NEUROSCIENCE of Consciousness

Neuroscience of Consciousness, 2026, **2026(1)**, niaf046
DOI: https://doi.org/10.1093/nc/niaf046
Registered Report (final)

Replicating the unconscious working memory effect: a multisite Registered Report

Alicia Franco-Martínez[1], Ricardo Rey-Sáez[1], Jesús Adrián-Ventura[2], Pietro Amerio[3], Ana Baciero[4], Amine Bennis[3], Fredrik Bergström[5], Axel Cleeremans[3], Laura Contu[6], Roberto Dell'Acqua[7], Xinping Deng[8], Fatma Nur Dolu[9,10], Filippo Gambarota[11], Yi Gao[12], Francisco Garre-Frutos[13], Anna Grubert[14], Ana Hernando[2], José A. Hinojosa[4,15,16], Asaf Hoory[17], ZhiKun Hou[18], Shao-Min Hung[19], Aine Ito[20], Mikel Jimenez[21], Alexandra I. Kosachenko[22], Merve Kulaksız[9], Daryl Y. H. Lee[23], Elmas Merve Malas[9], Simone Malejka[24], Pedro R. Montoro[21], Liad Mudrik[17,25], Yuri G. Pavlov[26], Gabriele Pesimena[6], Antonio Prieto[21], Dobromir Rahnev[12], Lais Ringenberg[5], Alejandro Sandoval-Lentisco[1,27], Akira Sarodo[28], Maor Schreiber[17], Paola Sessa[7], Pablo Solana[13], Dmitrii A. Tarasov[22], Miriam Tortajada[27], Kai Xue[12], Ziqian Xue[29], Yunxuan Zheng[12], Merve Çinici[9], David R. Shanks[23,*], David Soto[30,*], Miguel A. Vadillo[1,*]

[1] Departamento de Psicología Básica, Facultad de Psicología, Universidad Autónoma de Madrid, C. Iván Pavlov, 6, 28049, Madrid, Spain
[2] Department of Psychology and Sociology, University of Zaragoza, Spain
[3] Center for Research in Cognition & Neurosciences (CRCN), ULB Institute of Neuroscience (UNI), Université libre de Bruxelles, Belgium
[4] Departamento de Psicología Experimental, Procesos Psicológicos y Logopedia, Universidad Complutense de Madrid, Spain
[5] CINEICC, Faculty of Psychology and Educational Sciences, University of Coimbra, Portugal
[6] School of Psychological Sciences, University of Bristol, United Kingdom
[7] University of Padua, Department of Developmental and Social Psychology and Padova Neuroscience Center, Italy
[8] Shien-Ming Wu School of Intelligent Engineering, Guangzhou International Campus, South China University of Technology, Guangzhou 511442, China
[9] Konya Food and Agriculture University, Faculty of Social and Human Sciences, Department of Psychology, Konya, Türkiye
[10] Bilkent University, Graduate School of Engineering and Science, Department of Neuroscience, Ankara, Türkiye
[11] University of Padua, Department of Developmental and Social Psychology, Italy
[12] School of Psychology, Georgia Institute of Technology, Atlanta, GA, USA
[13] Department of Experimental Psychology, University of Granada, Spain
[14] Department of Psychology, Durham University, United Kingdom
[15] Instituto Pluridisciplinar, Universidad Complutense de Madrid, Spain
[16] Centro de Investigación Nebrija en Cognición (CINC), Universidad de Nebrija, Madrid, Spain
[17] School of Psychological Sciences, Tel Aviv University, Israel
[18] State Key Laboratory of Cognitive Neuroscience and Learning, Department of Psychology, Beijing Normal University, Beijing, China
[19] Waseda Institute for Advanced Study, Waseda University, Japan
[20] Department of English, Linguistics and Theatre Studies, National University of Singapore, Singapore
[21] Departamento de Psicología Básica I, UNED, Spain
[22] Laboratory of Neurotechnology, Ural Federal University named after the first President of Russia B N Yeltsin, Ekaterinburg, Russia
[23] Division of Psychology and Language Sciences, University College London, United Kingdom
[24] Department of Psychology, University of Cologne, Germany
[25] Sagol School of Neuroscience, Tel Aviv University, Israel
[26] University of Tuebingen, Institute of Medical Psychology and Behavioral Neurobiology, Germany
[27] Departamento de Psicología Básica y Metodología, Universidad de Murcia
[28] Faculty of Science and Engineering, Waseda University, Japan
[29] School of Automation Science and Engineering,Guangzhou International Campus, South China University of Technology, Guangzhou 511442, China
[30] Basque Center on Cognition, Brain and Language, Paseo Mikeletegi 69, 2° 20009 Donostia-San Sebastián, Spain

Alicia Franco-Martínez is a PhD candidate at the Autonomous University of Madrid. She has published research in psychometrics and statistical modelling.
Ricardo Rey-Sáez is a PhD candidate at the Autonomous University of Madrid. His research focuses on statistical modelling and Bayesian psychometrics in experimental psychology.
David R. Shanks is Professor of Psychology at University College London. He studies learning, memory, and decision making, with a particular focus on the impact of unconscious processes on behaviour.
David Soto is Research Professor at the Basque Center on Cognition, Brain and Language. His work focuses on understanding the distinction between conscious and non-conscious processes during perceptual selection, learning and memory, language, and metacognition.
Miguel A. Vadillo is Associate Professor at the Autonomous University of Madrid. His research focuses on the role of implicit learning processes in visual attention and also on reproducibility problems in psychological research.

Received 27 May 2025; **revised** 10 September 2025; **accepted** 16 October 2025

*Corresponding authors: Miguel A. Vadillo, Departamento de Psicología Básica, Facultad de Psicología, Universidad Autónoma de Madrid, C. Iván Pavlov, 6, 28049, Madrid, Spain. E-mail: miguel.vadillo@uam.es; David Soto, Basque Center on Cognition, Brain and Language, Paseo Mikeletegi 69, 2° 20009 Donostia-San Sebastián, Spain. E-mail: d.soto@bcbl.eu; David R. Shanks, Division of Psychology and Language Sciences, University College London, Room 230, 26 Bedford Way, London, WC1H 0AP, United Kingdom. E-mail: d.shanks@ucl.ac.uk

Abstract

Although in recent years some studies have found evidence suggesting that working memory (WM) may operate on unconscious perceptual contents, decisive demonstrations of the existence of unconscious WM are lacking. In the present Registered Report, we replicate the first study on this topic by Soto et al. (Working memory without consciousness. *Curr Biol* 2011;**21**:R912–3.): a visual discrimination task asking participants to report the direction in which a subliminal Gabor grating was rotated after a 2-s delay. We acquired a multisite sample from 19 laboratories, with a larger number of participants ($N = 531$) and trials (720 in two sessions) than those typically used in previous studies. As a result, a large-sample, international, and open-access dataset is now available for researchers and future analyses. Furthermore, some minimal baseline requirements were guaranteed for the experimental task (i.e. number of valid trials, motivation, and consistent labels for the Perceptual Awareness Scale). The results showed (1) above-chance WM performance in cue-present trials reported as unseen (.55 accuracy), (2) a significant positive correlation between WM performance and cue detection sensitivity ($r = .228$), and (3) a significant above-chance intercept in the regression of performance on sensitivity ($\beta_0 = .521$). These findings suggest that WM can operate on unconscious representations, although it remains positively associated with perceptual sensitivity. Crucially, because measurement error could compromise the interpretation of these three results, we provide evidence for our measures' excellent reliability and, more fundamentally, for their validity.

Keywords: unconscious; working memory; perceptual awareness; replication; multisite; Registered Report

One of the biggest challenges in modern science is disentangling the mechanisms of consciousness. Conscious awareness can be defined as 'the process leading to experience and reportability' (Soto and Silvanto 2016, p. 520). A popular strategy to study its functions and physiological basis is to determine which psychological processes can operate outside of this conscious awareness. The current evidence suggests that processes including perception, attention, and memory—or at least some of their subcomponents—may operate on the basis of information that is not consciously represented (Francken et al. 2022). In recent years, an especially provocative assertion states that working memory (WM) can operate unconsciously, namely, for items reported as *unseen* (e.g. Soto et al. 2011; Bergström and Eriksson 2014, 2015, 2018; Trübutschek et al. 2017, 2019; Persuh et al. 2018).

WM is defined as a 'limited capacity system, which temporarily maintains and stores information, [and] supports human thought processes by providing an interface between perception, long-term memory and action' (Baddeley 2003, p. 829). In its first formulations, Baddeley and Hitch (1974) did not explicitly focus on the relation between WM and consciousness. Later, WM was assumed to operate consciously, through the *episodic buffer*, proposed as a component that mediates the interaction between other subcomponents (i.e. the *phonological loop* and *visuospatial sketchpad*), forming conscious episodes (Baddeley 2003). Although some authors had speculated that WM operations might be independent from awareness (e.g. Baars 1997; Hassin et al. 2009), it was not until Soto et al.'s (2011) work that this issue began to be empirically tested.

These authors devised a simple discrimination experiment in which participants were presented with a Gabor grating (masked and lasting 16.67 ms, present on half of the trials and absent on the other half) and retained its orientation in WM across a 2-s delay period, after which they were asked to report whether a test grating was rotated to the left or right with respect to the initial grating cue. Participants were also asked to report the visibility of the memory cue on each trial using a Perceptual Awareness Scale (PAS, Ramsøy and Overgaard 2004; Sandberg and Overgaard 2015; Sandberg et al. 2010). Soto et al. (2011) found that in those trials where participants reported absence of awareness (PAS = 1, from now on, *unseen* trials), they performed above chance in the WM task (.59, $p = .006$), suggesting that these cues had been successfully encoded and retrieved from WM. Also, Soto et al. found a near-zero correlation ($r = -.18$, $p = .41$) between performance in the WM task and cue detection sensitivity of the memory cue (calculated with a pseudo-d', see Stein et al. 2016 for a critical discussion on this issue; see also Soto and Silvanto 2016), suggesting that 'memory discrimination dissociated from perceptual awareness' (p. R913). Furthermore, a regression of performance on sensitivity (Greenwald et al. 1995, 1996) returned a significant above-chance intercept ($\beta_0 = .60$), suggesting that at least statistically, performance in the WM task was predicted to be positive even for participants who showed no evidence of awareness. These results lead logically to the conclusion that WM can operate on stimuli that have not been consciously experienced.[1]

Since the publication of this seminal study by Soto et al. (2011), many other studies have reported similar effects in the same task (Dutta et al. 2014; King et al. 2016) and in other tasks with alternative masking methods (Sklar et al. 2012; Bergström and Eriksson 2014, 2015; Pan et al. 2014; Trübutschek et al. 2019; Nakano and Ishihara 2020). Recently, a meta-analysis by Gambarota et al. (2022), collating data from all the empirical studies that have addressed this question so far (38 independent effect sizes), found a medium-sized unconscious WM effect, Hedges' $g_z = 0.54$, with a 89% highest probability density interval between 0.32 and 0.77.

The aim of the present Registered Report is to test whether WM can operate unconsciously by replicating the original study of Soto et al. (2011) with a highly powered multisite sample. In doing so, we have responded to the need for more systematic and larger-scale studies in research on unconscious WM, as urged by Gambarota et al. (2022) in their meta-analysis. Moreover, the empirical evidence gathered so far poses concerns that we attempted to address and overcome, following recent methodological recommendations (e.g. Shanks et al. 2021). Below, the reader can find a list of these concerns, each followed by our proposed solution, across three major aspects: research biases and statistical power, minimal baseline requirements (i.e. number

[1] For an earlier attempt to test this possibility, see Hassin et al. (2009), although this study did not involve masking stimuli from visual awareness. Also, the seminal studies by Soto et al. (2011) bear some superficial resemblance to subliminal priming studies, but note that priming and WM are different processes: while priming operates on delays of a few hundreds of milliseconds and requires no cognitive operation on the prime, the WM task's delay is 2 s and requires maintenance of the memory cue and comparison to the target for a delayed decision.

of valid trials, motivation, and PAS labels), and potential hazards with data analyses.

Research biases and statistical power

The meta-analysis by Gambarota et al. (2022) found signs of publication bias in this literature (funnel-plot asymmetry), but did not attempt to apply any bias-correction method. This means that although Gambarota et al. found a medium-sized meta-analytic effect, this estimate is possibly inflated by the selective publication of positive findings (for instance, studies with null results may have been discarded as failed pilot experiments), but we cannot know the scope of the problem with certainty. Assessing the impact of publication or other reporting biases in the literature is a daunting challenge because different correction methods often yield different, even contradictory, results and usually it is difficult to decide which method's assumptions are most valid for the experimental setting at issue (Carter et al. 2019). Even so, it is advisable to apply some bias correction to the data because otherwise, in the presence of bias, the meta-analytic estimate is necessarily inflated to an unknown extent.

We applied a state-of-the-art method, robust Bayesian meta-analysis (RoBMA; Bartoš et al. 2021), to the Gambarota et al. dataset. The advantage of this method over alternative approaches is that RoBMA fits not just one but a set of bias-correction methods and returns a weighted average that is sensitive to how well each model fits the data. In our reanalysis of Gambarota et al.'s data, using default settings, we found an inclusion Bayes factor of 3.22 in favour of models including publication bias, confirming the suspicion that this literature is biased by the selective publication of positive findings. After correcting for bias, the evidence in favour of a positive effect is no longer substantial, with a point estimate of $g_z = 0.16$, 95% credible interval (CI) [−0.09, 0.55], and only provides inconclusive evidence for the existence of such an effect (inclusion Bayes factor = 1.003). The R script for this analysis is available at osf.io/xzv9t/.

Given the signs of bias in previous research conducted in this area, we believe that future studies should take measures to minimize the impact of publication or reporting biases. One of the most effective approaches to achieve this is to preregister the experimental protocol and analysis plan before the experiments are conducted. This allows the reader to distinguish purely confirmatory analyses that follow the registered protocol from exploratory analyses conducted *post hoc* that might be subject to bias. In addition, to minimize the impact of publication bias, the decision to publish a study or not should be based on the quality of the methods and not on the (significant or non-significant) results. Registered Reports are an ideal means to achieve this because, here, research proposals are assessed by editors and reviewers before data collection begins and therefore the decision to approve or reject the publication of the study cannot be contaminated by the empirical results (Chambers and Tzavella 2022). Consequently, this work follows a Registered Report format, the first one on unconscious WM (see Stage 1 manuscript at osf.io/nz6m5).

Also, this multisite study has the goal of creating a robust and open-access dataset from at least 10 laboratories in different countries since none of the previous studies in this field have made their data publicly available. Having access to the datasets is another tool against publication and reporting biases as it allows the scientific community to verify the robustness of the analyses and results, for instance, allowing multiverse analyses to be undertaken (Steegen et al. 2016; Simonsohn et al. 2020). Finally, the scripts for the task and data analysis are also available in the Open Science Framework repository: osf.io/xzv9t.

It is well known that one of the chief limitations in research on unconscious processing is low statistical power (Vadillo et al. 2016, 2020; Shanks et al. 2021). In the literature studying the unconscious WM effect, the median sample size is 17.5 participants (Gambarota et al. 2022), with a maximum of 38 in Trübutschek et al. (2019) and a minimum of 7 in Soto et al. (2011). With 17 participants, the minimum detectable effect size with a power of .95 is 0.932[2] in d_z units. This effect size could be very reasonable in the experimental context, but our reanalysis of Gambarota et al.'s meta-analysis yielded a bias-corrected effect of 0.16 (in fact, compatible with no true effect), thus suggesting that much larger samples might be needed to detect a true but small effect. In the present study, we initially planned to recruit at least 10 samples of at least 20 participants in two sessions, yielding a minimum of 200 participants. In the end, we collected 19 samples which sum 531 participants in total.

Another source of low statistical power is the number of trials included in the analyses. These experiments are usually long, intermixing trials in different conditions (e.g. absent or present cues, with or without distractors), but when only the valid trials are taken into account (i.e. reported as *unseen* when the cue was *present*), the number is typically not very large. Soto et al. (2011) analysed a mean across participants of only 26 valid trials from a total of 96 in their Experiment 1 (and a mean of 40 valid trials in both Experiments 2 and 3 with distractors), while other authors employing similar tasks analysed even fewer, specifically a median of 16 valid trials from a median of 96 total trials (Bona et al. 2013; Dutta et al. 2014; King et al. 2016; Taglialatela Scafati 2019). Soto et al. divided the trials into 50% *cue-present* and 50% *cue-absent*, thus initially losing half of the trials for the main analysis. To increase the percentage of valid trials in our design, we divided ours into 70% *cue-present* and 25% *cue-absent* trials, and an additional 5% of *supraliminal* trials. As a consequence, we presented a total of 360 trials per participant per session (252 *cue-present*), from which we expected to obtain ∼126 valid (*cue-present* but *unseen*) trials. Given that we planned to conduct the task over two sessions, we expected that each participant would provide ∼252 valid trials out of a total of 720 trials.

Minimal baseline requirements

One of the aspects that Gambarota et al. (2022) considered was heterogeneity between laboratories, accentuated by the differences between the tasks applied in those laboratories. Those two sources of heterogeneity were not readily separable in their meta-analysis. However, in our design, we obtained a pure estimation of between-laboratory heterogeneity since all of them used the same experimental task. Furthermore, we took measures to minimize the impact of undesirable sources of heterogeneity across experimental settings, ensuring minimal baseline requirements with respect to the number of valid trials, motivation, and PAS labels.

First, Gambarota et al. (2022) pointed out the often-neglected problem of the varying number of valid trials between tasks (discussed above). Even using the exact same experimental task, differences in the number of valid trials can occur as a consequence of distinct screen luminances, as Taglialatela Scafati

[2] Calculated with the function `pwr.t.test()` of the `{pwr}` package in R (Champely 2020), for a two-tailed *t*-test, the test commonly applied when contrasting the unconscious WM effect (e.g. Soto et al. 2011).

(2019) warned in her replications of Soto et al. (2011). She experienced difficulties when masking the cues and could not achieve >27% of *unseen* trials with a masked cue, against the 55% in Soto et al. For this reason, we calibrated the contrast of the memory cue to each participant at the beginning of the experiment with two sequential QUESTs and during the task by means of an adaptive staircase. This procedure aimed to yield at least 50% of *unseen* trials per participant in the *cue-present* condition as implemented in Jachs et al. (2015) and recommended by Shanks (2017).

Second, as pointed out by Dutta et al. (2014) and Soto and Silvanto (2016), experimenters need to ensure that participants' motivation is maintained throughout the task, so that this does not become an additional source of irrelevant between-participant heterogeneity. Soto et al. (2011) included two trial conditions in which the masked cue was present for 16.67 ms or absent. In this design, the masked cue was difficult to see and thus motivation could decline quickly. Even though we used a calibration method, our pilot participants also reported this motivational decline since almost every trial was near the subliminal threshold (Supplemental Material). Consequently, we ensured that 5% of trials were *supraliminal*. This way, participants find the experiment less frustrating and those trials can serve as evidence of response quality (see Method). This decision aims to overcome the 'task-difficulty artefact', under which awareness measured in *cue-present* trials may be underestimated (Pratte and Rouder 2009).

The third source of heterogeneity refers to the labels used for the PAS across experimental settings. The PAS has become the preferred scale for measuring perceptual awareness in the recent literature (Overgaard et al. 2010; Sandberg et al. 2010; Skóra et al. 2021), but a crucial and often-neglected feature concerns the labels for its categories, which often vary between studies. For example, the PAS labels applied in Soto et al. (2011, from 1 = 'did not see anything', 2 = 'maybe saw something', 3 = 'saw the stimulus but not its orientation', to 4 = 'saw the stimulus and its orientation') importantly differ from those applied in Nakano and Ishihara (2020, from 1 = 'did not see anything', 2 = 'did not see the target but felt existence of something', 3 = 'saw the target but it was not clear', to 4 = 'saw the target clearly') and from those employed in Taglialatela Scafati (2019, from 1 = 'nothing', 2 = 'glimpse', 3 = 'something', to 4 = 'well').[3] For instance, these latter PAS labels do not differentiate between seeing the cue and its orientation and might explain why their results failed to reproduce the original effect. In order to avoid this heterogeneity, we fixed the PAS labels across laboratories. After piloting the original PAS from Soto et al., we decided to adapt its labels to the ones used in Wiens et al. (2023). These authors applied an almost identical task and their PAS differentiated between detection and identification thresholds. Our PAS labels were 1 = 'I did not see the Gabor', 2 = 'I saw something, but did not identify the orientation', 3 = 'I saw the orientation vaguely', and 4 = 'I saw the orientation clearly' (see more details in Supplemental Material). By wording the categories in this way, future researchers will be able to run analyses based on both thresholds since PAS = 2 implies detection of the cue and PAS = 3 and 4 imply identification of its orientation.

Potential hazards with data analyses

As explained earlier, the conclusion that WM can operate over unconscious input is based on support for three statistical hypotheses: (1) above-chance WM performance in *cue-present* but *unseen* (PAS = 1) trials, (2) a non-significant correlation between participants' cue detection sensitivity and WM performance, and (3) a significantly above-chance intercept in a regression of performance on sensitivity. Hypothesis (1) gauges the main evidence for WM operating with unconsciously represented information. Hypothesis (2) is commonly interpreted as providing evidence that WM processing is independent from detection sensitivity, i.e. that participants showing better cue sensitivity do not show better WM performance and *vice versa*. A positive correlation between cue detection sensitivity and WM performance does not necessarily indicate that all knowledge is conscious. However, evidence for the absence of this association would be in line with a non-conscious account of WM performance because it discounts any contribution from conscious processes in cue detection. Hypothesis (3) is usually taken as proof that even an ideal participant showing no detection sensitivity would show above-chance WM performance. Nevertheless, interpretation of these patterns is not immune to criticism and their problems are closely intertwined.

All three results described above (i.e. above-chance performance when PAS = 1, non-significant correlation, and a significantly above-chance intercept) can serve as evidence of the effect, but *only* if we assume that the awareness measure is sufficiently reliable and that there is no measurement error in PAS responses or in the sensitivity indices computed from them (Shanks 2017; Shanks et al. 2021). The first analysis requires including only PAS = 1 trials, a method known as *post hoc selection*, under the logic of only analysing *unseen* trials. By doing this, we assume that every PAS = 1 trial was truly *unseen* and that no truly seen trial is (mis)reported as PAS = 1. Since no measures in the empirical world are perfectly reliable and the reliability of awareness tests is often quite modest (Rothkirch et al. 2022; Vadillo et al. 2022), awareness may have been underestimated in previous studies by *regression to the mean* (Vadillo et al. 2016; Shanks 2017): as a result of random error, some stimuli for which a degree of awareness is present will be mistakenly reported as *unseen*, thus contaminating the latter. In the same vein, measurement error can attenuate the observed correlation between detection sensitivity and WM performance, creating an illusion of a null correlation (Rouder and Haaf 2019; Malejka et al. 2021; Rouder et al. 2023). Also, it can induce a spurious intercept in the regression of sensitivity on performance, creating an illusion of significant performance when cue detection sensitivity is zero (Greenwald and Draine 1997; Klauer et al. 1998; Miller 2000). In brief, all three hypotheses taken as diagnostic of unconscious WM are potentially threatened by the presence of measurement error (Shanks et al. 2021).

In this field, it is not common practice to report an estimation of the reliability of the measures; for instance, none of the studies in Gambarota et al.'s (2022) meta-analysis reported reliability estimates. Also, *post hoc selection* is applied not only in most of the studies on unconscious WM (including all of the studies in Gambarota et al.'s meta-analysis), but also in most of the experimental literature that uses the PAS or similar measures of awareness (e.g. Sklar et al. 2012; Biderman and Mudrik 2018; Rowe et al. 2020). Thus, following Soto and Silvanto (2016) and Shanks et al. (2021), we explored if this problem of unreliability is sufficiently large to adversely affect the interpretation of our data. Specifically, we estimated reliability for the sensitivity measure

[3] In both Soto et al.'s (2011) and Nakano and Ishihara's (2020) PASs, category 1 measures the same degree of awareness. However, category 2 in Soto et al.'s scale is more similar to category 3 in Nakano and Ishihara's (the latter did not differentiate between seeing the stimulus and its orientation). Furthermore, the nuance captured in category 2 of Nakano and Ishihara's PAS, which they related to type-2 blindsight (and considered as 'unconscious experience' together with category 1), is not explicitly represented in the PAS of Soto et al. This detail may be key since Nakano and Ishihara did not find above-chance performance when discarding category 2 from their analyses, suggesting that a minimum level of conscious awareness is required for the WM effect.

d′ and performance in the WM task, calculating the permutated split-half reliability coefficient (Kahveci et al. 2022).

Method

Participants

This study consists of a multisite international collaboration that includes samples from 19 different laboratories (from 13 countries), each one contributing data from at least 20 valid participants, each of whom completed the task twice in two different sessions (run at different days separated by no more than 17 days and taking place at similar times of day). Of the 617 participants initially recruited, we ultimately collected complete datasets from 531 (median age: 21 years; 391 women, 136 men, and 4 classified as non-binary or 'other gender'). In the preregistered Stage 1 protocol, we determined a sample size of *at least* 20 participants per laboratory and *at least* 10 laboratories for reasons of affordability since the median sample size in previous research studying this effect was 17.5 (Gambarota et al. 2022). In order to study the adequacy of the number of trials, samples, and laboratories, we ran sensitivity analyses for all three hypotheses and concluded that the initial sample size planned (200 participants) afforded sufficient power to detect effects below the sizes often found in the literature, i.e. 54.2% accuracy in the WM task, a correlation of .23, and an intercept of .233 in d_z units. Given that we finally collected more than twice the planned sample size, the minimum detectable effect sizes are even lower. For a detailed report of these sensitivity analyses, see the Supplemental Material, accessible at osf.io/xzv9t.

The samples are mainly composed of Psychology students from different universities and research centres (Table 1). Participants were not included if they had already participated in similar experiments or if they reported any non-corrected visual or neurological disorder. Other participants were excluded because of computer problems, they did not attend the second session, or other technical reasons. All detailed reasons for exclusion are registered in the participants register uploaded at osf.io/xzv9t. Experiment instructions and materials were translated into 11 languages (Table 1) following a back-translation process. Each laboratory was free to reward participants either with course credits or with money. The Autonomous University of Madrid (UAM) obtained ethical approval for this study (Reference CEI-130-2689). Other laboratories were invited to rely on our approval or seek approval in their own institutions as required by local regulations.

Apparatus

The experiment ran on computers and monitors whose attributes are reported by each laboratory and detailed at osf.io/xzv9t. The screen had a grey background and the visual angles of the stimuli were standardized across laboratories. Laboratories followed the instructions available at osf.io/xzv9t. The experiment was programmed and executed using PsychoPy v.2022.2.3 (Peirce et al. 2019, the original experiment by Soto et al. 2011, was programmed in E-Prime). The scripts for the task in the 11 languages are available at osf.io/xzv9t.

Procedure and design

The protocol followed a similar design and procedure as that of Soto et al. (2011) and lasted from 1 to 1.5 h in each session. In both sessions, the experiment started with welcome and instruction screens, followed by two QUEST calibration tests (Watson and Pelli 1983) performed on each participant to estimate the contrast of the Gabor grating associated with a pThreshold of .55. The contrast estimated in these initial QUESTs served as the starting value of the following staircase calibration. During the experimental trials, the contrast of the Gabor dynamically varied across trials using an interleaved up-down staircase based on the participant's previous PAS rating, in order to maintain the desired distribution of responses across the different PAS visibility ratings. For more details regarding calibration, see the Supplemental Material.

In the WM task, participants attempted to retain the orientation of a Gabor that was *present* on 70% (with calibration), *absent* on 25%, and *supraliminal* on a further 5% (without calibration) of the trials. Participants received the instructions contained in the Supplemental Material. Figure 1 illustrates the trial structure. Before starting the experimental task, they completed two blocks of training trials, 24 trials in each block in the first session and only six trials in the second session. The first block consisted of supraliminal trials in which the cues were presented for 200 ms and performance feedback was provided to participants. After 12 trials of this training, participants were informed of their accuracy in the WM task and asked if they wanted to repeat another 12 trials more to ensure their competence in the task. The second training block consisted of trials in which half of them were supraliminal as above and the other half were briefly presented for 16.67 ms. Each training and experimental trial started with a 1-s fixation cross in the middle of the screen and a 500-ms blank screen. Then, the memory cue was presented (only in *cue-present* trials) for 16.67 ms and consisted of a Gabor grating with varying contrast (according to the staircase), spatial frequency of 1.5 cycle/deg, and diameter of 3.8 deg of visual angle from a viewing distance of 57 cm. In the *supraliminal* trials, this Gabor was presented for 200 ms with a constant contrast of 0.01. The cue's orientation was randomly chosen (10, 40, 70, 100, 130, or 160 deg from the vertical) in each trial and masked by a backward pattern of random dots the same size as the cue. The mask was generated as an 800 × 800 random noise texture displayed within a circular aperture of 3.8-deg diameter and was presented for 200 ms. A blank screen was presented for 2 s as a delay period to retain the orientation of the memory cue. After this, the test cue was presented for 200 ms and consisted of another Gabor grating (constant contrast of 0.01) randomly rotated 30 deg either to the right or left compared to the memory cue's orientation. The WM task for the participants was to detect the direction in which the test cue was rotated by clicking the left or right arrow on the keyboard, respectively (i.e. 2-alternative-forced-choice task). They were encouraged to try their best in each trial, even if they did not consciously see the cue. Except during the first training block, no feedback about their task performance was provided. In both sessions, trials appeared in a pseudo-random order (i.e. shuffling the permutations between all conditions) in 15 blocks of 24 trials each, and the inter-trial interval lasted 1 s.

Finally, participants reported their subjective awareness of the memory cue, by responding to a 4-point PAS in each trial (from 1 = 'I did not see the Gabor', 2 = 'I saw something, but did not identify the orientation', 3 = 'I saw the orientation vaguely', to 4 = 'I saw the orientation clearly', labels and stem shown on screen trial-by-trial), pressing the keyboard numbers using the little, ring, middle, and index fingers of their left hand, to standardize the participants' response strategy. At the beginning of the experiment, we presented six example trials to illustrate the perceptual experience of a *clearly visible* (PAS = 4) and an *unseen* (PAS = 1)

Table 1. Names of the collaborating laboratories, their number of valid participants, and the language used in the experiment

Laboratory ID	Name of the centre	Number of valid participants	Language
1	Universidad Autónoma de Madrid	25	Spanish
2	Universidad de Granada	20	Spanish
3	Universidad Nacional de Educación a Distancia—Universidad Complutense de Madrid	47	Spanish
4	Ural Federal University	42	Russian
5	University of Bristol	21	English
6	Durham University	21	English
7	Georgia Institute of Technology	40	English
8	Universidad de Zaragoza	20	Spanish
9	National University of Singapore	24	English
10	Tel Aviv University	20	Hebrew
11	Konya Food and Agriculture University	30	Turkish
12	Universidad de Murcia	28	Spanish
13	Université libre de Bruxelles	30	French
14	University of Cologne	21	German
15	Waseda University	20	Japanese
16	University College London	28	English
17	University of Padova	31	Italian
18	University of Coimbra	30	Portuguese
19	South China University of Technology	33	Chinese

Note. Due to the amount of ongoing workload, the Basque Center on Cognition, Brain and Language was unable to collect the data within the scheduled time frame.

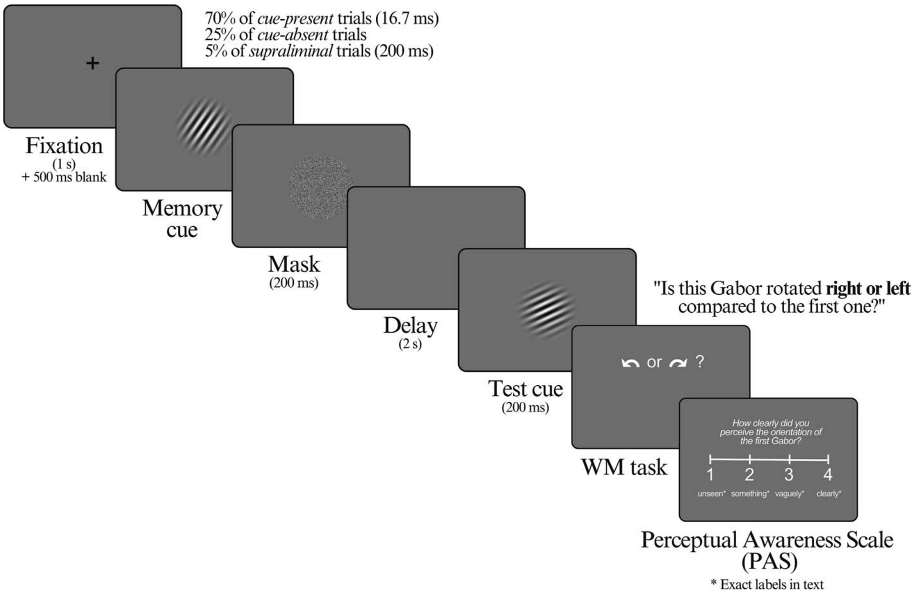

70% of *cue-present* trials (16.7 ms)
25% of *cue-absent* trials
5% of *supraliminal* trials (200 ms)
+
Fixation
(1 s)
+ 500 ms blank
Memory cue
Mask
(200 ms)
Delay
(2 s)
Test cue
(200 ms)
"Is this Gabor rotated **right or left** compared to the first one?"
or ?
WM task
How clearly did you perceive the orientation of the first Gabor?
1
2
3
4
unseen*
something*
vaguely*
clearly*
Perceptual Awareness Scale (PAS)
* Exact labels in text

Figure 1. Schematic of the trial procedure.

Gabor presentation.[4] From the total of 720 experimental trials across both sessions, the memory cue was either *present* (16.67 ms) in 504, *absent* (null contrast) in another 180, and *supraliminal* (200 ms) in 36 trials. Additionally, at the end of the second session, participants responded to the 20-item experiential subscale of the Rational Experiential Inventory (REI, Pacini and Epstein 1999), along with demographic information (age and gender), administered *via* Qualtrics.

Pilot study

Before collecting data from different laboratories, we ran a pilot study of the task on a sample of 31 participants at UAM. The goal of this pilot study was to detect any issues with the task and to check the functioning of the calibration algorithm. Pilot details and results can be checked in the Supplemental Material and pilot data are accessible at osf.io/xzv9t.

Results

Data pre-processing

Before analysing the data, we checked that no participant met the preregistered exclusion criteria: responses in the WM task with no variance (always pressing either the 'right' or 'left' button, as in Nakano and Ishihara 2020) and proportion of PAS = 1 responses

[4] The first three PAS = 4 examples illustrate to participants the perceptual experience of a *clearly visible* trial, presenting a Gabor for 200 ms at 0.1 contrast, for which only key '4' was available to press. The next three PAS = 1 examples illustrate to participants the perceptual experience of an *unseen* trial, explicitly informing them that no Gabor would be presented, for which only key '1' was available to press.

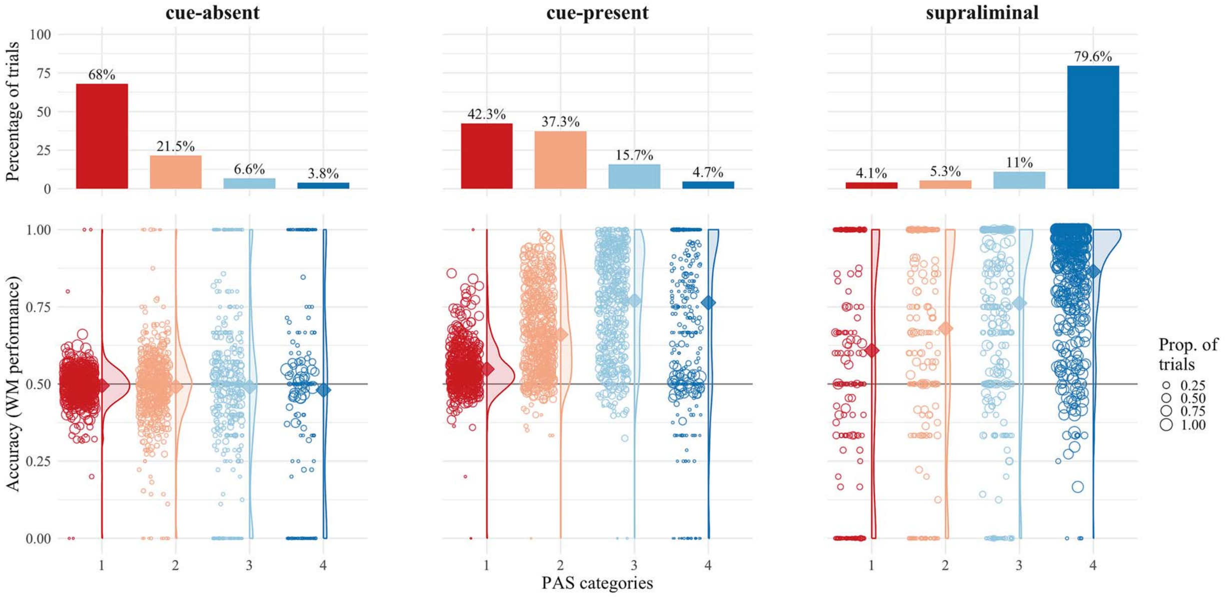


Figure 2. Top panel: Percentage of trials in each Perceptual Awareness Scale (PAS) category across the three cue conditions (cue-absent, cue-present, and supraliminal). Bottom panel: Working memory (WM) performance for each PAS category across cue conditions. Each point represents a participant; point size reflects the proportion of trials in that PAS category relative to all trials within the same cue condition. Diamonds in the violin plots represent the group mean. *Note.* Accuracy in *cue-absent* trials cannot be strictly interpreted as working memory (WM) performance since participants were not actually shown the first Gabor patch. However, stimuli were generated on these trials (though not displayed, i.e. contrast = 0). Accordingly, the y-axis represents *simulated* accuracy—i.e. the match between participants' responses and the generated orientation.

on *cue-absent* trials ≥3 SDs below the group mean (similar as in Dutta et al. 2014). There were no missing data in either test since answering was a condition for the experiment to continue. All pre-processing and analyses in this work were performed using the R language (version 4.4.0; R Core Team 2024) and are publicly available at osf.io/xzv9t.

Descriptive data

WM task performance and the proportions of PAS responses reported in the three conditions (*cue-present*, *cue-absent*, and *supraliminal*), both for the whole dataset (Fig. 2) and separately for each session are presented in Table 2. In the top panel of Fig. 2, the left- and right-most PAS response distributions confirm that the PAS labels were generally understandable: on average, participants reported a majority of PAS = 1 ratings ('I did not see the Gabor') in *cue-absent* trials, and a majority of PAS = 4 ratings ('I saw the orientation clearly') in *supraliminal* trials. Additionally, participants reported PAS = 1 in an average of 42.3% of the *cue-present* trials[5] and correctly identified the rotation of the Gabor in 88.7% of the *supraliminal* trials reported as PAS = 4, suggesting that, on average, they understood the WM task well. Moreover, these patterns are more prominent (and WM performance is consistently higher) in the second session (Table 2) suggesting that participants not only maintained compliance with the task but may have even improved due to increased familiarity or practice.

In the bottom panel of Fig. 2, the size of the points reflects the proportion of trials that each participant contributed within their respective cue condition. This allows us to identify participants with aberrant patterns. For instance, in *cue-present* and PAS = 1 trials, extreme WM performance, close to either 0 or 1 accuracy, corresponds to participants who rarely reported PAS = 1 ratings. Meanwhile, participants with a large proportion of PAS = 4 responses in *cue-present* trials (where a low proportion is expected because of the staircase procedure) exhibit near-chance WM performance, suggesting that these trials were essentially unseen, but PAS was not used appropriately. Performance in *cue-absent* trials was obtained as a manipulation check, for two reasons: first, to show that performance is indeed at chance when no cue is presented; and second, to illustrate how response variability increases as the number of trials decreases, gradually across PAS ratings. To do so, we relied on the fact that the script generated a grating in these trials, which was then presented at zero luminosity.

Preregistered inferential analyses

Two distinct questions were addressed: whether unconscious WM occurs or not and to what extent this effect is actually unconscious. The first question was addressed by testing (1) whether WM performance in *cue-present* but *unseen* trials was significantly above chance, removing PAS $\neq$ 1 trials. The second question was addressed by testing (2) if the correlation between participants' cue detection sensitivity and WM performance on *unseen cue-present* trials was significantly greater than zero, and (3) if the intercept in a regression of performance on sensitivity was significantly >.50. Sensitivity analyses ensuring that we had adequate statistical power for all tests are available in the Supplemental Material.

For test (1), we applied a three-level mixed logistic model, which accounted for the binary responses in the WM task (Level 1), as well as the heterogeneity between participants (Level 2) and between laboratories (Level 3), and determined whether the fixed intercept is significantly greater than zero (equivalent to an accuracy of .50). This model was fitted only with *cue-present* and *unseen* trials. In our most demanding simulations (with heterogeneity between subjects and between laboratories), we

[5] Although we used a staircase algorithm to calibrate the contrast of these trials so that 50% of them were *unseen*, some bias exists towards reporting PAS = 2 ('I saw something, but did not identify the orientation') when no cue was presented. For this reason, we could not obtain exactly 50% PAS = 1 responses in *cue-present* trials. This can be explained by some participants being too conservative in their PAS = 1 responses, by the label of PAS = 2 not being as precise as it should be, or by a potential postimage effect.

Table 2. Working memory (WM) performance means and standard deviations (SD) and relative frequencies of Perceptual Awareness Scale (PAS) responses in each category, across cue conditions (cue-absent, cue-present, and supraliminal), for both sessions and in each session separately

Cue condition	Session	WM performance Mean (SD)	Distribution of PAS responses			
			PAS = 1	PAS = 2	PAS = 3	PAS = 4
cue-absent	Both	.496 (0.500)	68%	21.5%	6.6%	3.8%
	Session 1	.498 (0.500)	67.5%	21.7%	6.7%	4.2%
	Session 2	.494 (0.500)	68.6%	21.2%	6.6%	3.5%
cue-present	Both	.626 (0.484)	42.3%	37.3%	15.7%	4.7%
	Session 1	.616 (0.486)	41.9%	38.2%	15%	5%
	Session 2	.635 (0.481)	42.6%	36.5%	16.3%	4.5%
supraliminal	Both	.848 (0.359)	4.1%	5.3%	11%	79.6%
	Session 1	.828 (0.377)	4.1%	5.3%	12.9%	77.7%
	Session 2	.868 (0.338)	4.1%	5.2%	9.2%	81.5%

had estimated that with $N = 200$ we would have >95% power to detect as significant a proportion of .542, whereas when assuming no between-laboratory heterogeneity, the minimum detectable proportion would be .517 (Supplemental Material). After estimating the model with $N = 525$ (the remaining six participants had no *unseen cue-present* trials), we obtained a significant positive fixed intercept (0.211 in log-odds, $Z = 15.66$, one-tailed $p < .001$, 95% CI [0.189, ∞]), equivalent to an accuracy of .552 (one-tailed 95% CI [.547, ≤ 1]), with a large between-participant variability (SD = 0.271), but zero between-laboratory variance. The latter means that there are no detectable differences across laboratories, thereby supporting the generalizability of our results.

To answer the second question and test hypotheses (2) and (3), we computed the cue detection sensitivity (measured with d') and WM performance for each participant on *unseen cue-present* trials (i.e. the accuracy on reporting the rotation of the Gabor in the test cue). Standard d' ($M_{d'} = 0.814$, 95% CI [0.765, 0.862][6]) was computed using the hit rate (P(H), i.e. the proportion of *seen*—PAS $\neq$ 1 responses—on *cue-present* trials) and the false-alarm rate (P(FA), i.e. the proportion of *seen*—PAS $\neq$ 1 responses—on *cue-absent* trials):[7]

$$d' = Z_{P(H)} - Z_{P(FA)} = Z_{P(PAS \neq 1 \mid present)} - Z_{P(PAS \neq 1 \mid absent)},$$

where Z is the standard score associated with the corresponding proportion (Stanislaw and Todorov 1999). First, we tested (2) whether the Pearson's correlation coefficient between these variables was equal to or greater than zero (one-tailed t-test with $\alpha = \beta = .05$) and obtained a significant positive correlation of $r = .228$ ($t_{(523)} = 5.358$, one-tailed $p < .001$, 95% CI [.159, ≤ 1]). Second, we tested (3) whether the intercept in a simple linear regression was $\geq .50$ in proportion units (one-tailed t-test with $\alpha = \beta = .05$) and obtained a significant above-chance intercept of $\beta_0 = .521$ ($t_{(523)} = 3.343$, one-tailed $P < .001$, 95% CI [.510, ≤ 1]). Figure 3 shows the scatter plot of this regression model.[8]

Reliability analyses

As explained in the Introduction, the analyses described in the previous section might be biassed by measurement error, thus affecting the validity of the inferences. To address this concern, we estimated the reliability for the two dependent measures used in the previous analyses (i.e. d' and WM performance[9]), using the permutated split-half reliability coefficient. This coefficient has shown a satisfactory performance in Kahveci et al. (2022), better than alpha or single split-half coefficients. Permutated split-half reliability is computed by averaging multiple correlations between random pairs of equal-sized splits, generated without replacement, and corrected with the Spearman–Brown formula. As recommended by Kahveci et al. (2022), considering our sample size we generated 1500 iterations with the function `rapidsplithalf()` from the R package `{rapidsplithalf}` (Kahveci 2025). No specific hypotheses were tested with the reliability estimates as preregistered analyses. For d', we obtained a split-half reliability of .94 with 95% CI [.94, .95] (.90 in the first session and .92 in the second one). However, for WM performance we could not compute the split-half reliability because there were some participants with only one valid trial (*unseen cue-present*). To handle this issue, we removed the nine participants with fewer than two valid trials in at least one of the two sessions. This selection did not affect the split-half reliability of d' meaningfully. The reliability of WM performance amongst the remaining participants was .73 with 95% CI [.65, .78] (.47 in the first session and .50 in the second).

Although this analysis was not preregistered in the Stage 1 protocol, we also assessed the test–retest reliability of both dependent variables across both sessions. The resulting estimates were .678 for d' and .463 for WM performance in valid trials (Fig. 4). Perhaps paradoxically, these estimations could be also contaminated by measurement error in each session. After correcting this with Bayesian hierarchical modelling, test–retest reliability increased to .745 for d' and .692 for WM performance (Supplemental Material, p. 18).

Deviations from the preregistration

All the analyses presented in the previous sections followed the preregistration in the Stage 1 manuscript, with the only exception

[6] Note that inferences about d' values were not preregistered, but the fact that the mean d' was significantly above zero ($t_{(524)} = 32.934$, one-tailed $P < .001$) might be informative for the reader.

[7] If any participant has a proportion of hits or false alarms equal to 0 or 1, we applied the log–linear correction method described in Hautus (1995). This correction consists of adding 0.5 to each cell in the SDT's contingency matrix, so that no hit or false-alarm proportion can result in 0 or 1.

[8] Although the preregistered protocol did not include any measure to ameliorate the impact of potential outliers, we nevertheless evaluated the presence of atypical or influential cases with hat values, Cook's distance, Mahalanobis distance, and DFBETAs. Even when all 49 influential participants are excluded from analyses, both the correlation and the intercept remain significant.

[9] Note that d' was calculated with all trials except for *supraliminal* ones, while WM performance includes only *unseen cue-present* trials.

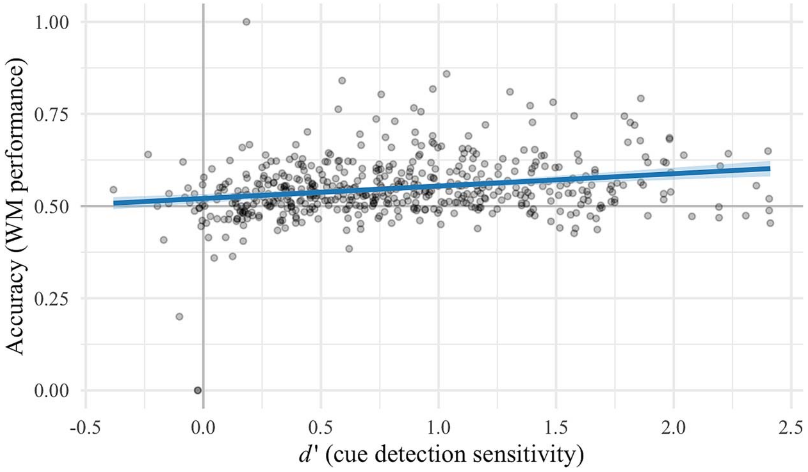


Figure 3. Scatter plot of a linear regression of working memory (WM) performance in cue-present and unseen trials on cue detection sensitivity (d'). *Note*. The shadowed area around the regression line represents the confidence interval.

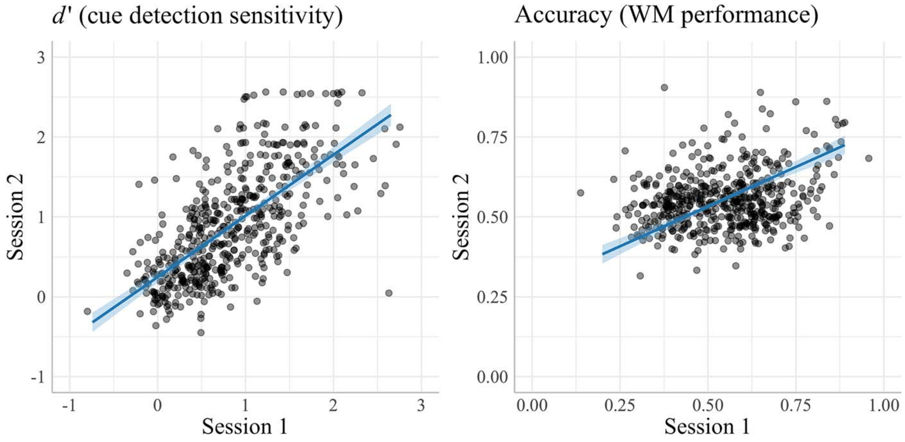


Figure 4. Scatter plots of test–retest reliability for cue detection sensitivity (d') and working memory (WM) performance across sessions.

that we decided to change the R package for estimating the permutated split-half reliability because the recently updated package {rapidsplithalf} is significantly quicker and more suitable for our variables.

Exploratory analyses

Validity analyses

The previous analyses suggest that the reliability of the two crucial dependent variables, d' and WM performance, is excellent, compared to the often poor reliabilities of behavioural measures (Rothkirch et al. 2022; Vadillo et al. 2022, 2024; Garre-Frutos et al. 2024; Yaron et al. 2024; Hernández-Gutiérrez et al. 2025). However, reliability is not all that matters in measurement; in fact, it is only one piece of the broader puzzle of validity (Kerschbaumer et al. 2025). Validity includes all aspects that guarantee a correct measurement process (i.e. the interaction between the participant and the task). For instance, did participants who barely reported *unseen* in *cue-present* and/or in *cue-absent* trials understand how to properly use the PAS?[10] Leaving aside the several reasons that could explain a low prevalence of PAS = 1 responses (e.g. conservative response bias or careless responding, amongst others), having a low number of valid trials clearly undermines the estimations. To explore this, we obtained both the permutated split-half and test–retest reliabilities, as well as the three preregistered hypotheses' estimates, using different exclusion criteria, based on varying the minimum required number of *unseen* trials (Table 3). As the exclusion criterion becomes more stringent (i.e. more participants without the minimum of *unseen* trials are excluded), the reliability of d' slightly decreases, but in contrast the reliability of WM performance improves. Meanwhile, inferences pertaining to all three hypotheses remain unchanged (all p values < .001). Notably, the unconscious WM effect (H1) slightly increases, the correlation (H2) decreases, and the intercept (H3) rises. This trend suggests that, as data validity improves, the results align more closely with those originally reported by Soto et al. (2011).

Correcting measurement error with Bayesian hierarchical modelling

Once researchers estimate the reliability of their measures, they can obtain corrected estimates of their statistics. These *post hoc* corrections are typically performed within the *Errors-in-Variables* framework (Carroll et al. 2006; Fuller 1987; Hunter and Schmidt 2015; Spearman 1904), which explicitly adjusts for bias in parameter estimates resulting from measurement error. Here, we follow this logic using a variant of the Bayesian hierarchical model proposed by Behseta et al. (2009). Matzke et al. (2017) showed that this model is particularly suitable when the 'observed variables' are not direct observations, but estimated parameters subject

[10] Note that all experimental stimuli in *cue-present* trials were calibrated trial-by-trial to obtain ∼50% of *unseen* responses. Not reporting PAS = 1 or doing so in a low number of *cue-present* trials implies atypical behaviour. This is particularly problematic in *cue-absent* trials, where all trials should ideally be reported as PAS = 1.

Table 3. Reliability (permutated split-half and test–retest coefficients) for d' and working memory (WM) performance and estimates for the three preregistered hypotheses, as a function of different exclusion criteria

Exclusion criteria	**N excluded**	**Reliability of d'**		**Reliability of WM performance**		**Hypotheses** **H1: unconscious WM** **H2: correlation** **H3: intercept**
		Split-half	**Test–retest**	**Split-half**	**Test–retest**	
None (excluded only if 0 valid trials)	6	.945 .90 in S1 .92 in S2	.674	NA (Ps with 1 valid trial)		**H1:** 55.25% **H2:** .228 **H3:** .5208
<2 *unseen* **only** in *cue-present* trials	15	.945 .90 in S1 .92 in S2	.678	.730 .47 in S1 .50 in S2	.463	**H1:** 55.24% **H2:** .220 **H3:** .5269
<10 *unseen* **only** in *cue-present* trials	25	.945 .90 in S1 .92 in S2	.675	.766 .60 in S1 .67 in S2	.482	**H1:** 55.28% **H2:** .218 **H3:** .5273
<20 *unseen* **only** in *cue-present* trials	34	.944 .90 in S1 .92 in S2	.673	.777 .63 in S1 .69 in S2	.488	**H1:** 55.31% **H2:** .216 **H3:** .5273
<2 *unseen* in **either** *cue-present* **and/or** *cue-absent* trials	18	.945 .90 in S1 .92 in S2	.674	.751 .57 in S1 .58 in S2	.466	**H1:** 55.24% **H2:** .224 **H3:** .5263
<10 *unseen* in **either** *cue-present* **and/or** *cue-absent* trials	37	.943 .90 in S1 .92 in S2	.664	.780 .63 in S1 .70 in S2	.488	**H1:** 55.32% **H2:** .216 **H3:** .5271
<20 *unseen* in **either** *cue-present* **and/or** *cue-absent* trials	62	.940 .90 in S1 .92 in S2)	.660	.791 .64 in S1 .72 in S2	.493	**H1:** 55.54% **H2:** .186 **H3:** .5315

Note. The exclusion criteria were based on the number of trials in *at least* one of the sessions being lower than the criterion. The minimum number of *unseen* trials was first applied **only** to *cue-present* trials (rows 2, 3, and 4) and later to **either** *cue-present* **and/or** *cue-absent* trials (last three rows). H1: unconscious working memory (WM) effect; H2: correlation between WM performance and cue detection sensitivity; H3: intercept in a regression of WM performance on sensitivity.

to uncertainty, like our cue detection sensitivity (d') and WM performance for each participant.[11] Our "Diamond" model (to distinguish it from the "Spade" model presented in the Supplemental Material) simultaneously provides corrected estimates of both the correlation between cue detection sensitivity (d') and WM performance (hypothesis 2) and the regression intercept for WM performance on cue detection sensitivity (hypothesis 3), as well as the uncertainty associated with these corrected estimates (see Fig. 5).

The Bayesian hierarchical model requires two key inputs from each participant: their performance and sensitivity measures, and the error variance associated with each measure. This allows for participant-specific reliability estimates rather than assuming a constant reliability across the sample. In Matzke et al. (2017), the model was applied using posterior means and standard deviations for each participant as inputs. In contrast, we directly input each participant's WM performance and d' estimates along with their corresponding frequentist standard errors.[12]

Given that these analyses were not preregistered, we used non-informative priors in all model parameters. As a result, the posterior mean and standard deviation closely resemble frequentist point estimates (Gelman et al. 2013), making the results easier to interpret for researchers unfamiliar with Bayesian inference. However, the posterior distributions obtained in this model can serve as informative priors for future studies adopting a fully Bayesian inference approach. Scripts with model implementation in JAGS (Plummer 2023) and Stan (Stan Development Team 2024) and detailed results for these estimations are available at osf.io/xzv9t.

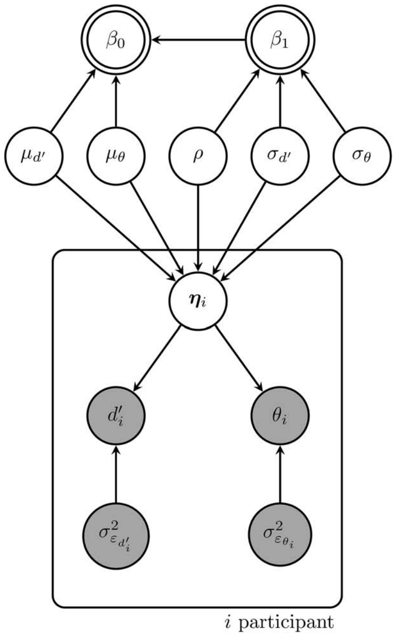
To avoid convergence issues, 12 participants with reliability values lower than or equal to zero or equal to one were excluded. For hypothesis (2), the correlation is corrected from the attenuated[13] $r = .222$ to a posterior mean $\rho = .248$ (one-tailed, 95% credibility interval [.166, ≤ 1]). For hypothesis (3), the regression intercept is corrected from the overestimated $b_0 = .527$ to a posterior mean $\beta_0 = .526$ (one-tailed, 95% credibility interval [.516, ≤ 1]). In conclusion, after taking into account measurement error, results remain practically unchanged and still support the inferences drawn from the preregistered analyses regarding hypotheses (2) and (3).

Since the previous model uses data from both experimental sessions, we decided to expand the Diamond model to assess whether there was an effect of the experimental session on both the correlation and the intercept (see the Spade model in the Supplemental Material, p. 18). Although the correlation did not significantly differ between sessions, there is weak evidence that the intercept (significant in both sessions) increased in the second session. Note that this is the model that allowed us to estimate the disattenuated test–retest reliability coefficients reported at the end of the Reliability analyses section.

[11] This Bayesian hierarchical model has been discussed and developed at least twice within the field of unconscious processing: Malejka et al. (2021) provided a detailed explanation and tutorial on its application, guiding researchers on best practices for analysing correlations in the presence of measurement error. Similarly, Goldstein et al. (2022) adapted it by modelling awareness scores as arising from two distinct subpopulations—i.e. conscious and unconscious perceivers.

[12] Since WM performance is a proportion of success, its standard error can be calculated as $\sqrt{\frac{\theta_i \times (1-\theta_i)}{n_i}}$, where θ_i is the proportion of success and n_i is the number of trials completed by participant i. For d', we have used the standard error formula proposed by Miller (1996, Equations 6–8), as previous research has identified it as the best-performing method (Suero et al. 2017).

[13] We re-computed the correlation and intercept again removing the same 12 participants to compare the estimations with and without measurement error.



$$
\begin{aligned}
\beta_0 &= \mu_\theta - \beta_1 \cdot \mu_{d'} \\
\beta_1 &= \rho \cdot \tfrac{\sigma_\theta}{\sigma_{d'}} \\
\mu_{d'} &\sim \text{Normal}(0,\ 100) \\
\mu_\theta &\sim \text{Uniform}(0,\ 1) \\
\rho &\sim \text{Uniform}(-1,\ 1) \\
\sigma_{d'} &\sim \text{Uniform}(0,\ 100) \\
\sigma_\theta &\sim \text{Uniform}(0,\ 0.5) \\
\boldsymbol{\eta}_i &\sim \text{MVN}\left(\begin{bmatrix} \mu_{d'} \\ \mu_\theta \end{bmatrix}, \begin{bmatrix} \sigma^2_{d'} & \rho\sigma_{d'}\sigma_\theta \\ \rho\sigma_{d'}\sigma_\theta & \sigma^2_\theta \end{bmatrix} \right) \\
d'_i &\sim \text{Normal}\left(\eta_{i1},\ \sigma^2_{\varepsilon_{d'_i}}\right) \\
\theta_i &\sim \text{Normal}\left(\eta_{i2},\ \sigma^2_{\varepsilon_{\theta_i}}\right)
\end{aligned}
$$

Figure 5. Diamond model: adaptation of Matzke et al.'s (2017) Bayesian hierarchical model diagram. *Note.* The graph structure indicates dependencies between the nodes. Shaded and unshaded nodes represent observed and latent variables, respectively. d' represents observed sensitivity; θ, observed WM performance. $\sigma^2_{\varepsilon d'}$ and $\sigma^2_{\varepsilon\theta}$ represent measurement error variances for d' and θ, respectively. $\boldsymbol{\eta}$ represents the true (i.e. free of measurement error) d' and θ values with true population means $\mu_{d'}$ and μ_θ, and true population variances $\sigma^2_{d'}$ and σ^2_θ, respectively. ρ represents the latent true correlation between d' and θ. Finally, β_0 and β_1 represent the latent true intercept and slope regression parameters of θ on d'. MVN: multivariate normal.

Discussion

It has been more than a decade since the first evidence about whether working memory can operate with unconscious representations was presented (Soto et al. 2011). Since then, many other studies have accumulated evidence in the same (Dutta et al. 2014; King et al. 2016) and related tasks (Sklar et al. 2012; Bergström and Eriksson 2014, 2015; Pan et al. 2014; Trübutschek et al. 2019; Nakano and Ishihara 2020). Although a meta-analysis (Gambarota et al. 2022) affirms an overall unconscious WM effect, these studies are individually under-powered (with a median sample size of 17.5 participants each contributing a median of 16 valid trials) and, as noted in the Introduction, display signs of publication or reporting bias.

The present study replicates Soto et al.'s (2011) task with considerable improvements. First, this work is less vulnerable to research bias than previous studies since all methods and analyses were preregistered, reviewed, and accepted for publication in the Stage 1 protocol, before the results were known. As a result, we offer a publicly available dataset from 19 laboratories worldwide with 531 participants, each with 720 trials in two different sessions (a median of 239 valid trials per participant), increasing the statistical power to detect the original effects. Note that the total sample size in Gambarota et al.'s meta-analysis was 689 participants, so our sample size is nearly as large as the total sample in their meta-analysis. Second, our decisions to guarantee the minimal baseline requirements (regarding number of valid trials, motivation, and PAS labels) led to a near-zero between-laboratory variability in our main analysis. Finally, unlike the previous literature in unconscious WM, we obtained reliability estimates for our measures and corrected the estimates of the inferential analyses for measurement error. Below, we present the main results and discuss them following the interpretation plan outlined in the Stage 1 Supplemental Material (Study Plan, p. 10–13).

Preregistered inferential analyses

Does the unconscious WM effect occur? (hypothesis 1)

The first and main question of this multisite study is whether the unconscious WM effect occurs. Soto et al. (2011) found significant above-chance WM performance ($M = .59$, aggregating participants' accuracies) in *unseen cue-present* trials. Here, we accounted for the same effect with a three-level mixed logistic model which isolates this effect from between-participant and between-laboratory variance (i.e. random intercepts). Our fixed intercept of .552 in proportion units was associated with a significant *p*-value, indicating that WM performance is above chance in *unseen cue-present* trials and, in turn, that Soto et al.'s finding was replicated. Our proportion is roughly the same as the value (.555) from Gambarota et al.'s (2022) meta-analysis with a chance level of .50, as in ours. This result serves as evidence against the theory that WM only operates on conscious representations.

To what extent is this effect actually unconscious? (hypotheses 2 and 3)

To explore more deeply to what extent this effect is *actually* unconscious, Soto et al. (2011) tested the correlation between cue detection sensitivity and WM performance (hypothesis 2). Soto et al. computed a pseudo-d', rightly criticized by Stein et al. (2016), so here we computed the common d' from signal detection theory (see Results section). In their experiments, Soto et al. did not find a significant correlation ($r = -.18$, $p = .41$), but we did ($r = .228$, $p < .001$). The reason why we did not reproduce Soto et al.'s null correlation is possibly due to their comparatively lower power (only 22 participants) which did not allow them to detect the subtle but significant correlation. With our initially planned sample size, we had 95% power to detect a minimum correlation of .23 (lower than the pooled correlation found in previous literature: .276). We eventually collected more than twice that sample size ($N_{\text{valid}} = 525$), so the final minimum detectable correlation[14] was .143. Our result serves as evidence against the theory that WM performance and sensitivity are independent, but, as noted in the Stage 1 Manuscript, it does not necessarily exclude an unconscious WM effect: WM could operate with unconscious representations while still improving with greater awareness. This result is consistent with the predictions of the SDT model by Sandberg et al. (2022), which predicts an increase in accuracy within our range of d' (i.e. from 0 to ∼2.5; see their figure 5B).

Another analysis conducted by Soto et al. (2011) tested whether the intercept of a linear regression of WM performance on cue detection sensitivity was above chance (hypothesis 3). They found that, at the point of zero d' (i.e. an ideally unconscious participant), WM performance was still above chance ($\beta_0 = .60$, no inferential information provided). With our dataset, we reproduced their result, obtaining a significant positive intercept of $\beta_0 = .521$ ($p < .001$), although smaller than the one they found. Our result could serve as evidence against the claim that an ideal observer with null sensitivity will perform at chance, but it could also reflect a statistical artefact (spurious intercept) in case the reliability of d' is insufficient. For this reason, it is crucial to combine these results with estimations of the measures' reliability.

Reliability (and validity) analyses

Few studies report the reliability of the dependent variables collected in experimental tasks—in fact, none of the studies in Gambarota et al.'s 2022 meta-analysis did—and, when reported, the reliability estimates often leave much to be desired (≈.54 in Garre-Frutos et al. 2024; <.53 in Hernández-Gutiérrez et al. 2025; ≈.52 in studies reviewed by Rothkirch et al. 2022; ≈.44 in Vadillo et al. 2022; ≈.34 in Vadillo et al. 2024, and <.50 in 14 of 18 datasets analysed by Yaron et al. 2024). In this context, we must be prepared for our measures to be somewhat unreliable, and consequently, for any subsequent inferences to be potentially undermined.

Nonetheless, with our large sample size and number of trials, our measures have excellent reliability: .94 for d' cue detection sensitivity and .73–.79 for WM performance. Note that, applying the Spearman–Brown prophecy formula, with the median number of valid trials in the literature, these reliabilities could have been near .393 for d' and .222–.253 for WM performance. We did not preregister any inferential analysis regarding reliability, but we explored a measurement error correction for the correlation and intercept of hypotheses (2) and (3), respectively, with a Bayesian hierarchical model. Once corrected, both inferences remained unchanged, suggesting that our current conclusions are quite robust. However, obtaining an estimation of the PAS measure's reliability itself is still a pending task.

In light of these findings, some researchers might misleadingly believe that measurement error is not as a major concern as has recently been suggested (e.g. Vadillo et al. 2022). However, we encourage readers to keep in mind the characteristics of the sample and measures analysed here. In our case, the experimental procedure was so robust that no differences were found in WM performance between the 19 laboratories that participated in data collection. This result is not coincidental. Since this project started, we spent over a year designing the task (enhancing aspects of the original by Soto et al. 2011), reviewing the Stage 1 protocol, and obtaining pilot data to ensure a good understanding of how participants would complete both the WM task and the PAS. Unlike in the original design by Soto et al., here we added a calibration algorithm during all experimental trials to ensure the contrast of the stimuli consistently elicited ∼50% unseen trials. We also included *supraliminal* trials to maintain motivation throughout the 360 experimental trials and collected data across two different sessions. Instructions, PsychoPy scripts, and REI items were carefully back-translated into 11 languages and checked with pilot data. Collaborators in all laboratories verified their computers and monitors (frame rate stability, Internet access, resolution, etc.) before data collection and submitted a video simulation with a fake participant to ensure the appropriateness of their setup. In summary, this meticulous attention to design details, combined with the efforts of international collaborators, contributed to what should be an expected consequence: highly reliable measures.

However, most of these details should be linked not only to reliability, but—beyond it—to validity. Imagine that, consistently, all participants understood the WM task the other way around (instead of paying attention to the top of the Gabors' lines to check rotation, they looked at the bottom of the lines), so when they report the Gabor rotated right, the Gabor actually rotated left, and *vice versa*. Since all participants' understanding is consistent, the reliability of the WM performance would be very high, suggesting that we can rely on whatever result we get. In this scenario, instead of obtaining an unconscious WM effect of 55% accuracy in valid trials as we did, this systematic error would result in accuracy of ∼45%. In other words, we would have wrongly concluded that WM accuracy is not above chance and that there is no evidence for an unconscious WM effect. This example highlights the importance of considering validity alongside reliability (Kerschbaumer et al. 2025). Because not many research teams can afford a multisite sample with double-session data, future research will need to unravel which of all design details contribute most to reliability and, ultimately, to validity.

Conclusion

This work offers evidence of a significant unconscious WM effect. Although higher cue detection sensitivity (d') was related to better WM performance—unlike the statistical independence found in Soto et al. (2011)—an ideally unconscious participant ($d' = 0$) would still perform significantly above chance in the WM task. These conclusions are framed under a particular task (visual discrimination of the direction in which a Gabor was rotated after a delay of 2 s) and a particular masking method (a pattern of random dots). Different stimuli (e.g. audible, linguistic, or even other types of visual stimuli…), different WM operations (i.e. maintenance versus manipulation), or different methods to manipulate awareness (i.e. beyond traditional masking techniques) might yield different patterns of results regarding the relationship between WM and conscious awareness.

Note that our experiment does not aim to investigate the mechanisms underlying this effect. While the findings support the idea of an unconscious WM effect, our procedure cannot fully discriminate between alternative explanations. For instance, we cannot ensure whether, after being presented with the subliminal Gabor, participants maintain the unconscious percept in WM or, instead, generate a conscious guess that is then maintained in WM (see Barton et al. 2022 for a similar concern). In either case, it could be argued that this process would still qualify as unconscious WM since the information being maintained originates from a *subjectively unseen* stimulus. Perhaps more importantly, our modelling results indicate that delayed

[14] Calculated with the function `pwr.t.test()` of the `{pwr}` package in R (Champely 2020), for a one-tailed t-test and $\alpha < .05$.

cue-target discrimination performance remains above chance even when perceptual sensitivity is at chance level. This result suggests that an unconscious representation of the cue can drive the effect on the subsequent discrimination: Since no reliable conscious guess can arise with null perceptual sensitivity of the cue, it is unlikely that participants kept in memory a conscious guess of the masked cue. Thus, we propose that the results are more compatible with the view that WM operates on unconscious input (i.e. maintaining a subliminal representation to guide subsequent perceptual decision making). Further explorations, for instance of the response times in seen and unseen trials, might provide useful evidence to this issue.

Additional analyses of this openly available dataset remain to be explored. For instance, with the REI scores (Pacini and Epstein 1999) already collected, we are studying whether participants' intuition predicts their unconscious WM performance. There may also be learning effects across sessions in how people accumulate evidence from non-conscious input to guide WM-based decisions, as suggested by the Spade model estimated in the Supplemental Material. Notably, this learning effect could be moderated by the length of the interval between sessions, which ranged from 1 to 17 days. This rich dataset may also allow modelling the entire PAS response distribution to avoid *post hoc* selection of PAS = 1 ratings. We hope that, after assessing this article, readers will have new ideas for further analyses or unanswered questions. All materials—including codebooks and self-explanatory scripts—are publicly available in our OSF repository (osf.io/xzv9t). We invite readers to access the dataset and build upon it in future investigations.

Acknowledgements

We would like to thank the coordinators of the Center for Open Science for their support with budget management, and Antón D. Pérez Rodríguez from UAM for his always kind assistance with administrative procedures. We extend special thanks to editor Zoltan Dienes for his meticulous and insightful guidance since the Stage 1 review.

Author contributions

Alicia Franco-Martínez (Conceptualization [equal], Data curation [lead], Formal analysis [equal], Funding acquisition [equal], Investigation [lead], Methodology [lead], Project administration [lead], Resources [lead], Software [equal], Visualization [lead], Writing—original draft [lead]), Ricardo Rey-Sáez (Data curation [equal], Formal analysis [equal], Investigation [equal], Methodology [equal], Software [equal], Validation [equal], Writing—review & editing [equal]), Jesús Adrián-Ventura (Investigation [supporting], Writing—review & editing [supporting]), Pietro Amerio (Investigation [supporting], Writing—review & editing [supporting]), Ana Baciero (Investigation [supporting], Writing—review & editing [supporting]), Amine Bennis (Investigation [supporting], Writing—review & editing [supporting]), Fredrik Bergström (Investigation [supporting], Writing—review & editing [supporting]), Axel Cleeremans (Investigation [supporting], Writing—review & editing [supporting]), Laura Contu (Investigation [supporting], Writing—review & editing [supporting]), Roberto Dell'Acqua (Investigation [supporting], Writing—review & editing [supporting]), Xinping Deng (Investigation [supporting], Writing—review & editing [supporting]), Fatma Nur Dolu (Investigation [supporting], Writing—review & editing [supporting]), Filippo Gambarota (Investigation [supporting], Writing—review & editing [supporting]), Yi Gao (Investigation [supporting], Writing—review & editing [supporting]), Francisco Garre-Frutos (Investigation [supporting], Writing—review & editing [supporting]), Anna Grubert (Investigation [supporting], Writing—review & editing [supporting]), Ana Hernando (Investigation [supporting], Writing—review & editing [supporting]), José A. Hinojosa (Investigation [supporting], Writing—review & editing [supporting]), Asaf Hoory (Investigation [supporting], Writing—review & editing [supporting]), ZhiKun Hou (Investigation [supporting], Writing—review & editing [supporting]), Shao-Min Hung (Investigation [supporting], Writing—review & editing [supporting]), Aine Ito (Investigation [supporting], Writing—review & editing [supporting]), Mikel Jimenez (Investigation [supporting], Writing—review & editing [supporting]), Alexandra I. Kosachenko (Investigation [supporting], Writing—review & editing [supporting]), Merve Kulaksiz (Investigation [supporting], Writing—review & editing [supporting]), Antonio Prieto (Investigation [supporting], Writing—review & editing [supporting]), Daryl Y.H. Lee (Investigation [supporting], Writing—review & editing [supporting]), Elmas Merve Malas (Investigation [supporting], Writing—review & editing [supporting]), Simone Malejka (Investigation [supporting], Writing—review & editing [supporting]), Pedro R. Montoro (Methodology [supporting], Writing—review & editing [supporting]), Liad Mudrik (Investigation [supporting], Writing—review & editing [supporting]), Yuri Pavlov (Investigation [supporting], Writing—review & editing [supporting]), Gabriele Pesimena (Investigation [supporting], Writing—review & editing [supporting]), Dobromir Rahnev (Investigation [supporting], Writing—review & editing [supporting]), Lais Ringenberg (Investigation [supporting], Writing—review & editing [supporting]), Alejandro Sandoval-Lentisco (Investigation [supporting], Writing—review & editing [supporting]), Akira Sarodo (Investigation [supporting], Writing—review & editing [supporting]), Maor Schreiber (Investigation [supporting], Writing—review & editing [supporting]), Paola Sessa (Investigation [supporting], Writing—review & editing [supporting]), Pablo Solana (Investigation [supporting], Writing—review & editing [supporting]), Dmitrii A. Tarasov (Investigation [supporting], Writing—review & editing [supporting]), Miriam Tortajada (Investigation [supporting], Writing—review & editing [supporting]), Kai Xue (Investigation [supporting], Writing—review & editing [supporting]), Ziqian Xue (Investigation [supporting], Writing—review & editing [supporting]), Yunxuan Zheng (Investigation [supporting], Writing—review & editing [supporting]), Merve Çinici (Investigation [supporting], Writing—review & editing [supporting]), David Shanks (Conceptualization [equal], Funding acquisition [supporting], Investigation [supporting], Methodology [equal], Resources [equal], Supervision [equal], Writing—review & editing [lead]), David Soto (Conceptualization [equal], Funding acquisition [supporting], Methodology [equal], Supervision [equal], Writing—review & editing [lead]), and Miguel Vadillo (Conceptualization [equal], Funding acquisition [equal], Investigation [equal], Methodology [equal], Project administration [equal], Resources [equal], Software [supporting], Supervision [lead], Writing—original draft [supporting], Writing—review & editing [lead])

Conflict of interest

The authors declare that they have no conflicts of interest.

Funding

This Registered Report was funded by the 'Funding Consciousness Research with Registered Reports' initiative from the Center for Open Science. A.F.M. was supported by grant PRE2021-097654 from the Spanish Ministry of Science and

Innovation. R.R.S. and M.A.V. were supported by grant CNS2022-135346 from Agencia Estatal de Investigación/FEDER UE. D.R.S. was supported by grant ES/S014616/1 from the UK Economic and Social Research Council. D.S. acknowledges support from the Basque Government through the BERC 2022–2025 programme, from Agencia Estatal de Investigación, through the 'Severo Ochoa' Programme for Centres/Units of Excellence in R&D (CEX2020-001010-S), and also from project grant PID2019-105494GB-I00. M.A.V. and A.F.M. were supported by grant PID2020-118583GB-I00 from Agencia Estatal de Investigación/FEDER UE. External collaborators were also supported by the following funders: P.A. was supported by an F.R.S.-FNRS Research Project T003821F (40003221) to A.C. F.B. was supported by Fundação para a Ciência e Tecnologia (CEECIND/03661/2017). A.C. is a Research Director with the F.R.S.-FNRS (Belgium) and a fellow of the Canadian Institute for Advanced Research (Brain, Mind and Consciousness programme), funded by European Research Council Advanced Grant #101055060 'EXPERIENCE'. F.G.F. was supported by a predoctoral fellowship from the Spanish Ministry of Science, Innovation and Universities (FPU20/00826). J.A.H., A.P., P.R.M., and M.J. were supported by grant PID2021-125842NB-I00 from Agencia Estatal de Investigación. S.M.H. was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI 'Early-Career Scientists' (24K16877). A.I. was supported by the HSS Seed Fund from the National University of Singapore (#A-8001370-00-00). A.I. would like to thank Shiyang Wu and Mingyuan Yang for help with data collection. P.S. was supported by FPU grant FPU20/01946 from the Spanish Ministry of Science and Innovation.

Data availability

All materials, including PsychoPy experiment versions, data, consent forms, codebooks and selfexplanatory scripts, are publicly available in our OSF repository: osf.io/xzv9t.

References

Baars BJ. In the theatre of consciousness. Global workspace theory, a rigorous scientific theory of consciousness. *J Conscious Stud* 1997;**4**:292–309.

Baddeley A. Working memory: looking back and looking forward. *Nat Rev Neurosci* 2003;**4**:829–39. https://doi.org/10.1038/nrn1201

Baddeley AD, Hitch G. Working memory. *Psychol Learn Motiv* 1974;**8**: 47–89. https://doi.org/10.1016/S0079-7421(08)60452-1

Barton AU, Valle-Inclán F, Cowan N *et al.* Unconsciously registered items reduce working memory capacity. *Conscious Cogn* 2022;**105**:103399. https://doi.org/10.1016/j.concog.2022.103399

Bartoš F, Maier M, Wagenmakers EJ *et al.* Robust Bayesian meta-analysis: model-averaging across complementary publication bias adjustment methods. *Res Synth Methods* 2021;**14**:99–116. https://doi.org/10.1002/jrsm.1594

Behseta S, Berdyyeva TK, Olson CR *et al.* Bayesian correction for attenuation of correlation in multi-trial spike count data. *J Neurophysiol* 2009;**101**:2186–93. https://doi.org/10.1152/jn.90727.2008

Bergström F, Eriksson J. Maintenance of non-consciously presented information engages the prefrontal cortex. *Front Hum Neurosci* 2014;**8**:1–10. https://doi.org/10.3389/fnhum.2014.00938

Bergström F, Eriksson J. The conjunction of non-consciously perceived object identity and spatial position can be retained during a visual short-term memory task. *Front Psychol* 2015;**6**:1–9. https://doi.org/10.3389/fpsyg.2015.01470

Bergström F, Eriksson J. Neural evidence for non-conscious working memory. *Cereb Cortex* 2018;**28**:3217–28. https://doi.org/10.1093/cercor/bhx193

Biderman N, Mudrik L. Evidence for implicit—but not unconscious—processing of object-scene relations. *Psychol Sci* 2018;**29**:266–77. https://doi.org/10.1177/0956797617735745

Bona S, Cattaneo Z, Vecchi T *et al.* Metacognition of visual short-term memory: dissociation between objective and subjective components of VSTM. *Front Psychol* 2013;**4**:1–6. https://doi.org/10.3389/fpsyg.2013.00062

Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. *Measurement Error in Nonlinear Models: A Modern Perspective* (2nd edition). New York: Chapman & Hall/CRC, 2006, https://doi.org/10.1201/9781420010138.

Carter EC, Schönbrodt FD, Gervais WM *et al.* Correcting for bias in psychology: a comparison of meta-analytic methods. *Adv Methods Pract Psychol Sci* 2019;**2**:115–44. https://doi.org/10.1177/2515245919847196

Chambers CD, Tzavella L. The past, present and future of registered reports. *Nat Hum Behav* 2022;**6**:29–42. https://doi.org/10.1038/s41562-021-01193-7

Champely S. pwr: basic functions for power analysis. 2020. R package version 1.3-0, Comprehensive R Archive Network (CRAN). https://CRAN.R-project.org/package=pwr

Dutta A, Shah K, Silvanto J *et al.* Neural basis of non-conscious visual working memory. *Neuroimage* 2014;**91**:336–43. https://doi.org/10.1016/j.neuroimage.2014.01.016

Francken JC, Beerendonk L, Molenaar D *et al.* An academic survey on theoretical foundations, common assumptions and the current state of consciousness science. *Neurosci Conscious* 2022;**2022**:niac011. https://doi.org/10.1093/nc/niac011

Fuller WA. *Measurement Error Models*. New York: John Wiley & Sons, 1987, https://doi.org/10.1002/9780470316665

Gambarota F, Tsuchiya N, Pastore M *et al.* Unconscious visual working memory: a critical review and Bayesian meta-analysis. *Neurosci Biobehav Rev* 2022;**136**:1–14. https://doi.org/10.1016/j.neubiorev.2022.104618

Garre-Frutos F, Vadillo MA, González F *et al.* On the reliability of value-modulated attentional capture: an online replication and multiverse analysis. *Behav Res Methods* 2024;**56**:5986–6003. https://doi.org/10.3758/s13428-023-02329-5

Gelman A, Carlin JB, Stern HS *et al. Bayesian Data Analysis* (3rd edition). New York: Chapman & Hall/CRC, 2013, https://doi.org/10.1201/b16018

Goldstein A, Sklar AY, Siegelman N. Accurately measuring nonconscious processing using a generative Bayesian framework. *Psychol Conscious Theory Res Pract* 2022;**9**:336–55. https://doi.org/10.1037/cns0000316

Greenwald AG, Draine SC. Do subliminal stimuli enter the mind unnoticed? Tests with a new method. In: Cohen JD, Schooler JW (eds.), *Scientific Approaches to Consciousness*, pp. 83–108. New York: Lawrence Erlbaum Associates, 1997.

Greenwald AG, Klinger MR, Schuh ES. Activation by marginally perceptible ("subliminal") stimuli: dissociation of unconscious from conscious cognition. *J Exp Psychol Gen* 1995;**124**:22–42. https://doi.org/10.1037/0096-3445.124.1.22

Greenwald AG, Draine SC, Abrams RL. Three cognitive markers of unconscious semantic activation. *Science* 1996;**273**:1699–702. https://doi.org/10.1126/science.273.5282.1699

Hassin RR, Bargh JA, Engell AD *et al.* Implicit working memory. *Conscious Cogn* 2009;**18**:665–78. https://doi.org/10.1016/j.concog.2009.04.003

Hautus MJ. Corrections for extreme proportions and their biasing effects on estimated values of *d*′. *Behav Res Methods Instrum Comput* 1995;**27**:46–51. https://doi.org/10.3758/BF03203619

Hernández-Gutiérrez D, Sorrel MA, Shanks DR *et al.* The conscious side of ‘subliminal’ linguistic priming: a systematic review with meta-analysis and reliability analysis of visibility measures. *J Cogn* 2025;**8**:13, 1–20. https://doi.org/10.5334/joc.419

Hunter JE, Schmidt FL. *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. 3rd edition. Thousand Oaks, California: SAGE Publications, 2015. https://doi.org/10.4135/9781483398105

Jachs B, Blanco MJ, Grantham-Hill S *et al.* On the independence of visual awareness and metacognition: a signal detection theoretic analysis. *J Exp Psychol Hum Percept Perform* 2015;**41**:269–76. https://doi.org/10.1037/xhp0000026

Kahveci S (2025). rapidsplithalf: a fast permutation-based split-half reliability algorithm. R package version 0.4, Comprehensive R Archive Network (CRAN). https://CRAN.R-project.org/package=rapidsplithalf

Kahveci S, Bathke A, Blechert J. Reliability of reaction time tasks: how should it be computed? *[preprint] PsyArXiv*. 2022. doi.org/10.31234/osf.io/ta59r

Kerschbaumer S, Voracek M, Aczél B *et al.* VALID: a checklist-based approach for improving validity in psychological research. *Adv Methods Pract Psychol Sci* 2025;**8**:1–16. https://doi.org/10.1177/25152459241306432

King JR, Pescetelli N, Dehaene S. Brain mechanisms underlying the brief maintenance of seen and unseen sensory information. *Neuron* 2016;**92**:1122–34. https://doi.org/10.1016/j.neuron.2016.10.051

Klauer KC, Draine SC, Greenwald AG. An unbiased errors-in-variables approach to detecting unconscious cognition. *Br J Math Stat Psychol* 1998;**51**:253–67. https://doi.org/10.1111/j.2044-8317.1998.tb00680.x

Malejka S, Vadillo MA, Dienes Z *et al.* Correlation analysis to investigate unconscious mental processes: a critical appraisal and mini-tutorial. *Cognition* 2021;**212**:104667. https://doi.org/10.1016/j.cognition.2021.104667

Matzke D, Ly A, Selker R *et al.* Bayesian inference for correlations in the presence of measurement error and estimation uncertainty. *Collabra Psychol* 2017;**3**:25. https://doi.org/10.1525/collabra.78

Miller J. Measurement error in subliminal perception experiments: simulation analyses of two regression methods. *J Exp Psychol Hum Percept Perform* 2000;**26**:1461–77. https://doi.org/10.1037/0096-1523.26.4.1461

Miller J. The sampling distribution of *d*′. *Percept Psychophys* 1996;**58**:65–72. https://doi.org/10.3758/BF03205476

Nakano S, Ishihara M. Working memory can compare two visual items without accessing visual consciousness. *Conscious Cogn* 2020;**78**:102859–11. https://doi.org/10.1016/j.concog.2019.102859

Overgaard M, Timmermans B, Sandberg K *et al.* Optimizing subjective measures of consciousness. *Conscious Cogn* 2010;**19**:682–4. https://doi.org/10.1016/j.concog.2009.12.018

Pacini R, Epstein S. The relation of rational and experiential information processing styles to personality, basic beliefs, and the ratio-bias phenomenon. *J Pers Soc Psychol* 1999;**76**:972–87. https://doi.org/10.1037/0022-3514.76.6.972

Pan Y, Lin B, Zhao Y *et al.* Working memory biasing of visual perception without awareness. *Atten Percept Psychophys* 2014;**76**:2051–62. https://doi.org/10.3758/s13414-013-0566-2

Peirce JW, Gray JR, Simpson S *et al.* PsychoPy2: experiments in behavior made easy. *Behav Res Methods* 2019;**51**:195–203. https://doi.org/10.3758/s13428-018-01193-y

Persuh M, LaRock E, Berger J. Working memory and consciousness: the current state of play. *Front Hum Neurosci* 2018;**12**:1–11. https://doi.org/10.3389/fnhum.2018.00078

Plummer M. *JAGS version* 4.3.2 [computer software]. 2023. Retrieved from https://mcmc-jags.sourceforge.io/

Pratte MS, Rouder JN. A task-difficulty artifact in subliminal priming. *Atten Percept Psychophys* 2009;**71**:1276–83. https://doi.org/10.3758/APP.71.6.1276

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2024. https://www.R-project.org/.

Ramsøy TZ, Overgaard M. Introspection and subliminal perception. *Phenomenol Cogn Sci* 2004;**3**:1–23. https://doi.org/10.1023/B:PHEN.0000041900.30172.e8

Rothkirch M, Shanks DR, Hesselmann G. The pervasive problem of post hoc data selection in studies on unconscious processing: a reply to Sklar, Goldstein, and Hassin (2021). *Exp Psychol* 2022;**69**:1–11. https://doi.org/10.1027/1618-3169/a000541

Rouder JN, Haaf JM. A psychometrics of individual differences in experimental tasks. *Psychon Bull Rev* 2019;**26**:452–67. https://doi.org/10.3758/s13423-018-1558-y

Rouder JN, Kumar A, Haaf JM. Why many studies of individual differences with inhibition tasks may not localize correlations. *Psychon Bull Rev* 2023;**30**:2049–66. https://doi.org/10.3758/s13423-023-02293-3

Rowe EG, Tsuchiya N, Garrido MI. Detecting (un)seen change: the neural underpinnings of (un)conscious prediction errors. *Front Syst Neurosci* 2020;**14**:1–15. https://doi.org/10.3389/fnsys.2020.541670

Sandberg K, Overgaard M. Using the perceptual awareness scale (PAS). In: Overgaard M (ed.) *Behavioural Methods in Consciousness Research*. USA: Oxford University Press, 2015, 181–96. https://doi.org/10.1093/acprof:oso/9780199688890.003.0011.

Sandberg K, Timmermans B, Overgaard M *et al.* Measuring consciousness: is one measure better than the other? *Conscious Cogn* 2010;**19**:1069–78. https://doi.org/10.1016/j.concog.2009.12.013

Sandberg K, Del Pin SH, Overgaard M *et al.* A window of subliminal perception. *Behav Brain Res* 2022;**426**:113842. https://doi.org/10.1016/j.bbr.2022.113842

Shanks DR. Regressive research: the pitfalls of post hoc data selection in the study of unconscious mental processes. *Psychon Bull Rev* 2017;**24**:752–75. https://doi.org/10.3758/s13423-016-1170-y

Shanks DR, Malejka S, Vadillo MA. The challenge of inferring unconscious mental processes. *Exp Psychol* 2021;**68**:113–29. https://doi.org/10.1027/1618-3169/a000517

Simonsohn U, Simmons JP, Nelson LD. Specification curve analysis. *Nat Hum Behav* 2020;**4**:1208–14. https://doi.org/10.1038/s41562-020-0912-z

Sklar AY, Levy N, Goldstein A *et al.* Reading and doing arithmetic nonconsciously. *Proc Natl Acad Sci* 2012;**109**:19614–9. https://doi.org/10.1073/pnas.1211645109

Skóra Z, Ciupińska K, Del Pin SH *et al.* Investigating the validity of the perceptual awareness scale–the effect of task-related difficulty on subjective rating. *Conscious Cogn* 2021;**95**:103–97. https://doi.org/10.1016/j.concog.2021.103197

Soto D, Silvanto J. Is conscious awareness needed for all working memory processes? *Neurosci Conscious* 2016;**2016**:1–3. https://doi.org/10.1093/nc/niw009

Soto D, Mäntylä T, Silvanto J. Working memory without consciousness. *Curr Biol* 2011;**21**:R912–3. https://doi.org/10.1016/j.cub.2011.09.049

Spearman C. The proof and measurement of association between two things. *Am J Psychol* 1904;**15**:72–471. https://doi.org/10.2307/1412159

Stan Development Team. *Stan Version* 2.36.0 *[Computer Software Manual]* 2024. Retrieved from https://mc-stan.org

Stanislaw H, Todorov N. Calculation of signal detection theory measures. *Behav Res Methods Instrum Comput* 1999;**31**:137–49. https://doi.org/10.3758/BF03207704

Steegen S, Tuerlinckx F, Gelman A *et al.* Increasing transparency through a multiverse analysis. *Perspect Psychol Sci* 2016;**11**:702–12. https://doi.org/10.1177/1745691616658637

Stein T, Kaiser D, Hesselmann G. Can working memory be non-conscious? *Neurosci Conscious* 2016;**2016**:1–3. https://doi.org/10.1093/nc/niv011

Suero M, Privado J, Botella J. Methods to estimate the variance of some indices of the signal detection theory: a simulation study. *Psicologica* 2017;**38**:149–75.

Taglialatela Scafati I. Is there evidence for non-conscious processing in working memory? [Ph.D. dissertation, University of Edinburgh] 2019. http://hdl.handle.net/1842/36170.

Trübutschek D, Marti S, Ojeda A *et al.* A theory of working memory without consciousness or sustained activity. *eLife* 2017;**6**:1–29. https://doi.org/10.7554/eLife.23871

Trübutschek D, Marti S, Dehaene S. Temporal-order information can be maintained in non-conscious working memory. *Sci Rep* 2019;**9**: 6484–10. https://doi.org/10.1038/s41598-019-42942-z

Vadillo MA, Konstantinidis E, Shanks DR. Underpowered samples, false negatives, and unconscious learning. *Psychon Bull Rev* 2016;**23**:87–102. https://doi.org/10.3758/s13423-015-0892-6

Vadillo MA, Linssen D, Orgaz C *et al.* Unconscious or underpowered? Probabilistic cuing of visual attention. *J Exp Psychol Gen* 2020;**149**: 160–81. https://doi.org/10.1037/xge0000632

Vadillo MA, Malejka S, Lee DY *et al.* Raising awareness about measurement error in research on unconscious mental processes. *Psychon Bull Rev* 2022;**29**:21–43. https://doi.org/10.3758/s13423-021-01923-y

Vadillo MA, Malejka S, Shanks DR. Mapping the reliability multiverse of contextual cuing. *J Exp Psychol Learn Mem Cogn* 2024;**51**:910–27. https://doi.org/10.1037/xlm0001410

Watson AB, Pelli DG. Quest: a Bayesian adaptive psychometric method. *Percept Psychophys* 1983;**33**:113–20. https://doi.org/10.3758/BF03202828

Wiens S, Andersson A, Gravenfors J. Neural electrophysiological correlates of detection and identification awareness. *Cogn Affect Behav Neurosci* 2023;**23**:1303–21. https://doi.org/10.3758/s13415-023-01120-5

Yaron I, Zeevi Y, Korisky U *et al.* Progressing, not regressing: a possible solution to the problem of regression to the mean in unconscious processing studies. *Psychon Bull Rev* 2024;**31**:49–64. https://doi.org/10.3758/s13423-023-02326-x

Neuroscience of Consciousness, 2026, **2026(1)**, **2026**, niaf046
https://doi.org/10.1093/nc/niaf046
Registered Report (final)