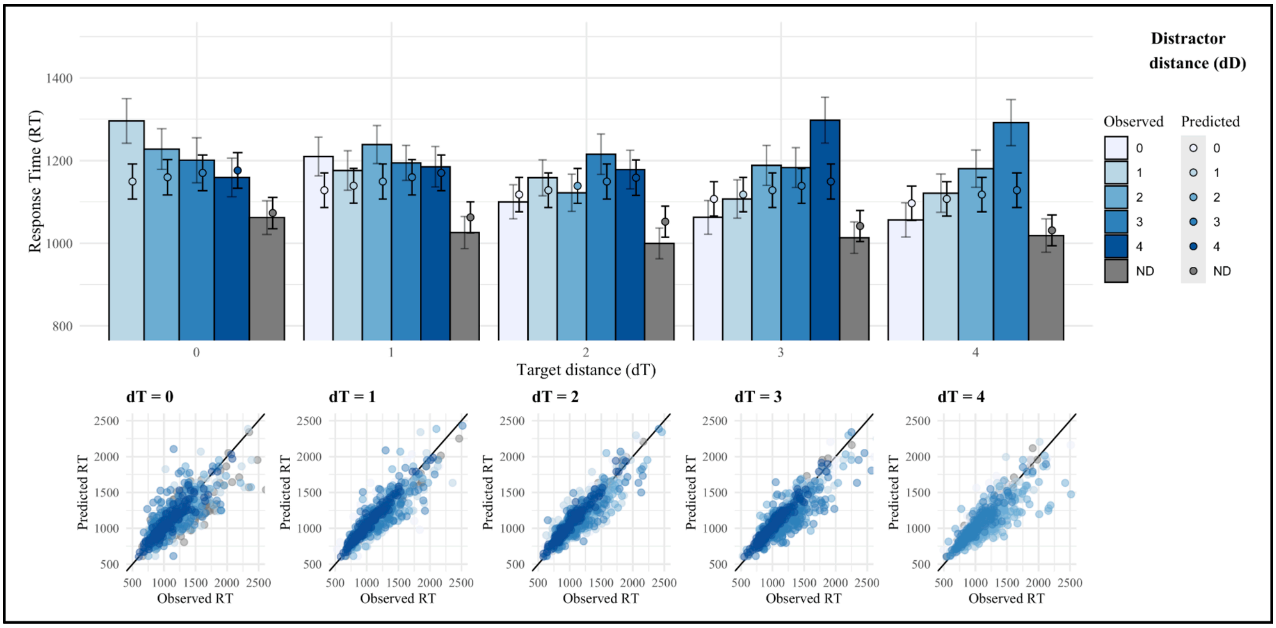

Fig. 5. *Top panel: Comparison between observed (bars) and predicted (dots) response times (RTs), across dT (distance from target to high-probability location) and dD (distance from distractor to high-probability location) conditions. Bottom panel: Fit between observed and predicted mean RTs for each participant in each dT and dD condition Note.* ND: no-distractor trial. Due to differences in trial numbers in each condition and participant, reported observed RTs are the fixed intercepts from an intercepts-only mixed model for each condition (Level 1) and participant (Level 2), which isolates the within-participants variability. The model predicts the same RT for every trial from the same condition and participant, so the dots are the mean across participant predictions in each condition (one per participant).

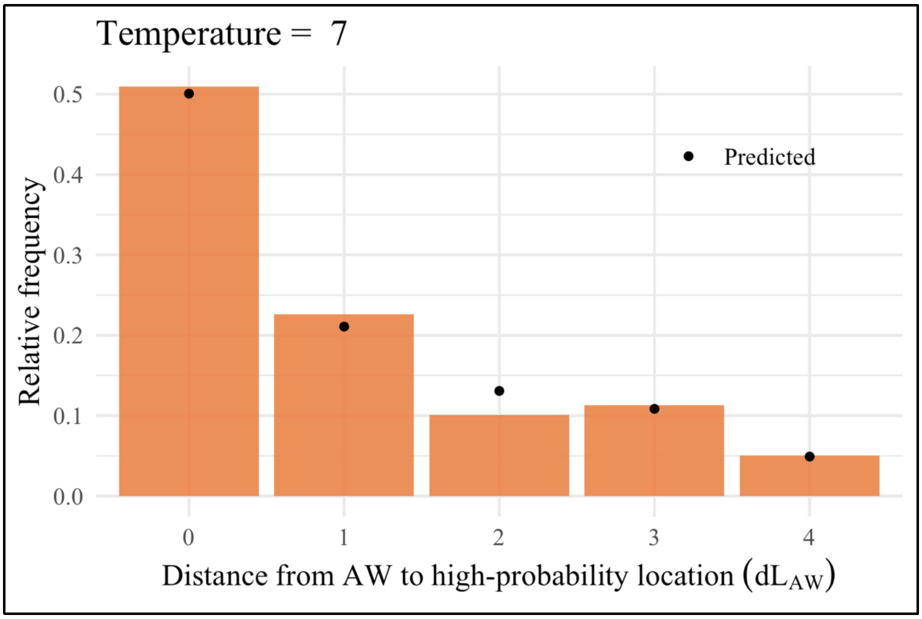

Fig. 6. *Comparison of relative frequencies of selected awareness response (AW) in observed (bars) and predicted (dots) data, as a function of their distances from this location to the high-probability location (dL$_{AW}$)*

reliability (.450) when assuming all participants completed 660 trials (330 per condition).[11] Unfortunately, none of the previous methods for estimating reliability are applicable for the second variable, AW, since empirically there is only one response to the awareness test per participant. Luckily, we still can approximate these estimations following the modeling approach.

In the following section we simulate data according to the model described above. The model assumes a true interaction because learning is characterized by a single latent process, represented by the parameter β_1, which affects both RTs and awareness. Hence this simulation will quantify the statistical power to detect this interaction under realistic sampling and measurement noise.

In the previous section, we described the model's ability to predict search times and awareness by analytically calculating the mean RTs

[11] The signal-to-noise ratio was computed as the square root of the ratio between the variance of the random slope for the trials condition (HP and LP) and the residuals variance, based on a mixed model with the function `lmer()` of the R package `{lme4}`. We made two noteworthy assumptions: that all participants completed the same number of trials (660 per participant) and that trial conditions were balanced (330 per condition), although in fact there were twice as many HP than LP trials in the empirical dataset.

and probabilities of choosing each location as AW. These predictions are noiseless since there is a unique predicted value for each condition and participant. For the following illustration, we more appropriately simulate data with trial noise, randomly sampling each RT or AW from its corresponding distribution (the ex-Gaussian for RTs and the multinomial for AW).

Sampling noise

Fig. 7 shows the mean predicted interaction effects across 1,000 simulated samples (A) with the same design properties (trial conditions, and numbers of trials, and participants) as in Vicente-Conesa et al.'s (2023) dataset, and (B) with a random subsample of 24 participants.

Further analysis reveals that, for the complete sample, the interaction effect was significant in 99.9 % of the simulated samples. While this level of power seems reassuring at first glance, it is based on a very large sample ($N = 159$). More importantly, the power to detect an effect of this size ($\eta_p^2 = .0013$ or $d_s = 0.79$) with a sample equal to the median sample size in this literature (24 participants among all studies reviewed in the Supplementary Material of Vicente-Conesa et al., 2023) is merely 45.3 %. This implies that obtaining a non-significant interaction (as has been consistently observed in previous studies; e.g., Failing et al., 2019; Gao & Theeuwes, 2022; Lin et al., 2021) is more likely than obtaining a significant one, even when this interaction actually exists. If a researcher wants to achieve acceptable statistical power to detect this (plausible) interaction,[12] a sample of at least 69 participants is required to guarantee 90 % power. Fig. 8 illustrates the power curve for sample sizes from 24 to 160.

Measurement noise

Note that the predicted effect size of the interaction is medium-sized across simulations. This is the main reason why larger sample sizes are needed to guarantee sufficient power. However, increasing sample size is not the only means to avoid false-negative conclusions. Indeed, remember that caring only about having a large sample while neglecting measurement noise can inappropriately increase our confidence about an attenuated effect (Rouder et al., 2023b). To explore to what extent this statistical artifact might be affecting the simulated variables, we estimated their reliability by correlating each variable (suppression score, AW accuracy, and distance from the AW to the HP location) in two different replications. Each violin plot in Fig. 9 represents the correlations between 1,000 random pairs of replications.

Regarding measurement noise in RTs, the leftmost plots in Fig. 9 illustrate what we suspected: there is a reliability problem in suppression scores (Pearson's correlation, $M_r = .317$, $sd_r = 0.085$). In fact, this mean reliability predicted by the model is included within the 95 % CI [.200, .597] for split-half reliability previously estimated in the data.

While regarding measurement noise in AW, we already mentioned that the empirical reliability of the awareness test cannot be computed, since it consists of a single-trial measure. Fortunately, our model allows us to study its consistency across simulated replications: both for the binary AW accuracy variable (phi coefficient, $M_r = .476$, $sd_r = 0.067$, middle plot of Fig. 9) and for the distances (0–4) from the AW response to the HP location (Spearman's correlation, $M_r = .513$, $sd_r = 0.060$, right plot), the latter being a marginally more robust measure of awareness compared to the first. Nonetheless, with such measurement uncertainty, again, a non-significant interaction is almost the only empirical scenario a researcher should expect.

What if a researcher obtained half of the measurement noise that we estimate for Vicente-Conesa et al.'s (2023) data? Or a quarter? Translucid plots in Fig. 9 represent these scenarios. We simulated the same original data, but reducing the impact of measurement noise both in RTs (by multiplying the original values of σ_{RT} and τ_{RT} by the square root of 0.50 or 0.25[13]) and in AW (by multiplying Q by 0.50 or 0.25). The result, as expected, is an increase in the measures' reliability, but also in the effect size of the interaction. With half the measurement noise, the interaction increases from $d_s = 0.79$ to $d_s = 1.08$, for which 38 participants would be sufficient to achieve 90 % power. In turn, with a quarter of the measurement noise, the interaction reaches $d_s = 1.25$ and only 29 participants (close to the median sample size in the literature employing the additional singleton task) would finally be sufficient to detect it with 90 % power.

Discussion

The modeling approach presented in this work allows researchers to assess the power of inferences in an experimental context in a twofold manner. First, our model allows us to estimate the amount of measurement noise in data collected in the additional singleton task, given a specified level of sampling noise. When these data lead to a non-significant interaction, as in Vicente-Conesa et al.'s (2023) dataset, the model permits us to know if the data are compatible with the theoretical claim that location probability learning and awareness truly interact, but the observed interaction is attenuated by measurement and sampling noise. The model fit of Vicente-Conesa et al.'s dataset is adequate in most of its predictions. Additionally, it allowed us to ensure that having this large sample size (159 participants is far more than what other studies with the same paradigm have reached) is sufficient to achieve high statistical power to detect a small-to-moderate interaction. However, samples like the ones prevalent in the current literature, with a median size of 24 participants, are insufficient (power of 43 % to detect the interaction). This is likely due to a serious problem of measurement noise in the suppression and awareness variables, as they have been usually measured.

Second, our model allows us to simulate data for which the correct theoretical model is known, so we can evaluate how sample sizes and reliabilities impact our capacity to infer this model. For the additional singleton task, previous authors have systematically interpreted a null interaction between suppression and awareness as evidence of location probability learning being unconscious. Consistent with Hedge et al. (2018) and LeBel and Paunonen (2011), our results suggest that, even under a model where that interaction is true, the power to detect it rapidly decays as a function of measurement and sampling noise, so obtaining a null interaction becomes a very likely outcome in empirical contexts. This is particularly discouraging for research teams constrained in their ability to access large samples. Instead, they need to obtain measures with little noise to be relatively sure they are able to detect their effects of interest. For instance, one relatively simple way of achieving this is to include more trials per participant. Additionally, researchers should explore whether the overall reliability of the task can be improved with adjustments in the procedure (e.g., Vadillo et al., 2024; Viviani et al., 2024) and in the dependent variable (e.g., Kim et al., 2025). This work also illustrates how easily an effect size can be vulnerable to sampling noise. Hedge et al. (2018) found that even assuming moderate effect sizes, the required sample size for 80 % power still exceeds the typically used sample sizes. Consequently, future research will need to exhaustively study how the true size of the intended effect moderates the relationship between measurement/sampling noise and power.

The modeling approach we adopted opens up alternative methods to

[12] These power calculations were obtained with the function `pwr.t.test()` from the R package `{pwr}` (Champely, 2020). Note that, to obtain these numbers, an interaction of 0.79 in d_s units (mean value obtained across simulated samples) was taken as the population effect size and that balanced groups (*aware* and *unaware*) were assumed.

[13] The justification for using the square root is to transform the scaling factor (0.50 or 0.25) to the original metric of parameters σ_{RT} and τ_{RT}.

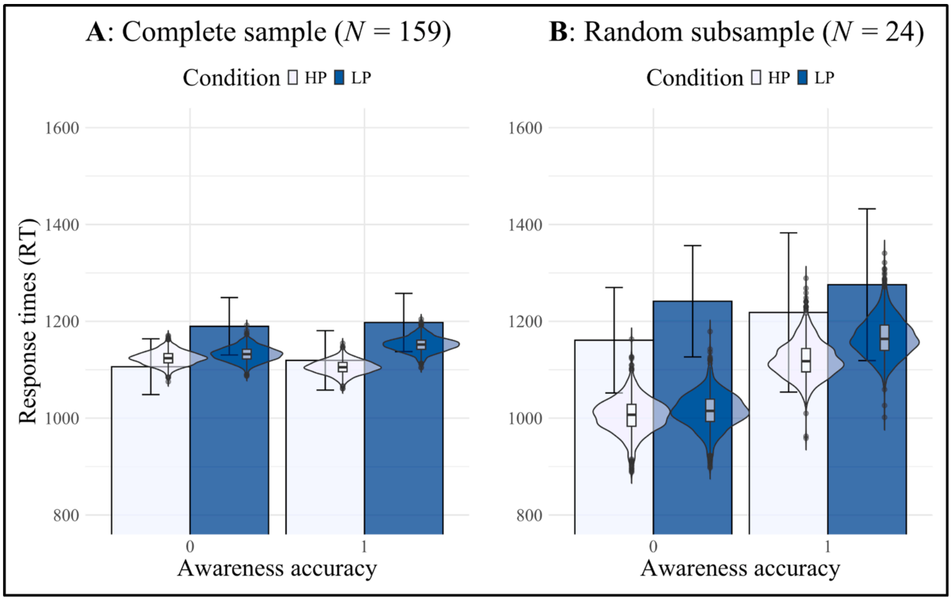

Fig. 7. *Interaction effect in response times (RTs) between trial condition (high-probability, HP, and low-probability, LP) and awareness accuracy (0, failing to select the HP location, and 1, correctly selecting the HP location). Bars represent the observed data by* Vicente-Conesa et al. (2023) *and the violins the distribution of simulated results Note.* Error bars represent the 95% CI. Panel A includes the complete sample of 159 participants, while Panel B includes a random subsample of 24 participants from the complete sample. In both cases the aware and unaware groups were almost balanced (81 vs. 78 and 13 vs. 11, respectively).

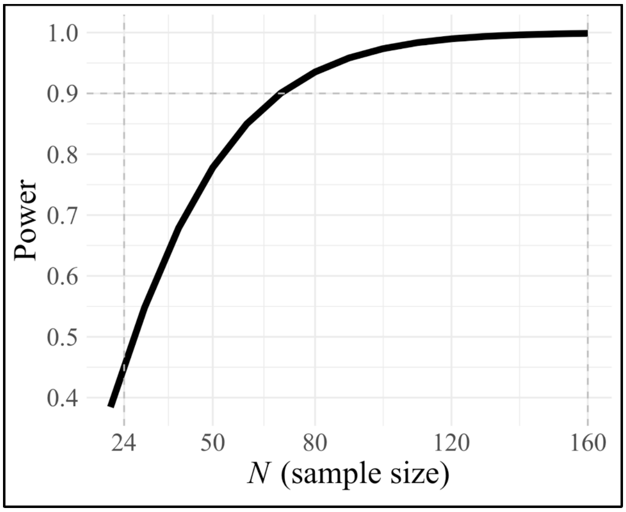

Fig. 8. *Statistical power to detect the interaction predicted by the model across a range of sample sizes*

estimate reliability in cases where standard methods (e.g., split-half reliability) are inapplicable, as when collecting a single-item measure. The model-based reliability calculated here, computed as the average correlation across pairs of simulated replications, provides a rough estimation in the absence of better evidence. However, we strongly recommend that future designs for the additional singleton task include more informative measures of awareness, if such a construct – and its measures' reliability – is to be properly studied (Adams & Gaspelin, 2020; Cronbach, 1957). For instance, researchers could include multiple awareness measures across blocks to avoid relying on a single item, similar to the online measurement proposed by Lu et al. (2022). Intuitively, repeatedly asking participants about their awareness may influence their awareness itself. To our knowledge, no study has yet examined the extent of this potential influence. Another way to collect more informative awareness measures, proposed by one of the reviewers of this paper, is to use thought probes during the tasks, as has been done in mind-wandering research (Weinstein, 2018).

Limitations and future enhancements

Although we have offered this model as an alternative way of studying the reliability of a single-item measure (such as awareness in this task), this still implies irreconcilable issues, not only computationally, but also theoretically. Remember we have presented the temperature parameter, Q, as an attenuation factor that can be understood as measurement noise. After all, its influence is analogous to adding some amount of Gaussian noise to a true variable (see changes from Fig. 3B to 3C). Of course, Q could be reinterpreted to align more with the claim that learning is unconscious (e.g., that is another unconscious process that obscures the true learning). While this interpretation is logically possible, we doubt it is consistent with the original intent of researchers in this field. An unconscious model like this could explain any pattern of association between distractor suppression and awareness, from no relationship to perfect alignment, depending on the value assigned to Q. If taken seriously, this would mean that none of the many distractor suppression studies assessing awareness actually tested unconscious learning, as this claim becomes unfalsifiable with such a methodology. Note that, despite superficial similarities, this does not apply to our conscious model in the same way. In our framework, Q has a very specific meaning, measurement noise, and therefore any manipulation that demonstrably reduces measurement error should favor the observation of an association between learning and awareness. Nevertheless, we emphasize that our model and analyses should not be taken as evidence that learning is conscious. Rather, this study points to a troubling conclusion: with current designs, we cannot yet distinguish unconscious processing from noisy data and additional empirical evidence is needed to disentangle the possible interpretations of this Q parameter.[14]

Finally, both our model and simulations offer numerous opportunities for enhancement. First, the memory-guided visual search and

[14] For example, in an experiment that manipulates the participant's level of awareness through instructions, varying the explicitness of the HP location. If Q truly reflects measurement noise, its value should remain constant regardless of the instruction condition, and differences should only be observed in β_1.

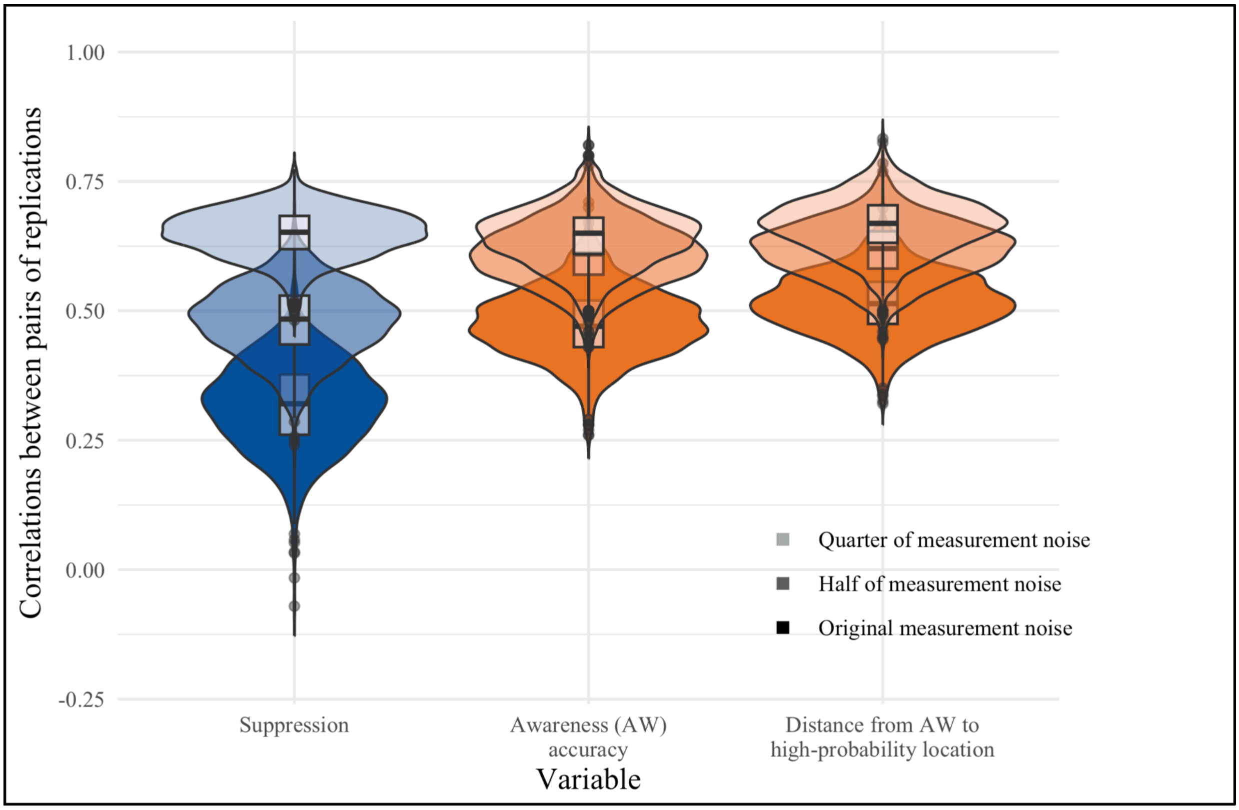

Fig. 9. *Reliability of simulated variables (suppression, AW accuracy, and distance from the AW response location to the high-probability location, dL_{AW}), computed as the correlation between 1,000 random pairs of replications. The more transparent the violin plots, the lower the measurement noise of the simulated variables*
Note. We used Pearson's correlation for suppression scores (continuous variable), the phi coefficient for AW accuracy (binary variable), and Spearman's correlation for dL_{AW} (categorical variable).

awareness model presented here represents just one of several options we considered (see Supplementary Material) and other empirical datasets should be used to gain more evidence on which model fits the data better. Also, it would be profitable to check if other non-linear functions better represent the distances from target and distractor to HP location (e.g., an exponential function). Note that here we limited our estimations to the participant level, but other approaches such as multilevel models would offer some interesting additional information (e.g., Rouder & Haaf, 2019, found that their effect estimated from a hierarchical model was more accurate than conventional aggregated estimates). We suspect that, to fulfil the requirements of multilevel models, a Bayesian approach will be more suitable and could overcome potential convergence difficulties (Levy & McNeish, 2023).

We conclude that the heart of the problem in failing to detect an effect does not lie in the reliability of experimental measures themselves. Within the very nature of studying individual differences there is a core concept: between-participants variability. Probably the main scientific task in this field is to evaluate potential relationships between psychological variables. For instance, when studying intelligence or personality traits, no one will doubt that a homogeneous sample would pose a challenge in finding correlates with other variables (i.e., a variable with small variance is unlikely to correlate with others). In contrast, experimental psychology seeks other goals such as finding robust basic effects, prevalent across most individuals within a population. This goal inherently conflicts with variability (Borsboom et al., 2009; Cronbach, 1957) and, when it is low or absent, our measures can neither correlate with others, nor correlate with themselves.

In this latter scenario, a measure that does not correlate with itself entails a measure that cannot, by definition, be reliable. This is a key message: the measure can be unreliable, not because it is affected by measurement noise in the population, but because the sample in which it was measured has little variability. The increasingly widespread practice of reporting estimations of reliability is revealing that many common experimental tasks have substantial amounts of measurement noise (Hedge et al., 2018; Paap & Sawi, 2016), although not always (Viviani et al., 2024). However, as Hedge et al. and Rouder et al. (2023a; 2023b) suggest, we are not yet sure whether this implies that our measures need to be enhanced or that their nature in experimental settings makes the classic study of reliability unfeasible.

Some design and statistical artifacts can produce this variability reduction: range restriction, dichotomization, and unbalanced designs. First, if a sample does not represent the range of effects in the whole population, reliability will be underestimated (Fife et al., 2012; Franco-Martínez et al., 2023; Hunter et al., 2006). Since distractor suppression is such a robust effect, it would not be surprising if a lack of participants representing ranges of lower (or even null) suppression is making this variable more unreliable. Second, in the additional singleton task, previous authors have commonly dichotomized both variables involved in their analysis, losing all the ordinal information the task inherently provides (Hunter & Schmidt, 1990). Authors can gain some variability if they systematically manipulate all levels of distance from the distractor to the HP location, instead of only manipulating the binary variable of whether the distractor appears at the exact HP location or not. The same applies with the awareness measure: the distance from participants' response in the awareness test to the actual HP location is richer in variability than the binary variable of whether they correctly select the HP location or not. Third, another possible source for low empirical reliability would be, as Rouder et al. (2023a, 2023b) called it, excessive trial noise. In Vicente-Conesa et al.'s (2023) design, there are approximately 312 trials for the HP condition but only 168 for the LP condition, per participant. Consequently, this unbalanced design makes trial noise in the latter condition higher than in the former.

Conclusion

In sum, this work provides an illustration that data appearing to offer evidence in favor of a certain theory (i.e., that location probability learning is independent of awareness) may be more consistent with the opposite theory (i.e., that location probability learning and awareness share a common underlying process) if such data are contaminated with measurement and sampling noise. Future simulation studies should explore the influence of alternative design and statistical artifacts (i.e., range restriction, dichotomization, and imbalance) on statistical power. All these, together with the already explored sampling and measurement noise, could provide a wider picture on what obstacles we should avoid when studying individual differences in experimental psychology.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used GPT-4 in order to occasionally resolve English language doubts. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

CRediT authorship contribution statement

Alicia Franco-Martínez: Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Francisco Vicente-Conesa:** Resources. **David R. Shanks:** Writing – review & editing, Supervision, Methodology. **Miguel A. Vadillo:** Writing – review & editing, Supervision, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

AFM was supported by PhD fellowship PRE2021-097654 from the Spanish Ministry of Science and Innovation. FVC was supported by PhD fellowship PRE2018-085148 from Ministerio de Ciencia e Innovación. DRS was supported by grant ES/Y002482/1 from the UK Economic and Social Research Council. MAV was supported by grants PID2020-118583 GB-I00 and CNS2022-135346 from Agencia Estatal de Investigación / FEDER UE. Special acknowledgement is extended to Javier Revuelta for his assistance with the technical aspects of model estimation, to our colleagues and friends Ricardo Rey-Sáez and Ignacio Castillejo for their reflections and ideas, and to both reviewers for their helpful suggestions.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jml.2025.104621.

Data availability

The data and scripts used to reproduce the findings of this study are openly available in the Open Science Framework at osf.io/y8v74/.

References

Adams, O. J., & Gaspelin, N. (2020). Assessing introspective awareness of attention capture. *Attention, Perception, & Psychophysics, 82*, 1586–1598. https://doi.org/10.3758/s13414-019-01936-9

Anderson, B. A., & Kim, H. (2019). Test-retest reliability of value-driven attentional capture. *Behavior Research Methods, 51*(2), 720–726. https://doi.org/10.3758/s13428-018-1079-7

Arnon, I. (2020). Do current statistical learning tasks capture stable individual differences in children? An investigation of task reliability across modality. *Behavior Research Methods, 52*(1), 68–81. https://doi.org/10.3758/s13428-019-01205-5

Awh, E., Belopolsky, A. V., & Theeuwes, J. (2012). Top-down versus bottom-up attentional control: A failed theoretical dichotomy. *Trends in Cognitive Sciences, 16*(8), 437–443. https://doi.org/10.1016/j.tics.2012.06.010

Bacon, W. F., & Egeth, H. E. (1994). Overriding stimulus-driven attentional capture. *Perception & Psychophysics, 55*(5), 485–496. https://doi.org/10.3758/BF03205306

Balota, D. A., Yap, M. J., Cortese, M. J., & Watson, J. M. (2008). Beyond mean response latency: Response time distributional analyses of semantic priming. *Journal of Memory and Language, 59*(4), 495–523. https://doi.org/10.1016/j.jml.2007.10.004

Bogaerts, L., Siegelman, N., Ben-Porat, T., & Frost, R. (2018). Is the Hebb repetition task a reliable measure of individual differences in sequence learning? *Quarterly Journal of Experimental Psychology, 71*(4), 892–905. https://doi.org/10.1080/17470218.2017.1307432

Borsboom, D., Kievit, R. A., Cervone, D., & Hood, S. B. (2009). The two disciplines of scientific psychology, or: The disunity of psychology as a working hypothesis. In J. Valsiner, P. Molenaar, M. Lyra, & N. Chaudhary (Eds.), *Dynamic Process Methodology in the Social and Developmental Sciences* (pp. 67–97). New York, NY: Springer. https://doi.org/10.1007/978-0-387-95922-1_4

Champely, S. (2020). *pwr: Basic Functions for Power Analysis. R package version 1.3-0*. https://CRAN.R-project.org/package=pwr.

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

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

Colagiuri, B., & Livesey, E. J. (2016). Contextual cuing as a form of nonconscious learning: Theoretical and empirical analysis in large and very large samples. *Psychonomic Bulletin & Review, 23*, 1996–2009. https://doi.org/10.3758/s13423-016-1063-0

Covington, N. V., Brown-Schmidt, S., & Duff, M. C. (2018). The necessity of the hippocampus for statistical learning. *Journal of Cognitive Neuroscience, 30*(5), 680–697. https://doi.org/10.1162/jocn_a_01228

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12*(11), 671–684. https://doi.org/10.1037/h0043943

Di Caro, V., Theeuwes, J., & Della Libera, C. (2019). Suppression history of distractor location biases attentional and oculomotor control. *Visual Cognition, 27*(2), 142–157. https://doi.org/10.1080/13506285.2019.1617376

Draheim, C., Mashburn, C. A., Martin, J. D., & Engle, R. W. (2019). Reaction time in differential and developmental research: A review and commentary on the problems and alternatives. *Psychological Bulletin, 145*(5), 508–535. https://doi.org/10.1037/bul0000192

Duncan, D., & Theeuwes, J. (2020). Statistical learning in the absence of explicit top-down attention. *Cortex, 131*, 54–65. https://doi.org/10.1016/j.cortex.2020.07.006

Enkavi, A. Z., Eisenberg, I. W., Bissett, P. G., Mazza, G. L., MacKinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2019). Large-scale analysis of test–retest reliabilities of self-regulation measures. *Proceedings of the National Academy of Sciences, 116*(12), 5472–5477. https://doi.org/10.1073/pnas.1818430116

Erickson, L. C., Kaschak, M. P., Thiessen, E. D., & Berry, C. A. (2016). Individual differences in statistical learning: Conceptual and measurement issues. *Collabra: Psychology, 2*(1), 14. https://doi.org/10.1525/collabra.41

Failing, M., Wang, B., & Theeuwes, J. (2019). Spatial suppression due to statistical regularities is driven by distractor suppression not by target activation. *Attention, Perception, & Psychophysics, 81*, 1405–1414. https://doi.org/10.3758/s13414-019-01704-9

Ferrante, O., Patacca, A., Di Caro, V., Della Libera, C., Santandrea, E., & Chelazzi, L. (2018). Altering spatial priority maps via statistical learning of target selection and distractor filtering. *Cortex, 102*, 67–95. https://doi.org/10.1016/j.cortex.2017.09.027

Fife, D. A., Mendoza, J. L., & Terry, R. (2012). The assessment of reliability under range restriction: A comparison of α, ω, and test–retest reliability for dichotomous data. *Educational and Psychological Measurement, 72*(5), 862–888. https://doi.org/10.1177/0013164411430225

Franco-Martínez, A., Alvarado, J. M., & Sorrel, M. A. (2023). Range restriction affects factor analysis: Normality, estimation, fit, loadings, and reliability. *Educational and Psychological Measurement, 83*(2), 262–293. https://doi.org/10.1177/00131644221081867

Gao, Y., & Theeuwes, J. (2020). Learning to suppress a distractor is not affected by working memory load. *Psychonomic Bulletin & Review, 27*, 96–104. https://doi.org/10.3758/s13423-019-01679-6

Gao, Y., & Theeuwes, J. (2022). Learning to suppress a location does not depend on knowing which location. *Attention, Perception, & Psychophysics, 84*(4), 1087–1097. https://doi.org/10.3758/s13414-021-02404-z

Garre-Frutos, F., Vadillo, M. A., González, F., & Lupiáñez, J. (2024). On the reliability of value-modulated attentional capture: An online replication and multiverse analysis.

Behavior Research Methods, 56, 5986–6003. https://doi.org/10.3758/s13428-023-02329-5

Gaspelin, N., & Luck, S. J. (2018). The role of inhibition in avoiding distraction by salient stimuli. *Trends in Cognitive Sciences, 22*(1), 79–92. https://doi.org/10.1016/j.tics.2017.11.001

Geng, J. J., & Behrmann, M. (2005). Spatial probability as an attentional cue in visual search. *Perception & Psychophysics, 67*(7), 1252–1268. https://doi.org/10.3758/BF03193557

Giménez-Fernández, T., Luque, D., Shanks, D. R., & Vadillo, M. A. (2023). Rethinking attentional habits. *Current Directions in Psychological Science, 32*(6), 494–500. https://doi.org/10.1177/09637214231191976

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

Hernández-Gutiérrez, D., Sorrel, M. A., Shanks, D. R., & Vadillo, M. A. (2025). The conscious side of 'subliminal' linguistic priming: A systematic review with meta-analysis and reliability analysis of visibility measures. *Journal of Cognition, 8*(1), 13. https://doi.org/10.5334/joc.419

Hunter, J. E., & Schmidt, F. L. (1990). Dichotomization of continuous variables: The implications for meta-analysis. *Journal of Applied Psychology, 75*(3), 334. https://doi.org/10.1037/0021-9010.75.3.334

Hunter, J. E., & Schmidt, F. L. (2015). *Methods of meta-analysis: Correcting error and bias in research findings. Third edition.* Sage Publications, Ltd. https://doi.org/10.4135/9781483398105.

Hunter, J. E., Schmidt, F. L., & Le, H. (2006). Implications of direct and indirect range restriction for meta-analysis methods and findings. *Journal of Applied Psychology, 91* (3), 594. https://doi.org/10.1037/0021-9010.91.3.594

Jiang, Y. V. (2018). Habitual versus goal-driven attention. *Cortex, 102*, 107–120. https://doi.org/10.1016/j.cortex.2017.06.018

Jiang, Y. V., Sha, L. Z., & Sisk, C. A. (2018). Experience-guided attention: Uniform and implicit. *Attention, Perception, & Psychophysics, 80*, 1647–1653. https://doi.org/10.3758/s13414-018-1585-9

Jiang, Y. V., Swallow, K. M., Won, B. Y., Cistera, J. D., & Rosenbaum, G. M. (2015). Task specificity of attention training: The case of probability cuing. *Attention, Perception, & Psychophysics, 77*, 50–66. https://doi.org/10.3758/s13414-014-0747-7

Kalra, P. B., Gabrieli, J. D., & 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

Karr, J. E., Areshenkoff, C. N., Rast, P., Hofer, S. M., Iverson, G. L., & Garcia-Barrera, M. A. (2018). The unity and diversity of executive functions: A systematic review and re-analysis of latent variable studies. *Psychological Bulletin, 144*(11), 1147–1185. https://doi.org/10.1037/bul0000160

Kaufman, S. B., DeYoung, C. G., Gray, J. R., Jiménez, L., Brown, J., & Mackintosh, N. (2010). Implicit learning as an ability. *Cognition, 116*(3), 321–340. https://doi.org/10.1016/j.cognition.2010.05.011

Kim, A. J., Grégoire, L., & Anderson, B. A. (2025). Reliably measuring learning-dependent distractor suppression with eye tracking. *Behavior Research Methods, 57* (1), 1–9. https://doi.org/10.3758/s13428-024-02552-8

Krishnan, S., Carey, D., Dick, F., & Pearce, M. T. (2022). Effects of statistical learning in passive and active contexts on reproduction and recognition of auditory sequences. *Journal of Experimental Psychology: General, 151*(3), 555–577. https://doi.org/10.1037/xge0001091

Lakens, D. (2022). Sample size justification. *Collabra: Psychology, 8*(1), 33267. https://doi.org/10.1525/collabra.33267

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

Levy, R., & McNeish, D. (2023). Perspectives on Bayesian inference and their implications for data analysis. *Psychological Methods, 28*(3), 719. https://doi.org/10.1037/met0000443

Lien, M. C., Ruthruff, E., & Tolomeo, D. (2024). Evidence that proactive distractor suppression does not require attentional resources. *Psychonomic Bulletin & Review, 31* (3), 1376–1386. https://doi.org/10.3758/s13423-023-02422-y

Lin, R., Li, X., Wang, B., & Theeuwes, J. (2021). Spatial suppression due to statistical learning tracks the estimated spatial probability. *Attention, Perception, & Psychophysics, 83*, 283–291. https://doi.org/10.3758/s13414-020-02156-2

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). *Statistical theories of mental test scores.* Addison-Wesley.

Lu, F., Huang, C., Zhu, C., He, Y., Shu, D., & Liu, D. (2022). Exploring an online method of measuring implicit sequence-learning consciousness. *Experimental Brain Research, 240*(12), 3141–3152. https://doi.org/10.1007/s00221-022-06535-z

Luce, R. D. (1986). *Response times: Their role in inferring mental organization.* Oxford, UK: Oxford University Press.

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal, 7*(4), 308–313. https://doi.org/10.1093/comjnl/7.4.308

Paap, K. R., & Greenberg, Z. I. (2013). There is no coherent evidence for a bilingual advantage in executive processing. *Cognitive Psychology, 66*(2), 232–258. https://doi.org/10.1016/j.cogpsych.2012.12.002

Paap, K. R., & Sawi, O. (2016). The role of test-retest reliability in measuring individual and group differences in executive functioning. *Journal of Neuroscience Methods, 274*, 81–93. https://doi.org/10.1016/j.jneumeth.2016.10.002

Palmer, E. M., Horowitz, T. S., Torralba, A., & Wolfe, J. M. (2011). What are the shapes of response time distributions in visual search? *Journal of Experimental Psychology: Human Perception and Performance, 37*(1), 58. https://doi.org/10.1037/a0020747

R Core Team. (2023). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/

Rey-Mermet, A., Gade, M., Souza, A. S., von Bastian, C. C., & Oberauer, K. (2019). Is executive control related to working memory capacity and fluid intelligence? *Journal of Experimental Psychology: General, 148*(8), 1335–1372. https://doi.org/10.1037/xge0000593

Ross, D. A., Richler, J. J., & Gauthier, I. (2015). Reliability of composite-task measurements of holistic face processing. *Behavior Research Methods, 47*, 736–743. https://doi.org/10.3758/s13428-014-0497-4

Rothkirch, M., Shanks, D. R., & Hesselmann, G. (2022). The pervasive problem of post hoc data selection in studies on unconscious processing. *Experimental Psychology, 69*, 1–11. https://doi.org/10.1027/1618-3169/a000541

Rouder, J. N., Chávez De la Peña, A. F., Mehrvarz, M., & Vandekerckhove, J. (2023a). *On Cronbach's merger: Why experiments may not be suitable for measuring individual differences. PsyArXiv*. https://doi.org/10.31234/osf.io/8ktn6

Rouder, J. N., & Haaf, J. M. (2019). A psychometrics of individual differences in experimental tasks. *Psychonomic Bulletin & Review, 26*(2), 452–467. https://doi.org/10.3758/s13423-018-1558-y

Rouder, J. N., Kumar, A., & Haaf, J. M. (2023b). Why many studies of individual differences with inhibition tasks may not localize correlations. *Psychonomic Bulletin & Review, 30*(6), 2049–2066. https://doi.org/10.3758/s13423-023-02293-3

Sagan, C. (1977). *The Dragons of Eden: Speculations on the Evolution of Human Intelligence.* Random House.

Salvador, A., Berkovitch, L., Vinchier, 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

Siegelman, N., & Frost, R. (2015). Statistical learning as an individual ability: Theoretical perspectives and empirical evidence. *Journal of Memory and Language, 81*, 105–120. https://doi.org/10.1016/j.jml.2015.02.001

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

Soto, D., Mäntylä, T., & Silvanto, J. (2011). Working memory without consciousness. *Current Biology, 21*(22), R912–R913. https://doi.org/10.1016/j.cub.2011.09.049

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

Stilwell, B. T., Bahle, B., & Vecera, S. P. (2019). Feature-based statistical regularities of distractors modulate attentional capture. *Journal of Experimental Psychology: Human Perception and Performance, 45*(3), 419. https://doi.org/10.1037/xhp0000613

Theeuwes, J. (1992). Perceptual selectivity for color and form. *Perception & Psychophysics, 51*(6), 599–606. https://doi.org/10.3758/BF03211656

Turk-Browne, N. B., Jungé, J., & Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology: General, 134*(4), 552–564. https://doi.org/10.1037/0096-3445.134.4.552

Vadillo, M. A., Konstantinidis, E., & Shanks, D. R. (2016). Underpowered samples, false negatives, and unconscious learning. *Psychonomic Bulletin & Review, 23*, 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. https://doi.org/10.1037/xge0000632

Vadillo, M. A., Malejka, S., Lee, D. Y., Dienes, Z., & Shanks, D. R. (2022). Raising awareness about measurement error in research on unconscious mental processes. *Psychonomic Bulletin & Review, 29*, 1–23. https://doi.org/10.3758/s13423-021-01923-y

Vadillo, M. A., Malejka, S., & Shanks, D. (2024). Mapping the reliability multiverse of contextual cuing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Advance online publication. https://doi.org/10.1037/xlm0001410.

Van Moorselaar, D., & Theeuwes, J. (2021). Statistical distractor learning modulates perceptual sensitivity. *Journal of Vision, 21*(12), 3. https://doi.org/10.1167/jov.21.12.3

Van Moorselaar, D., & Theeuwes, J. (2022). Spatial suppression due to statistical regularities in a visual detection task. *Attention, Perception, & Psychophysics, 84*, 450–458. https://doi.org/10.3758/s13414-021-02330-0

Van Moorselaar, D., & Theeuwes, J. (2024). Transfer of statistical learning between tasks. *Journal of Experimental Psychology: Human Perception and Performance, 50*(7), 740–751. https://doi.org/10.1037/xhp0001216

Vicente-Conesa, F., Castillejo, I., & Vadillo, M. A. (2024). Working memory load does not interfere with distractor suppression in the additional singleton task. *Attention, Perception, & Psychophysics.*. https://doi.org/10.3758/s13414-024-02940-4

Vicente-Conesa, F., Giménez-Fernández, T., Luque, D., & Vadillo, M. A. (2023). Learning to suppress a distractor may not be unconscious. *Attention, Perception, & Psychophysics, 85*(3), 796–813. https://doi.org/10.3758/s13414-022-02608-x

Viviani, G., Visalli, A., Finos, L., Vallesi, A., & Ambrosini, E. (2024). A comparison between different variants of the spatial Stroop task: The influence of analytic flexibility on Stroop effect estimates and reliability. *Behavior Research Methods, 56* (2), 934–951. https://doi.org/10.3758/s13428-023-02091-8

Von Bastian, C. C., Blais, C., Brewer, G. A., Gyurkovics, M., Hedge, C., Kałamała, P., & Wiemers, E. A. (2020). *Advancing the understanding of individual differences in attentional control: Theoretical, methodological, and analytical considerations*. *PsyArxiv*. https://doi.org/10.31234/osf.io/x3b9k

Wang, B., & Theeuwes, J. (2018a). How to inhibit a distractor location? Statistical learning versus active, top-down suppression. *Attention, Perception, & Psychophysics, 80*, 860–870. https://doi.org/10.3758/s13414-018-1493-z

Wang, B., & Theeuwes, J. (2018b). Statistical regularities modulate attentional capture. *Journal of Experimental Psychology: Human Perception and Performance, 44*(1), 13–17. https://doi.org/10.1037/xhp0000472

Wang, B., & Theeuwes, J. (2018c). Statistical regularities modulate attentional capture independent of search strategy. *Attention, Perception, & Psychophysics, 80*, 1763–1774. https://doi.org/10.3758/s13414-018-1562-3

Wang, B., van Driel, J., Ort, E., & Theeuwes, J. (2019). Anticipatory distractor suppression elicited by statistical regularities in visual search. *Journal of Cognitive Neuroscience, 31*(10), 1535–1548. https://doi.org/10.1162/jocn_a_01433

Weinstein, Y. (2018). Mind-wandering, how do I measure thee with probes? Let me count the ways. *Behavior Research Methods, 50*, 642–661. https://doi.org/10.3758/s13428-017-0891-9

West, G., Vadillo, M. A., Shanks, D. R., & Hulme, C. (2018). The procedural learning deficit hypothesis of language learning disorders: We see some problems. *Developmental Science, 21*(2). https://doi.org/10.1111/desc.12552

Yaron, I., Zeevi, Y., Korisky, U., Marshall, W., & Mudrik, L. (2024). Progressing, not regressing: A possible solution to the problem of regression to the mean in unconscious processing studies. *Psychonomic Bulletin & Review, 31*(1), 49–64. https://doi.org/10.3758/s13423-023-02326-x