Journal of Experimental Psychology: General

Advancing Research on Unconscious Priming: When Can Scientists Claim an Indirect Task Advantage?

Sascha Meyen, Iris A. Zerweck, Catarina Amado, Ulrike von Luxburg, and Volker H. Franz

Online First Publication, July 15, 2021. http://dx.doi.org/10.1037/xge0001065

AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

ISSN: 0096-3445

https://doi.org/10.1037/xge0001065

Advancing Research on Unconscious Priming: When Can Scientists Claim an Indirect Task Advantage?

Sascha Meyen[1], Iris A. Zerweck[1], Catarina Amado[1], Ulrike von Luxburg[1, 2], and Volker H. Franz[1]
[1] Department of Computer Science, University of Tübingen
[2] Max Planck Institute for Intelligent Systems, Tübingen, Germany

Current literature holds that many cognitive functions can be performed outside consciousness. Evidence for this view comes from unconscious priming. In a typical experiment, visual stimuli are masked such that participants are close to chance performance when directly asked to which of two categories the stimuli belong. This close-to-zero sensitivity is seen as evidence that participants cannot consciously report the category of the masked stimuli. Nevertheless, the category of the masked stimuli can indirectly affect responses to other stimuli (e.g., reaction times or brain activity)—an effect called priming. The priming effect is seen as evidence for a higher sensitivity to the masked stimuli in the indirect responses as compared with the direct responses. Such an apparent difference in sensitivities is taken as evidence that processing occurred unconsciously. But we show that this "standard reasoning of unconscious priming" is flawed: Sensitivities are not properly compared, creating the wrong impression of a difference in sensitivities even if there is none. We describe the appropriate way to determine sensitivities, replicate the behavioral part of a landmark study, develop methods to estimate sensitivities from reported summary statistics of published studies, and use these methods to reanalyze 15 highly influential studies. Results show that the interpretations of many studies need to be changed and that a community effort is required to reassess the vast literature on unconscious priming. This process will allow scientists to learn more about the true boundary conditions of unconscious priming, thereby advancing the scientific understanding of consciousness.

Keywords: consciousness, unconscious priming, reanalysis, indirect task advantage, signal detection theory

Supplemental materials: https://doi.org/10.1037/xge0001065.supp

Research on consciousness and its cerebral substrates has far-reaching implications and received substantial attention in recent years (Michel et al., 2019). A driving factor comes from reports that masked stimuli that are not consciously perceived can nevertheless affect behavioral responses and brain activity (Kouider & Dehaene, 2007; Van den Bussche et al., 2009). The exciting claim here is that unconscious processing might be more than a mere residue of conscious processing and may be performed by different neuronal processes than conscious processing. Such results impact current theories about the functional role of consciousness (Dehaene et al., 2017; Hassin, 2013; Kouider & Dehaene, 2007; Sklar et al., 2012; Van den Bussche et al., 2009), might suggest parallel neuronal routes for unconscious versus conscious processing (Morris et al., 1999), and might support theories of superior

Sascha Meyen https://orcid.org/0000-0001-6928-4126
Iris A. Zerweck https://orcid.org/0000-0002-6043-0550
Catarina Amado https://orcid.org/0000-0002-5768-0428
Volker H. Franz https://orcid.org/0000-0002-8457-0611

We have no known conflict of interest to disclose.

Sascha Meyen, Volker H. Franz, and Ulrike von Luxburg developed the methods. Iris A. Zerweck conducted the experiment. All authors wrote the article.

The authors declared no conflicts of interest with respect to the authorship or the publication of this article.

All data and analyses are available at the Open Science Framework (https://osf.io/kp59h), including R code to reanalyze additional studies.

Supported by Deutsche Forschungsgemeinschaft (DFG/German Research Foundation) CRC 1233 "Robust Vision" Project 276693517; Institutional Strategy of University of Tübingen (DFG, ZUK 63); Cluster of Excellence Machine Learning: New Perspectives for Science, EXC 2064/1, Project 390727645.

A previous version of this article has been published on arXiv (https://arxiv.org/pdf/2004.14987.pdf). The results were discussed in a poster session at the (a) GK Doctoral Symposium on Cognitive Science, Tübingen, Germany in 2020; (b) Annual Meeting of Vision Sciences Society, Florida, United States in 2019; (c) European Conference on Visual Perception, Trieste, Italy in 2018; and Tagung Experimentell Arbeitender Psychologen, Marburg, Germany in 2018.

Correspondence concerning this article should be addressed to Sascha Meyen, Department of Computer Science, University of Tübingen, Sand 6, 72076 Tübingen, Germany. Email: sascha.meyen@uni-tuebingen.de

unconscious processing (Custers & Aarts, 2010; Dijksterhuis et al., 2006; ten Brinke et al., 2016).

Here, we scrutinize one of the most frequently used approaches in this field. We show that the *standard reasoning* in the dissociation paradigm (Hannula et al., 2005; Holender, 1986; Schmidt & Vorberg, 2006; Simons et al., 2007) is flawed for mathematical reasons. It fails to provide meaningful interpretation of the data, and needs to be replaced by an *appropriate analysis*. Because many studies have used the standard reasoning, a large body of literature needs reassessment. This has the potential to drastically change our views on unconscious processing and its neuronal underpinnings. The fallacy we expose affects a wide range of research areas because the standard reasoning has been employed on such diverse topics as, for example, unconscious processing of semantic meaning (Dehaene et al., 1998), motivation (Pessiglione et al., 2007), emotion (Morris et al., 1998), cognitive control (van Gaal et al., 2010), and detection of lies (ten Brinke et al., 2014).

To assess how seriously the literature is affected, we proceeded in three strands: (a) We replicated the behavioral part of a landmark study (Dehaene et al., 1998) and showed that the appropriate analysis of the data does not support unconscious priming (in contrast to the claims of the original study). (b) We developed statistical methods to reanalyze published studies based on the reported t and F statistics (because access to the full trial-by-trial data is often lacking). We validated this approach by showing that our reanalysis of the published data of Dehaene et al. (1998) is consistent with the results of our replication. (c) We used our methods to reanalyze 15 highly influential studies (with a total of 3,277 citations in Web of Science). Even though all these studies used the standard reasoning to infer unconscious processing, their data tell a different story.

The Standard Reasoning of Unconscious Priming

As a typical example for a study using the standard reasoning, consider the study by ten Brinke et al. (2014) who reported that humans can detect liars better unconsciously than consciously: "[T]he unconscious mind identifies and processes cues to deception . . . more efficiently and effectively than the conscious mind" (p. 1104). In the following, we will describe the specifics of this study as well as the general aspects that are typical for studies using the standard reasoning.

Participants of ten Brinke et al. (2014) first watched videos of suspects who were either lying or telling the truth. Then participants performed two tasks: The direct and the indirect task. These tasks were supposed to measure conscious and unconscious lie detection, respectively.

In the direct task, participants judged which suspects had been lying or telling the truth. Participants performed poorly with an accuracy of only 49.62%-correct (with chance level being 50%), which was taken by ten Brinke et al. (2014) as evidence that participants could not consciously detect liars with more than a poor sensitivity. In the same way, studies using the standard reasoning typically let participants directly discriminate stimuli belonging to one of two categories (see Figure 1). Participants' performance—measured by the proportion of correct responses or by the sensitivity index, d', from signal detection theory (Green & Swets, 1988)—is typically found to be close to chance level. This result is then taken as evidence that conscious discrimination of the presented stimuli is poor at best.

In the *indirect task* of ten Brinke et al. (2014), participants categorized target–words, such as "deceitful" or "honest," into two categories: lying or truth–telling. Before each target–word, a masked picture of one of the suspects was briefly presented in order to affect (or "prime") the responses to the target words (therefore those masked stimuli are often called the "primes"). ten Brinke et al. (2014) found that participants' reaction times (RTs) to the target words were faster when the primes were congruent with the targets (e.g., the picture of a lying suspect before a lie–related word) than when the primes were incongruent with the targets. That is, ten Brinke et al. (2014) found a congruency effect between primes and targets in the indirect task. In the same way, studies using the standard reasoning typically employ an indirect task attempting to find such congruency effects (see Figure 1). These congruency effects could be on RTs (as in the case ten Brinke et al., 2014), but also on other behavioral responses (e.g., skin conductance) or neurophysiological measures (e.g., in EEG or fMRI).

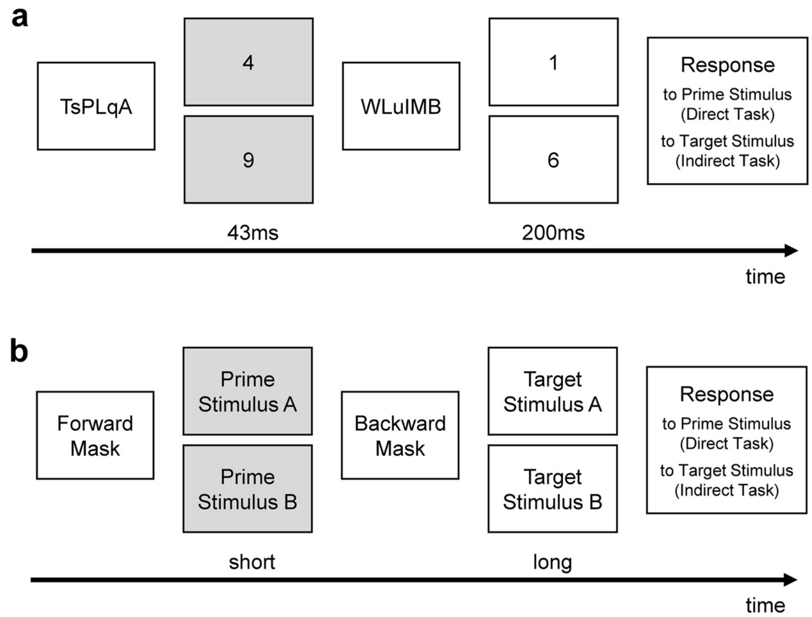
Taken together, ten Brinke et al. (2014) found the typical pattern of results for the unconscious priming paradigm: (a) a poor accuracy, or sensitivity in the direct task and (b) a clear congruency effect in the indirect task. Based on this pattern, they concluded that participants' indirect task revealed more accurate lie detection than the direct task: "[I]ndirect measures of deception detection are significantly more accurate than direct measures" (p. 1098; Abstract). In the same way, studies using the standard reasoning infer from such a pattern of results better sensitivity for the primes in the indirect task than in the direct task (see Figure 2). We dubbed this situation the *indirect task advantage*, or *ITA*. It is important to note that the claim of an ITA is, in this phase of the reasoning, independent of any considerations about conscious or unconscious processing. We call this descriptive phase of the standard reasoning *Step 1*.

In *Step 2* of the standard reasoning, ten Brinke et al. (2014) used the presumed ITA to conclude superior unconscious processing: "[A]lthough humans cannot consciously discriminate liars from truth tellers, they do have a sense, on some less-conscious level, of when someone is lying" (p. 1103). The authors thereby followed the typical assumption that direct and indirect tasks measure conscious and unconscious processing, respectively. Based on the supposed ITA from Step 1, these assumptions lead to the typical conclusion that participants processed the category of the masked stimuli better unconsciously than they can consciously report.

The standard reasoning is summarized for example by Dell'Acqua and Grainger (1999): "The present work follows the tradition of providing evidence for a dissociation between direct and indirect effects of unconsciously presented stimuli (Draine & Greenwald, 1998; Greenwald et al., 1995). More specifically, null effects are sought in direct measures (i.e., where subjects respond directly to the unconsciously presented stimuli) accompanied by non-null indirect effects (i.e., priming effects)" (p. B2). For further description of the standard reasoning see also Merikle (1992) and Simons et al. (2007). Even though some studies may not state an ITA as explicitly as shown here, it is nevertheless necessarily implied when claims about unconscious processing are made because Step 1 is a necessary condition for Step 2.

But note that the standard reasoning infers better sensitivity in the indirect task than in the direct task (i.e., an ITA) without ever

Figure 1
Typical Study Design to Infer an Indirect Task Advantage (ITA)



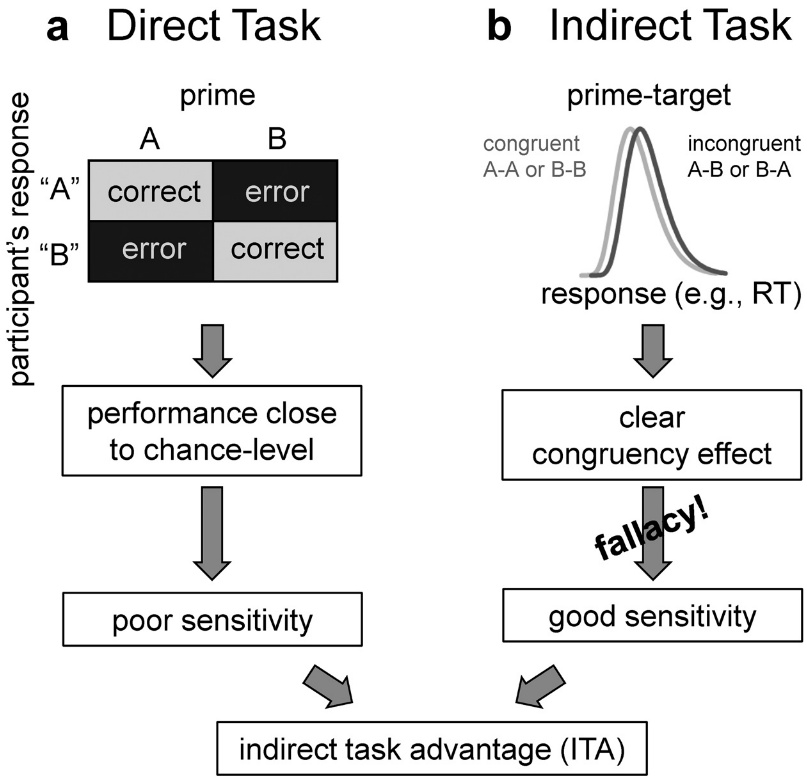
Note. (a) Example study: The study of Dehaene et al. (1998) is a prototypical example for unconscious priming with number stimuli. In each trial, a masked prime stimulus is presented for a short duration followed by a well visible target stimulus. In the direct task, participants discriminated the primes and performance was close to chance level. In the indirect task, participants discriminated the target stimuli by deciding whether they were smaller or larger than the number 5. Reaction times (RTs) were faster and lateralization of brain activity was larger when prime and target stimuli were congruent (both smaller or both larger than 5) than when they were incongruent (one larger one smaller). Dehaene et al. (1998) followed the standard reasoning to infer a higher sensitivity for the primes in the indirect task than in the direct task (i.e., an ITA) and conclude that the primes were processed in the absence of conscious awareness. (b) General design: In general, prime and target stimuli each come from one of two categories, A or B. In the direct task, participants discriminate the prime (e.g., guess whether it is from Category A or B) with a poor sensitivity. In the indirect task, the same stimuli are presented and participants now discriminate the target. In this task, the prime is shown to influence responses resulting in faster RTs for congruent (A–A, or B–B) than incongruent trials (incongruent: A–B, or B–A). From this pattern of results, the standard reasoning infers an ITA (cf. Figure 2).

calculating sensitivity (or accuracy) in the indirect task to compare against that in the direct task. For example, ten Brinke et al. (2014) only demonstrated a congruency effect on RTs. *However, if this congruency effect indicated accurate unconscious lie detection, we should be able to use the RT data to determine which of the suspects were lying with a higher accuracy than in the direct task.* Otherwise the congruency effect does not truly provide evidence for better accuracy in the indirect than in the direct task (i.e., for an ITA).

Because ten Brinke et al. (2014) laudably followed an open-data policy, Franz and von Luxburg (2015) were able to reanalyze how much evidence the RT data truly provided for better accuracy in the indirect than in the direct task. To assess this, they determined statistically optimal classifiers, used the RT of each trial in the indirect task to classify ("predict" in the nomenclature of statistical learning) which of the suspects were lying, and found the accuracy in the indirect task to be only at 50.6%-correct (SEM = .3%; see below for more details on the methods used). This value is very similar to—and not significantly different from—the accuracy in the direct task (which was 49.62%-correct; SEM = 1.4%). Therefore, ten Brinke et al.'s (2014) inference in Step 1 was flawed: Their data did not provide evidence for better accuracy in the indirect than in the direct task. In our words, there was no evidence for an ITA. Because the existence of an ITA in Step 1 is a necessary condition for Step 2 of the standard reasoning, inferences about unconscious processing were not warranted.

In the following section, we show in detail why claiming an ITA based on the standard reasoning is flawed. Note, that our critique focuses on how an ITA is established in Step 1 and is therefore independent of any assumptions about conscious versus unconscious processing, which are relevant only in Step 2 and for which different, sometimes contentious approaches exist (e.g., Eriksen, 1960; Erdelyi, 1986; Holender, 1986; Reingold & Merikle, 1988, 1990;

Figure 2
Standard Reasoning to Infer an Indirect Task Advantage (ITA)



Note. (a) In the direct task, the standard reasoning infers from close-to-chance performance that there was poor sensitivity for the primes, if any at all. (b) In the indirect task, the standard reasoning infers from a clear congruency effect that sensitivity was relatively good. Based on this pattern of results the standard reasoning makes two inference steps: In Step 1, it incorrectly infers that participants' responses in the indirect task were more sensitive to the primes than responses in the direct task (ITA). In Step 2, it attributes this difference to unconscious processing. However, already Step 1 of this reasoning is flawed because a clear congruency effect does not necessarily indicate good sensitivity. It could be caused by a sensitivity as poor as (or even worse than) the sensitivity in the direct task! Because Step 1 is independent of any (sometimes contentious) assumptions about conscious versus unconscious processing, our critique is also independent of any such assumptions.

Schmidt & Vorberg, 2006). We avoid these discussions by focusing on an empirical investigation of Step 1 which makes our critique very general.

The Standard Reasoning Is Flawed and Fails to Provide Evidence for an ITA

The standard reasoning is intuitively very appealing, which seems to be one reason for its popularity. The colloquial version of the arguments to infer an ITA in Step 1 goes like this: "Participants have a very hard time to discriminate the masked stimuli in the direct task. They are very close to zero sensitivity and usually not significantly above chance. Nevertheless we find clear and highly significant congruency effects in the indirect task. Therefore, it seems obvious, that the indirect task responses are more sensitive to the masked stimuli than the direct task responses."

However, this intuition is misguided. To see this, consider what happens if we increased the number of observations (number of participants or trials). The poor sensitivity in the direct task (Figure 2a) will only be measured more precisely but will still be poor. In contrast, the congruency effect in the indirect task (Figure 2b) becomes clearer because it is based on the difference between congruent and incongruent condition means: With more observations, the variability of the two means becomes smaller, such that the difference between them becomes clearer. Therefore, a clear congruency effect can be generated by a good underlying sensitivity (corresponding to, say, $d' = 5$ or 99% correct) but it can also be generated by a very poor sensitivity (say, $d' = 0.05$ or 51% correct). In cases where the sensitivity of the indirect task is as poor as in the direct task, there is no ITA and further interpretations about unconscious processing are unwarranted. Not recognizing this is the *main fallacy of the standard reasoning*. We demonstrate this problem by using a toy example.

Toy Example With Baby Weights

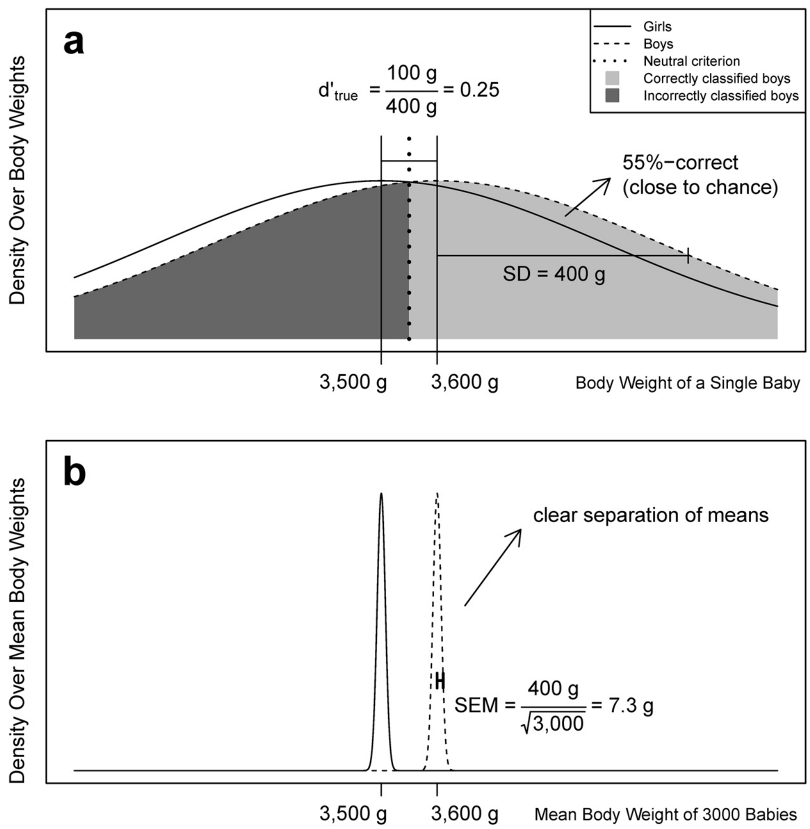
To illustrate the problem of the standard reasoning, consider an example in which responses in the direct and indirect tasks are based on the *exact same* underlying sensitivity. Nevertheless, the standard reasoning would erroneously infer that responses in the indirect task were *more* sensitive than responses in the direct task (i.e., an ITA). Consider participants measured the birth weights of newborn girls (Category A) and boys (Category B), such that they only knew the weight of the babies but not the biological sex. This would be all the information participants had in both, direct and indirect, tasks.

In the direct task, participants would use this weight information to guess whether a baby is a girl or a boy (newborn girls weigh a little less than boys). Due to the large overlap between the weight distributions (Figure 3a), participants would be correct in only approximately 55% of the cases even when using an ideal decision criterion. This corresponds to a poor performance that is close-to-chance level (50%). Following the standard reasoning, an experimenter would correctly infer a poor sensitivity in this direct task (Figure 2a).

In the indirect task, participants would simply report the numerically measured weight of the babies. The experimenter would *average* those responses across groups of baby girls and

Figure 3

Toy-Example Demonstrating Fallacy of Standard Reasoning



Note. We show that even when responses in the direct and indirect tasks are based on the *exact same* information the standard reasoning would nevertheless infer an indirect task advantage (ITA): a *higher* sensitivity in the indirect as compared with the direct task. Consider participants of a hypothetical experiment measured the birth weight of babies but did not know the babies' sex. (a) In the *direct task*, participants used the weight of an individual baby to guess whether it is a girl or a boy. The weight distributions overlap heavily such that sensitivity would be poor ($d'_{\text{true}} = 0.25$; corresponding to 55% correct). (b) In the *indirect task*, participants responded by simply stating the measured weights. The experimenter would average those responses across many trials (e.g., across 3,000 girls and 3,000 boys). The resulting group means are much less variable than the individual weights such that the experimenter would obtain a clear difference between the two group means (this corresponds to a clear congruency effect in the priming paradigm). Based on this result, the standard reasoning would erroneously infer that participants had relatively good sensitivity about whether a baby was a girl or a boy in the indirect task—better than in the direct task. That is, the standard reasoning would infer an ITA even though the *exact same* information created the responses in both tasks. Weight–data based on Janssen et al. (2007).

boys and would calculate the difference of the *mean* responses to those two groups. With increasing group sizes, the experimenter would eventually find a clear difference (corresponding to the clear congruency effect in the priming paradigm). Figure 3b shows this for 3,000 observations in each group, which is a typical number of observations in the indirect task (e.g., when 10 participants perform 300 trials in each condition, the number of observations per condition is $10 \times 300 = 3{,}000$). Following the standard reasoning, the experimenter would incorrectly infer a good sensitivity in this indirect task (Figure 2b).

Here is the catch: The standard reasoning would incorrectly interpret this pattern of results as evidence for better sensitivity in the indirect task than in the direct task (i.e., for an ITA). However, this inference is wrong because participants gave responses in both tasks based on exactly the same information: In both tasks they knew only the weight of the babies. The illusion of an ITA is generated by the different data-analysis strategies of the experimenter in the two tasks and by the fact that the experimenter never attempted to estimate the sensitivity in the indirect task.

Further Details on the Standard Reasoning

We have shown that the standard reasoning is flawed because it infers an ITA in Step 1 even when there is none. The problem is that the standard reasoning calculates two very different things in the direct and indirect tasks: In the direct task, it calculates how well each stimulus can be classified on a trial-by-trial level. In the indirect task, it assesses whether there is a difference in mean responses. These are two very different things and it is a priori to be expected that the sensitivity in single trials can be poor while mean responses can nevertheless be clearly separated between the two categories given enough trials. A more appropriate analysis to determine whether there is an ITA would need to estimate sensitivities in both tasks and compare them. Before we present such an analysis, we want to first discuss some details of the standard reasoning.

True Zero-Sensitivity in the Direct Task

Consider that the true sensitivity in the direct task were known to be exactly zero and that there were at the same time a clear congruency effect in the indirect task. This ideal situation is typically sought—but typically not fully achieved—in the dissociation paradigm (Hannula et al., 2005; Schmidt & Vorberg, 2006; Simons et al., 2007). In this case (and only in this case), the standard reasoning would be justified in claiming that responses in the indirect task were somehow more sensitive than responses in the direct task. This is so, because a positive (larger than zero) sensitivity—even if it is minute—is required to produce a congruency effect and therefore the indirect task sensitivity must be larger than zero. However, there are a number of problems with this scenario: (a) It is unrealistic. Typically, studies either find some small, residual sensitivity in the direct task or they do not find a congruency effect (Zerweck et al., 2021). (b) One cannot be certain of a true zero sensitivity. Instead, sensitivity in the direct task always needs to be measured (and is therefore affected by measurement error). Thus, we would still need to establish that the sensitivity in the indirect task is indeed larger than that in the direct task (e.g., by a significance test on the difference). (c) The sensitivity in the indirect task could still be so low, that it would be close–enough to the zero sensitivity of the direct task to not allow for strong conclusions (e.g., consider a sensitivity that corresponded to 50% correct in the direct task and to 51% correct in the indirect task).

Significance Testing Versus Bayesian Methods

Until now, we purposefully did not talk about statistical significance testing because we wanted to focus on the main fallacy of the standard reasoning. Because significance testing and its applications have been heavily—and often rightfully—criticized since the very inception of the concept (Boring, 1919; Cumming, 2014; Dienes, 2011; Morrison & Henkel, 1970); it might be tempting to attribute the main fallacy of the standard reasoning also to significance testing. However, the problem of the standard reasoning is not so much that the statistical tools were wrong, but that the wrong statistical question is asked for the indirect task: The standard reasoning asks whether there is a true difference in means between congruent and incongruent conditions. However, the correct question to ask would be what the sensitivity in the indirect task is and whether this sensitivity is higher than in the direct task (such that an ITA can be concluded). Therefore, it would not help to simply replace the frequentist significance testing by Bayesian methods. Because researchers are interested in establishing an ITA (i.e., a difference in sensitivities) it does not suffice to evaluate both tasks in isolation. We must test directly for a difference in sensitivities between the two tasks. Failure to do so can lead to serious errors no matter whether we used significance testing (cf. Appendix B of Franz & Gegenfurtner, 2008 and Nieuwenhuis et al., 2011) or Bayesian methods (cf. Supplement F and Palfi & Dienes, 2020).

Direct Task is Typically Underpowered

An additional problem in the application of the standard reasoning arises from the widespread use of seriously underpowered direct tasks (Buchner & Wippich, 2000; Vadillo et al., 2016; Vadillo et al., 2020). When the direct task is sampled with fewer participants and trials than the indirect task (as is often the case), a nonsignificant direct task result may not indicate that the true sensitivity is close to or exactly zero but rather that statistical power is low. Moreover, participants are required to give binary responses in the direct task in contrast to the continuous measures in the indirect task (e.g., RTs). Because participants have some continuous sense (confidence) about their responses (Rausch et al., 2018; Zehetleitner & Rausch, 2013), the binary response format forces them to discard this information (Cohen, 1983), which further decreases the statistical power in the direct task. Therefore, even if the same sensitivity underlies responses in both tasks, it is a priori to be expected that the direct task produces less often significant results than the indirect task.

Appropriate Analysis: Calculate Sensitivities and Test for a Difference

We have shown that the standard reasoning is flawed and that researchers must compare sensitivities of both tasks if they want to infer an ITA. In this section, we describe more appropriate analyses. First, we assume that trial–by–trial data are available (this analysis was used by Franz & von Luxburg, 2015). Then we describe our newly developed method to reanalyze studies when only summary statistics are available. For detailed mathematical derivations see the online supplementary materials.

In deriving our methods, we unavoidably were confronted with degrees of freedom when choosing the details of our analysis strategy. In these cases, we chose strategies that favored finding an ITA. That is, we followed a *benefit-of-the-doubt* approach, thereby increasing the chances of confirming an ITA. We adopted this approach because we are criticizing a large body of literature. Therefore, it seemed necessary and reasonable to adopt such a liberal bias in confirming ITAs (and thereby being conservative in our critique) at this stage of the scientific discussion. It is understandable that researchers who have spent years using the standard reasoning might be reluctant to accept our arguments if our methods were too restrictive. This approach makes our arguments even stronger when we nevertheless do not find evidence for ITAs.

Sensitivity Comparison When Trial-By-Trial Data Are Available

The appropriate method directly compares sensitivities in the direct and indirect tasks. Different than the standard reasoning, the appropriate analysis equates analysis steps for both tasks such that the calculated statistics are comparable. Then, a test of the difference between the two tasks is applied. Similar approaches have been used in previous—albeit very few—studies (Dulaney & Eriksen, 1959; Franz & von Luxburg, 2015; Klotz & Neumann, 1999; Kunst-Wilson & Zajonc, 1980; Schmidt, 2002) in accordance with the long standing (but often ignored) request for both tasks to be measured using the "same metric" (Reingold & Merikle, 1988).

In both tasks, we compute d' using signal detection theory (Green & Swets, 1988) and then test for a difference between them. In the *direct task*, participants typically classify the primes in each trial and a d' value is often already reported by the studies using the standard reasoning. In the *indirect task*, however, the standard reasoning computes a congruency effect on continuous measures (e.g., RTs or brain activity as measured by EEG or fMRI). For a proper comparison, we have to transform these continuous measures into classifications (predictions) for each trial. There are different ways to achieve this. We suggest to use the optimal classifier for the given setup. This gives the indirect task the best possible performance and increases the chances of finding an ITA following the *benefit-of-the-doubt* approach.

Which classifier is best? We have shown that under typical conditions with equal number of congruent and incongruent trials, the median-split classifier is optimal (Franz & von Luxburg, 2015; see our Supplement B for details and proof). The classification proceeds as follows: For each participant, we determine the median RT and classify ("predict" in the nomenclature of statistical learning) all trials with smaller RTs as congruent, and trials with larger RTs as incongruent. Then, we compare these classifications to the true labels (congruent/incongruent) evaluating for each trial whether the classification was correct or not, and we then compute a d' value as in the direct task. Finally, we compare the d' values of the direct and indirect task and test for a difference.

Some details: (a) Instead of computing d' values, the analysis could also be based on % correct values. Assuming a neutral observer predicting both categories equally often in the direct task, both approaches produce the same results and we later report both measures to foster intuition. (b) Dichotomization of the continuous, indirect measures will result in a loss of information (Cohen, 1983). However, the direct task also requires participants to give binary responses. Converting indirect task responses into a binary response format using our median split approach only equates this dichotomization to make responses in both tasks comparable. (c) We classify the trials of the indirect task according to the labels congruent/incongruent and not according to the prime Category A/B, as is typically asked in the direct task. This is so because studies typically find a congruency effect between prime and target (and not a mere effect of the prime being in Category A or B). For a comparison to the direct task, we would ideally transform the congruency classification into a classification of the prime category (A vs. B). For simplicity, we assume an optimal transformation here (without errors). This is plausible, because the target stimuli are typically fully visible to the participants, such that errors are rare. Again, our approach increases the chances of finding an ITA following the benefit-of-the-doubt approach.

Sensitivity Comparison When Only Summary Statistics Are Available

Because the standard reasoning to infer an ITA is flawed, many already published studies on unconscious priming need reassessment. However, the appropriate analysis as described in the previous section would require full trial-by-trial data. Unfortunately, trial-by-trial data can be difficult or impossible to obtain for published studies (Wicherts et al., 2006). For the older—but nevertheless influential—studies, those data might not even exist anymore. Therefore, we developed an approach that allows to estimate the results of the appropriate analysis without access to trial-by-trial data and solely based on the typically reported statistics from the standard reasoning. Here, we sketch the central approach of this analysis; details are given in Supplement C.

The overall aim of this reanalysis is, again, to estimate sensitivities for the direct and indirect tasks (i.e., to either calculate d' from signal detection theory or % correct assuming a neutral observer). The direct task typically already provides d' or % correct values. In the indirect task, studies typically report t or F values from a repeated measures design for the congruency effect. In this design, we show how F values can be translated to t values. We then derive an estimator for the underlying sensitivity that takes the form of a constant c_{N,K,q^2} multiplied onto the reported t value. This constant will include the number of participants N and trials K from the indirect task because t values become larger the more observations are made. Additionally, because this reanalysis can only use the reported statistics, one free parameter needs to be estimated: the ratio of between- versus within-subject variances, which we denote by q^2. We estimated this parameter based on (a) our own replication experiment, (b) a literature review, and (c) extensive simulations (see Supplement D). By assuming the largest plausible value for q^2, we again maximize the estimated sensitivity, d', in the indirect task and therefore increase the likelihood of confirming an ITA. Here, we again follow the benefit-of-the-doubt approach.

Replication of a Landmark Study Finds No ITA

We are now equipped with the appropriate tools that allow us to analyze typical settings and tasks that have been investigated in the context of unconscious priming. In this section, we will focus on one highly influential study on unconscious semantic priming of numbers (Dehaene et al., 1998). We will first describe the study and how its conclusions depend crucially on the flawed standard reasoning. Then, we will describe a replication experiment of the behavioral part of this study and analyze the trial-by-trial data. In the next section, we will then reanalyze the published results of this and other studies (15 in total). Overall, we will conclude that the results of our replication are similar to those of the original study. Both, our replication and our reanalysis of the original study, give reason to seriously doubt the existence of an ITA, questioning the authors' interpretation in the original study.

Dehaene et al. (1998) were interested in the question of whether the semantic meaning of numbers can be processed outside conscious awareness. They employed a prototypical priming experiment with stimuli shown in Figure 1a and applied the standard reasoning: In the direct task, participants discriminated features of masked numbers and performed poorly (d' = .2; corresponding to 54% correct). In the indirect task, participants were again presented with the masked numbers (now serving as primes), but decided whether subsequent target numbers were smaller or larger than five. Participants responded approximately 24 ms faster when prime and target were congruent (both larger or smaller than five) than when they were incongruent (one smaller and one larger than five). Similar congruency effects were found for brain activity in EEG and fMRI (i.e., larger lateralization of brain activity in congruent than incongruent trials).

Dehaene et al. (1998) interpreted these results according to the standard reasoning: In *Step 1*, they inferred an ITA. That is, higher sensitivity in the indirect task than in the direct task: "[participants] could neither reliably report [the prime's] presence or absence nor discriminate it from a nonsense string [. . .] Nevertheless, we show here that the prime is processed to a high cognitive level" (p. 597). In *Step 2*, they argued that "the prime was unconsciously processed" (p. 597) because participants were at chance performance in the direct task. Overall, they concluded: "By showing that a large amount of cerebral processing, including perception, semantic categorization and task execution, can be performed in the absence of consciousness, our results narrow down the search for its cerebral substrates" (p. 599). In short, Dehaene et al. (1998) employed a prototypical version of the standard reasoning to infer an ITA and unconscious processing exactly as described above. To assess the validity of these claims, we first replicate the behavioral part of that study, later we will reanalyze the published data.

Disclosures

Data, Materials, and Online Resources

The experimental material, data, and the scripts for the analyses reported in this article have been made available on the Open Science Framework (OSF), at https://osf.io/kp59h; Meyen et al., 2021. We also provide an online tool to apply our reanalysis to other data at http://www.ecogsci.cs.uni-tuebingen.de/ITAcalculator/.

Reporting

We report how we determined our sample size, all data exclusions, and all measures in the study.

Method

Twenty-four volunteers participated in our study (13 female, five left-handed, age range: 19–27 years; M = 21.5, SD = 1.9). All had normal or corrected-to-normal vision, signed written informed consent and were naive to the purpose of the experiment. In the original study by Dehaene et al. (1998), six and seven participants took part in the first and second direct task, respectively, and 12 participants took part in the indirect task.

We took great care to make stimuli and timings as similar as possible to those of the original study. Each trial consisted of: fixation cross (417 ms), forward mask (67 ms), prime (42 ms), backward mask (67 ms), and target (200 ms). In the original study, those values were: forward mask (71 ms), prime (43 ms), backward mask (71 ms), and target (200 ms; cf. Figure 1a). Slight differences in timing are due to slightly different refresh rates of the monitors used. The prime duration of 43 ms was chosen by the original authors because it was the longest duration that produced nonsignificant results in the direct tasks. Primes and targets were numbers (1, 4, 6, or 9) that were either presented as digit (e.g., "1") or word (e.g., "EINS"; German for "ONE"). The original study used the same numbers in English, a follow–up used French (Kouider & Dehaene, 2009). As in the original study, primes and targets could be congruent (both smaller or both larger than five) or incongruent (one smaller, one larger). Masks were composed of seven randomly drawn characters from {a–z, A–Z} mirroring the original study's masks. Participants were seated in front of a monitor (VIEWPixx/3D, VPixx Technologies Inc., Montreal, Canada), effective refresh rate 120 Hz at a viewing distance of approximately 60 cm in a sound- and light-protected cabin. In the original study, the monitor was a cathode-ray tube (CRT) with a refresh rate of 70 Hz. Stimuli were presented centrally as white text (69 cd/m^2; character height: 1°; width: .5° visual angle; font: Helvetica) on a black background (.1 cd/m^2). These luminance values were not specified in the original study so that we chose the most plausible settings for our experiment.

In the direct task, participants classified whether the prime was smaller or larger than five. We used this particular task because the original authors argued in a subsequent study that it is "better matched with the [indirect] task" (Naccache & Dehaene, 2001, p. 227). In the original study by Dehaene et al. (1998), two direct tasks were used, that produced similar results: In their first direct task, the prime stimulus was omitted in some trials and participants had to discriminate their presence versus absence. In the second direct task, the prime stimuli were replaced by random letter strings and participants had to discriminate between numbers versus random strings.

In the indirect task, participants decided as quickly as possible whether the target was smaller or larger than five; as was the case in the original study. Each participant performed 256 trials per task, preceded by 16 practice trials in each task. In contrast, in the original study, participants performed only 96 and 112 trials in the first and second direct tasks, respectively, and 512 trials in the indirect task.

We repeated indirect task trials with RTs that were too slow (>1 s) or too fast (<100 ms). The original study also rejected too slow trials (>1 s) but was more restrictive in terms of fast trials: They rejected with RT <250 ms. However, we only found eight out of 6,144 trials in our data to be above 100 ms but below 250 ms so that we obtained very similar results when applying their criterion. The indirect task was performed before the direct task (as is common practice in this paradigm) to prevent participants attending to the prime in the indirect task. In the original study, the direct and indirect tasks were performed by different groups.

The number of participants and trials were chosen to produce a statistical power of above 95% to find a difference between sensitivities and confirm an ITA if it is there (see Supplement A). For this power estimation, we assumed a true sensitivity of $d'_{\text{true, direct}} = 0$ in the direct task versus $d'_{\text{true, indirect}} = .25$ in the indirect task (values based on our reanalysis of Dehaene et al., 1998; see below). To our knowledge, the original study did not perform a power analysis. A post hoc power analysis revealed that the original study had a statistical power of only 46% to find an ITA using the appropriate analysis (again, assuming $d'_{\text{true, direct}} = 0$ and $d'_{\text{true, indirect}} = .25$). This low power is due to a small number of direct task participants and trials.

Results and Discussion

Our analysis proceeded in two strands: First, we perform the traditional analysis which forms the basis for the standard reasoning. Second, we perform the appropriate analysis.

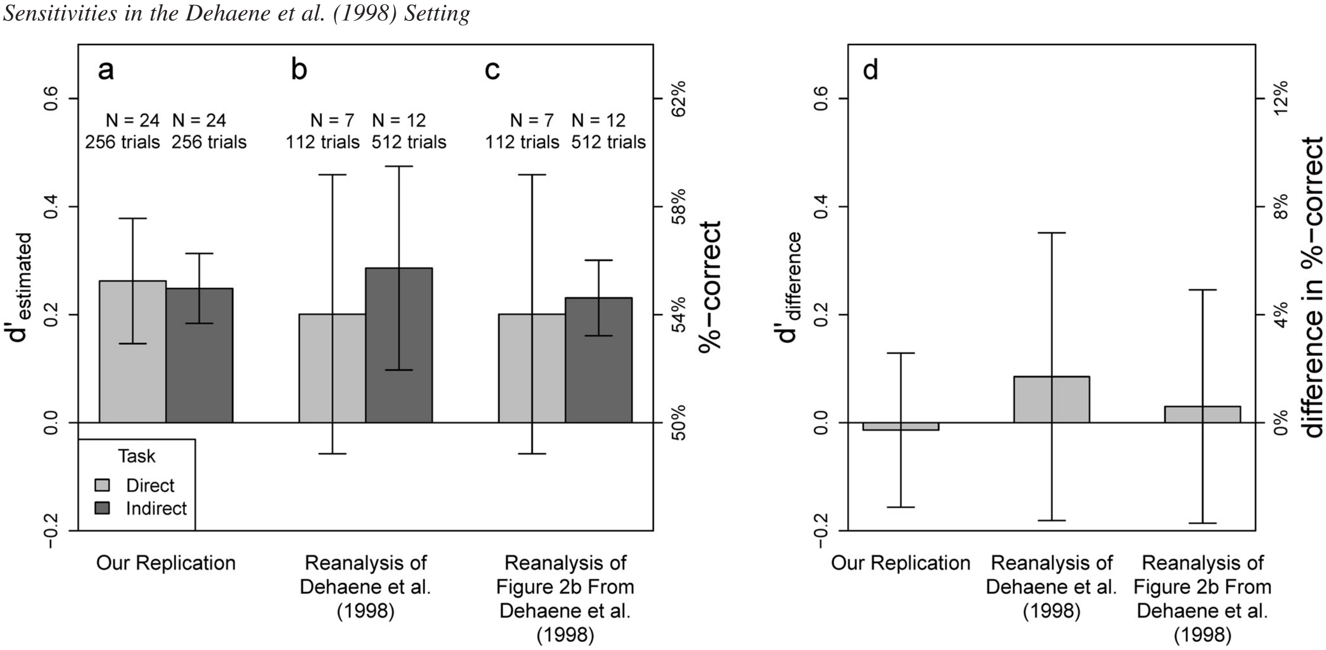
Standard Reasoning

The direct task sensitivity was $d' = .26$ ($SD = .27$), $t(23) = 4.68$, $p < .001$, corresponding to an accuracy in prime identification of $M = 54.87\%$ correct ($SD = 4.9$), $t(23) = 4.82$, $p < .001$. This is exactly in the range of sensitivities reported in the original study's direct tasks ($d' = .3$ in the first and $d' = .2$ in the second direct task). For a graphical depiction of these results, compare the bars corresponding to the direct tasks in Figures 4a and 4b.

Note that, in contrast to the original study, our direct task sensitivity is significantly above zero. This is so, because we sampled much more participants and trials than in the original study. Therefore, we had much higher statistical power. To simulate the lower power of the original study, we discarded data from participants and trials to match the same number of observations as in the original study: We kept only the first $N = 7$ participants and the first 112 trials of each participant. This leads to a nonsignificant result, $d' = .31$ ($SD = .39$), $t(6) = 2.06$, $p = .085$, as was the case in the original study. Therefore, it is plausible that it was the low statistical power in the original study (and not the sensitivity being exactly at zero) that was the reason for the nonsignificant result in the direct task of the original study.

Figure 4

Sensitivities in the Dehaene et al. (1998) Setting



Note. (a) Results of our replication study, based on our full trial-by-trial data. (b) Results of our reanalysis approach based on the published statistics from Dehaene et al. (1998). (c) Reanalysis results from digitizing Figure 2b from Dehaene et al. (1998) showing histograms of indirect task's RT data. For the comparison, we used the same direct task results herein (c) as we used in (b). Comparing (a–c) we see that our replication closely matches the results of the original study. (d) Difference in sensitivities between direct and indirect tasks: There is no significant difference in sensitivities in our replication study or in our reanalyses of Dehaene et al. (1998). That is, there is no evidence for an ITA. The reanalysis result from (b) is also shown in the large summary in Figure 5. Error bars indicate 95% confidence intervals.

In the indirect task, the congruent condition yielded faster RTs (M = 445 ms, SD = 42) than the incongruent condition (M = 457 ms, SD = 37), resulting in a clear and highly significant congruency effect of M = 12 ms (SD = 11.8), $t(23) = 4.95$, $p < .001$. That is, we found a highly significant congruency effect on RTs, as did the original study.

There is one potential caveat here: The congruency effect in the original study was larger than that in our replication (24 ms vs. 12 ms, respectively). However, we will show that sensitivities in our replication and our reanalysis of the original study are very consistent, see Figures 4a and 4b. This can be explained by larger trial-by-trial variability in the original study counteracting the larger RT effect: The original study, despite using 512 trials per participant, observed SD = 13.5 while we observed SD = 11.8 in our replication with only 256 trials per participant. Generally, more trials per participant should make individual RT effects more precisely measured. Thus, the standard deviation in the original study should be smaller than in our replication. But the opposite is the case! This can be explained by a larger trial-by-trial variance in the original study. Larger effect and larger variability in the original study cancel out such that sensitivities are in fact quite comparable to our replication, see below. Further research employing systematic variation of stimulus parameters can further clarify this situation. For example, we are currently determining the role of an ITA in the particular setting of Dehaene et al. (1998) in a more extensive study, see Zerweck et al. (2021).

In summary, we found a similar pattern of results as in the original study: A very poor direct task performance and a clear congruency effect in the indirect task. Based on this pattern of results many researchers would have applied the standard reasoning and inferred an ITA.

Sensitivity Comparison

The appropriate analysis compares sensitivities in direct and indirect tasks. We have already described in the last section that the direct task in our experiment yielded a sensitivity of $d' = .26$ (SD = .27), corresponding to an accuracy of M = 54.87% correct. For the indirect task, we obtained a sensitivity of $d' = .25$ (SD = .15), corresponding to an accuracy of M = 54.93% correct (SD = 3.03).

Inspection of Figure 4a shows that these sensitivities in direct and indirect tasks are very similar, see their difference plot in Figure 4d. We found virtually no difference between these sensitivities, $M = -.01$ (SD = .34), $t(23) = -.2$, $p = .844$. That is, there is no indication for an ITA.

In conclusion, our results are similar to the typical pattern of results found by Dehaene et al. (1998) and many researchers would have inferred an ITA. However, the appropriate analysis yields no evidence for an ITA: The sensitivities in both tasks are essentially identical.

Reanalysis of 15 Influential Studies Finds Hardly Any ITA

After having demonstrated that the problems of the widely used standard reasoning are indeed serious, we now apply our approach to a sample of 15 highly relevant studies in the field of unconscious priming.

Method

Selection Criteria for Reanalyzed Studies

We focused on studies that applied the standard reasoning and claimed an ITA. First, we selected eight studies by hand that are particularly relevant. These studies and their number of citations in Web of Science (Clarivate Analytics, Philadelphia, PA, USA) are: Finkbeiner and Palermo (2009, 56 citations); Finkbeiner (2011, 13 citations); Mattler (2003, 76 citations); Pessiglione et al. (2007, 352 citations); Sumner (2008, 34 citations); van Gaal et al. (2010, 154 citations); Wang et al. (2017, zero citations); Wójcik et al. (2019, one citation),

Second, we searched for English articles in Web of Science with the topic “unconscious priming.” We selected all studies with more than 150 citations that applied the standard reasoning and claimed an ITA. This resulted in seven additional studies: Damian (2001, 178 citations); Dehaene et al. (1998, 662 citations); Dehaene et al. (2001, 770 citations); Kiefer (2002, 237 citations); Kunde et al. (2003, 217 citations); Naccache and Dehaene (2001, 214 citations); Naccache et al. (2002, 313 citations). Overall, these 15 studies received a total of 3,277 citations. See Supplement E for details on these studies.

Details of Analysis When Only Summary Statistics Are Available

Our reanalysis method estimates and compares sensitivities for direct and indirect tasks. Here, we sketch some technical details of the analysis. A detailed account with mathematical derivations is given in Supplement C. We denote the estimated sensitivities in the direct and indirect tasks by $d'_{\text{estimated,direct}}$ and $d'_{\text{estimated, indirect}}$, respectively. For the direct task, the typically reported statistics are average d' or % correct values. Therefore, our estimate is simply the measured sensitivity,

$$d'_{estimated,direct} = d',$$

or a well-known conversion of % correct values to d' values assuming neutral observers (Green & Swets, 1988),

$$d'_{estimated,direct} = 2\Phi^{-1}(\%\,\text{correct}),$$

where Φ^{-1} is the inverse of the normal cumulative density function.

In the indirect task, statistics for the congruency effect are typically reported by t values from a paired t test or F values from a repeated–measures ANOVA. In this setting, F values can be translated into t values by $|t| = \sqrt{F}$. From a t value, we estimate the sensitivity by

$$d'_{estimated,indirect} = t \cdot c_{N,K,q^2} \text{ with } c_{N,K,q^2} = \sqrt{\frac{q^2 + 4/K}{N}}\sqrt{\frac{2}{N-1}}\,\frac{\Gamma\left(\frac{N-1}{2}\right)}{\Gamma\left(\frac{N-2}{2}\right)}.$$

where Γ is the gamma distribution. The constant c_{N,K,q^2} corrects for the fact that t values increase with increasing number of participants (N) , increasing number of trials (K), and that they depend on the ratio of between- and within-subject variance, which we denote by q^2.

The parameter q^2 is the only free parameter we need to estimate for our approach. It is reasonable to assume that this ratio is at most $q^2 = .0225$ given our replication study, a literature review (see Supplement D) and extensive simulations (see Supplement A). Assuming the largest plausible value for q^2, increases the likelihood of finding an ITA thereby following the benefit-of-the-doubt approach.

From the estimated sensitivities, we compute the difference

$$d'_{difference} = d'_{difference,\,indirect} - d'_{estimated,direct}$$

and construct a 95% confidence interval using the corresponding standard errors (derived in Supplement C). This allows to test for an ITA: If the confidence interval lies above 0 (that is, it has the form [a, b] with $a > 0$), the reported result is significant and an ITA is confirmed, otherwise there is not sufficient evidence to claim an ITA.

We demonstrate in Appendix that confidence intervals based on our reanalysis method are quite comparable with those based on the trial-by-trial analysis. For the study of ten Brinke et al. (2014), the trial-by-trial analysis versus our reanalysis method using summary statistics produced 95% CI [−.07, .23] and [−.11, .25], respectively (Figure A1). In our replication, the two methods produced 95% CI [−.15, .12] and [−.20, .06], respectively (Figure A2). Thus, our reanalysis method produces consistent results with the analysis based on trial-by-trial data.

Results and Discussion

We first describe our reanalysis in detail for the study of Dehaene et al. (1998) and then use the same methods for all the other studies.

Reanalysis of Dehaene et al. (1998)

As discussed in our replication, the study reported two direct tasks with sensitivities of $d' = .2$ and $d' = .3$, respectively. We used the results of the first task, because it had the smaller sensitivity, thereby, increasing the chances of our reanalysis to confirm an ITA and following the benefit-of-the-doubt approach.

In this direct task, $N = 7$ participants were sampled in $K = 112$ trials and a sensitivity of $d' = .2$ was reported, see light gray bar in Figure 4b. From these values, our reanalysis method estimates the standard error to be $SE = .11$.

In the indirect task, the study reported on average a congruency effect of 24 ms with a standard deviation of 13.5 ms in a sample of $N = 12$ participants sampled in $K = 512$ trials each. This equals a t value of $t = 24\text{ ms}/(13.5\text{ ms}/\sqrt{12}) = 6.12$ from which our reanalysis method estimates the sensitivity to be

$d'_{\text{estimated, indirect}} = t \cdot c_{N,K,q^2} = .29$ ($SE = .09$), see dark gray bar in Figure 4b. Taken together, the sensitivities in both tasks are very similar with no clear difference between them, $d'_{\text{difference}} = .09$, $SE = .14$, see Figure 4d. The confidence interval for the difference includes zero, 95% CI [−.18, .35], thereby indicating that the sensitivity difference did not deviate significantly from zero. That is, there is no evidence for an ITA.

We were able to reanalyze the results from Dehaene et al. (1998) in an additional way. They depicted summary histograms of RTs in their Figure 2b visualizing that congruent and incongruent RT distributions are similar in shape but only shifted because incongruent RTs were slower than congruent RTs. Despite the shift, RT distributions largely overlap. We digitized the histogram and split RTs along the median as described in the appropriate analysis section. From this, we estimated the indirect task sensitivity to be $d' = .23$ ($SE = .03$). Again, we find no difference to their first direct task's sensitivity ($d' = .2$, $SE = .11$) because zero is included in the confidence interval of the difference, 95% CI [−.19, .25], see Figure 4c and 4d. Note that this approach deviated from our appropriate analysis in that it does not compute the median for each individual participant but uses a grand median across participants because the published histogram pools all participants' RT data. This approach ignores between-subjects variance leading to a slight underestimation of the indirect task's sensitivity. Nevertheless, this additional reanalysis provides converging evidence complementing our previous results.

The results from our reanalysis of the original study (Figure 4b and 4c) and the results from our replication experiment (Figure 4a) are very consistent. Estimates for the sensitivities are very stable. This corroborates the validity of our reanalysis approach as well as of our replication experiment (see Supplement A for further validation of our reanalysis approach).

To summarize, both, our reanalysis of Dehaene et al. (1998) as well as our replication of the behavioral responses, suggest that there is no ITA in the behavioral part of that study. This demonstrates the fundamental flaw of the standard reasoning and suggests that similar problems might exist in other studies.

Reanalysis of All 15 Studies

We now apply our reanalysis in a similar way to all other studies. For this, we present the data in a more compact fashion in Figure 5. For example, what we showed in Figures 4b and 4d for the study of Dehaene et al. (1998) now corresponds to the Lines 7 and 8 in Figure 5, showing the sensitivities for each task in Figure 5a and the difference of sensitivities in Figure 5b.

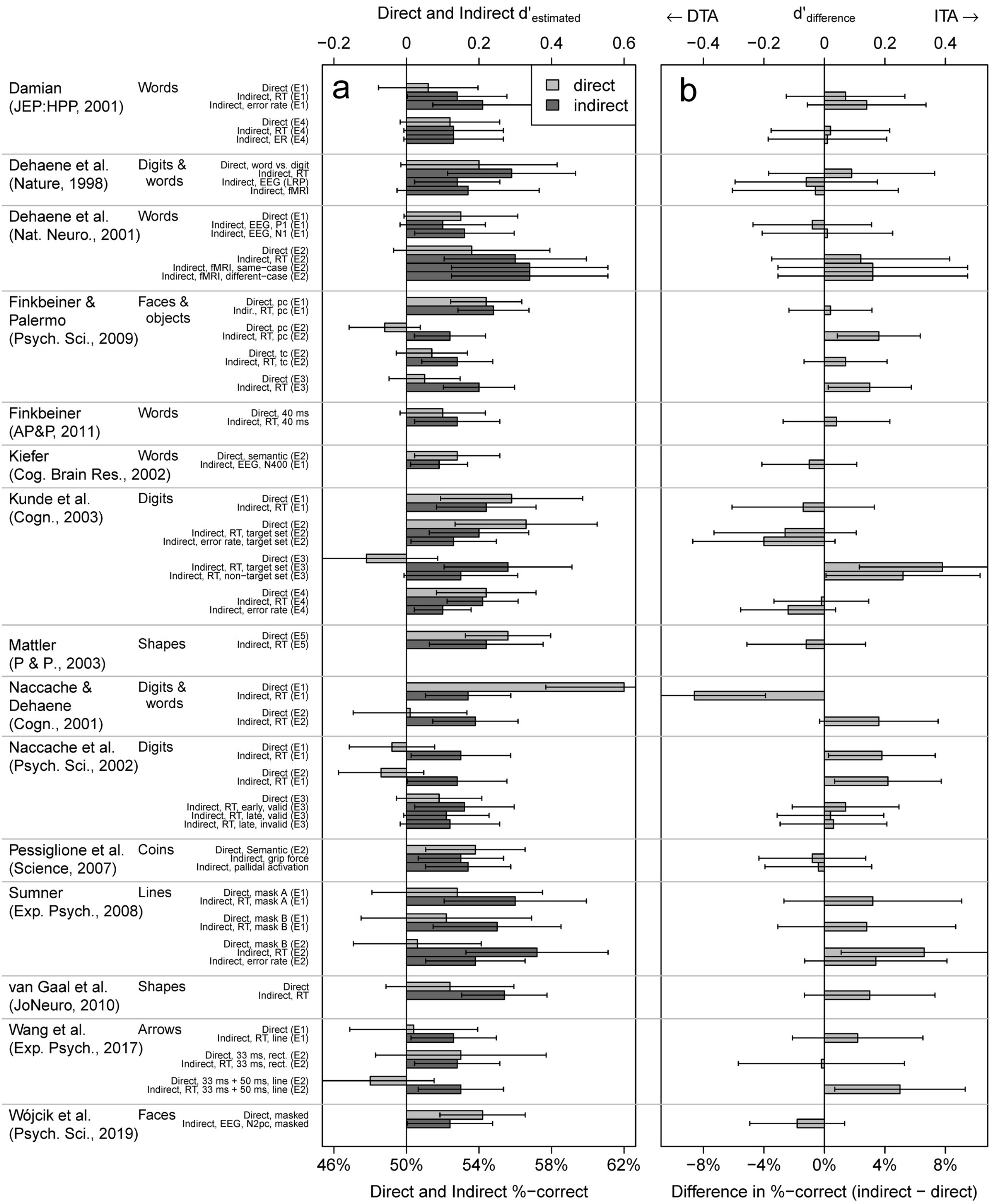
When evaluating this figure, it is important to be aware that we used our benefit-of-the-doubt approach. For example, Dehaene et al. (1998) had two direct tasks, resulting in $d' = .2$ and $d' = .3$, respectively. As described above, we used the smaller of those values, thereby increasing the chances of finding an ITA, which makes our arguments stronger if we nevertheless do not find an ITA (cf. General Discussion).

Inspecting the figure shows that in most studies the sensitivities of direct and indirect tasks have comparable sensitivities, such that the differences are small and not significantly different from zero. This is the case for 35 of the 44 differences between direct and indirect tasks (Figure 5b). This is in stark contrast to the fact that all studies claimed ITAs in all these cases.

Only in eight of the 44 differences there is a significant difference in the direction of an ITA, such that the indirect task has higher sensitivity than the direct task. These results are, however, intermixed with inverted differences in the same studies. For example, although Kunde et al. (2003) have two significant differences in the direction of an ITA, there are five differences pointing in the opposite direction within the same study (albeit those are not significantly different from zero).

Figure 5

Reanalysis of Influential Studies Reporting Indirect Task Advantages (ITAs)



Finally, the largest of all differences is even inverted: In Experiment 1 of Naccache and Dehaene et al. (2001) there is a significantly higher sensitivity in the direct task than in the indirect task, just the opposite of an ITA.

To summarize, our reanalysis found significant ITAs in only eight out of 44 instances, which are spread across five different studies (Finkbeiner & Palermo, 2009; Kunde et al., 2003; Naccache et al., 2002; Sumner, 2008; Wang et al., 2017). Note that for multiple hypothesis testing, one would expect at least some false positive results. These results are intermixed with 35 inconclusive results and even an opposite result where the direct task showed significantly higher sensitivity than the indirect task (Naccache & Dehaene et al., 2001). Inspecting Figure 5 shows that there is no consistent evidence for an ITA in any of the reanalyzed studies. Not a single study showed significant ITAs in all conditions, albeit all studies claimed ITAs for all reanalyzed conditions.

Let us stress that our goal was not to investigate whether there exists a "general" ITA across all studies with their vastly different stimuli, experimental setups, tasks, and scientific questions. Therefore, we did not perform a meta-analysis or correct for multiple testing. This had several reasons. First, our reanalysis favored finding an ITA by using our benefit-of-the-doubt approach. Second, there are additional methodological issues in the reanalyzed studies that introduce further biases, and for which we cannot correct in our reanalysis (see General Discussion). Considering these two biases toward finding an ITA, a meta-analysis could misleadingly produce the impression that there is a slight ITA present across all reanalyzed studies. An ITA might exist but perhaps only for some particular stimuli and setups. Given that the evidence for an ITA in each individual study is now in question, the research goal should be to differentiate under which conditions a reliable ITA can be obtained and under which conditions this is not possible. A meta-analysis would not serve this differentiating purpose.

In summary, reanalyzing the results from studies on unconscious priming shows that there is little to no evidence for ITAs in those studies despite them claiming ITAs for all conditions. Sensitivities in the indirect tasks are not consistently larger than sensitivities in the direct task as one would expect, given that unconscious processing was inferred using the standard reasoning that necessarily implies ITAs. This demonstrates how seriously the literature on unconscious priming is affected by the flaws of the standard reasoning.

General Discussion

Many studies on consciousness that investigate a wide range of cognitive functions are based on the flawed standard reasoning. The main fallacy occurs when the standard reasoning infers an ITA. That is, a higher sensitivity for masked stimuli in the indirect task as compared with the direct task. In an earlier reanalysis of ten Brinke et al. (2014) by Franz and von Luxburg (2015), in our replication of the behavioral part of Dehaene et al (1998), and in our reanalysis of 15 highly influential studies we found that none of these studies can overall truly claim evidence for an ITA. To the contrary, responses in the indirect task often show a similar sensitivity as compared with the direct task. This casts serious doubt on the evidence for unconscious processing that exceeds conscious reportability in these studies.

The fallacy of the standard reasoning has serious consequences for the trustworthiness of the scientific literature on consciousness. It also takes away from the appeal of many claims in the field: For example, it would be an interesting result if lie detection and semantic meaning of numbers could be processed outside of awareness. But such strong claims require substantive empirical evidence, which we did not find because the reanalyzed studies employed the flawed standard reasoning. The appropriate analysis yields results that may be considered as less exciting because—under scrutiny—participants' responses did not seem to be affected by processing beyond what they can consciously report.

Besides theoretical issues, there are also additional methodological problems that can systematically bias the results and lead to claims of an ITA even if the true underlying sensitivities in the direct and the indirect task are perfectly equal.

First, a common practice is to exclude participants with a good direct task sensitivity. The researchers' motivation here is to avoid including the subset of participants who are consciously aware of the masked stimuli. However, this practice bears the problem of regression to the mean (Barnett et al., 2005; Schmidt, 2015; Shanks, 2017). Thus, this practice is biased toward finding a smaller sensitivity in the direct task and thus biased toward finding an ITA even if there is none. Several studies in our reanalysis have this problem (Finkbeiner, 2011; Mattler, 2003; Pessiglione et al., 2007; Sumner, 2008; van Gaal et al., 2010). This can explain why these studies produced some of the largest differences in our reanalysis in Figure 5.

Second, in some experimental procedures participants have to respond to the target stimulus (indirect task) first and only then respond to the masked stimulus (direct task) all within the same trial (see Finkbeiner & Palermo, 2009; Peremen & Lamy, 2014). Because the cognitive impact of a masked stimulus decays quickly after 300 ms (Mattler, 2005; Wolfe, 1999), this procedure makes the direct task more difficult. Participants have to memorize the masked stimulus while performing the indirect task until they can

Figure 5 *(opposite) Note.* The 15 studies used the standard reasoning to infer an ITA in 44 conditions. (a) We reanalyzed the sensitivities and, to foster intuition, we also show %-correct values assuming a neutral observer. (b) We reanalyzed the difference in sensitivities: In each group of bars from (a), the indirect task is compared with the corresponding direct task yielding the differences shown in (b). Only if a confidence interval (error bars) around the difference lies to the right and does not contain 0, there is evidence for an ITA. Only in very few cases (eight out of 44), there is evidence for an ITA, while in most cases (35 out of 44) there is no evidence. There is even one case with a significant opposite result, an advantage of the direct task (DTA). Not a single study provides consistent evidence for ITAs across its experiments and conditions in which it claimed ITAs. Moreover, these results are obtained under most favorable conditions for finding an ITA: Our reanalysis overestimates the indirect task sensitivities and therefore the evidence for an ITA due to our conservative choice of analysis strategies. Additionally, some of the reanalyzed studies apply problematic methodology that further biases the results towards finding an ITA even if there is none, see Discussion. This pattern of results casts serious doubts on the existence of ITAs in most, if not all, of the studies. Error bars represent 95%-confidence intervals.

give a direct task response. This may decrease the direct task sensitivity due to the additional difficulty, which can produce misleading ITA results. It is somewhat impressive that, even under these favorable circumstances, none of these reanalyzed studies provide consistent evidence for an ITA.

Nevertheless, our results do not necessarily rule out the possibility that ITAs exist in some cases. But the existence of an ITA may depend on the particular task and stimuli used. It might not be as ubiquitous as previously thought. Albeit the long standing request to use the same metric for both tasks (Reingold & Merikle, 1988) has often been ignored, there are some studies that provide evidence for an ITA using the appropriate analysis. For example, the setting of Schmidt (2002)—color stimuli served as primes and targets—found a distinct ITA result. Another example is the study by Kunst-Wilson and Zajonc (1980) using geometric shapes (but see also de Zilva et al., 2013; Seamon et al., 1983).

Therefore, we do not claim that there are no instances in which an ITA exists. Such a claim would be far beyond the scope of a single scientific study. But we do claim that one of the most prevalent methods in the wide research area of unconscious priming is fundamentally flawed. This flaw affects and potentially invalidates interpretations of many studies. As a consequence, the field has to reassess the situation of ITAs by applying the appropriate analysis to substantiate or refute previously made claims.

In deriving our appropriate methods, we have chosen strategies that favored finding an ITA. That is, we have followed the benefit-of-the-doubt approach to increase the chances of confirming an ITA. From such an approach, one would have expected clear evidence for an ITA in each of the reanalyzed studies. But because we nevertheless did not find consistent evidence for ITAs, having followed the benefit-of-the-doubt approach makes our arguments even stronger.

However, in future research, we hope that the benefit-of-the-doubt approach will no longer be necessary because it has a drawback: It would be inappropriate to simply revert the reasoning and use our liberal method to establish evidence *for* an ITA. To provide convincing evidence for an ITA, we would need a more balanced approach, one that might have not convinced researchers in the current situation (because they might have rejected it for being too conservative in terms of finding an ITA). For example, we used a clearly fail-safe estimate for q^2 in our reanalysis, that was chosen to be *larger* than all reported values on which this estimate is based. A more balanced approach would use a smaller estimate, which would reduce the chances to find an ITA, see our additional reanalyses in Supplement D for a figure like Figure 5 but with a more balanced estimate of q^2. Of course, trial-by-trial data should be used whenever possible.

To summarize, what we suggest is a research program: Given the tremendous interest in unconscious priming and the far-reaching inferences based on studies using the standard reasoning, researchers should reinvestigate the most relevant cases of claimed ITAs and clarify to which degree the claims in those studies are truly warranted. In those cases where an ITA is properly established, researchers can then start to draw further reaching conclusions about conscious versus unconscious processing (Eriksen, 1960; Erdelyi, 1986; Holender, 1986; Reingold & Merikle, 1988; Schmidt & Vorberg, 2006). An ITA is only a prerequisite but not a sufficient condition for the inferences that are typically drawn about unconscious processing.

In short, the literature needs a serious and concerted reassessment that would go well beyond the scope of a single study and will also require—in critical cases—the collection of new data. In many cases where superior unconscious processing already seemed an established fact (e.g., Hassin, 2013), we expect that this view needs to be revised. In other cases, researchers might still be able to establish such a relationship—which will then be even more interesting and foster the theoretical understanding of when exactly conscious processing is vital for a cognitive function and when it is not.

Context

Unconscious processing has been investigated for a long time. A common notion is that more information is processed unconsciously than consciously accessible. A main body of evidence for this comes from unconscious priming, where a standard reasoning is used to provide evidence for unconscious processing that exceeds consciously reportable processing. We show that the standard reasoning is flawed for statistical reasons. We introduce an analysis that is more appropriate. Using this analysis, we find that interpretations about unconscious processing break down. That is, even though the standard reasoning produced the notion that participants process more information unconsciously than they can consciously report, we show that there is inconsistent evidence for this in many studies. This lack of supposed evidence for superior unconscious processing has far-reaching consequences: It questions the idea that conscious processing is not necessary for many cognitive processes. We call for a community effort to apply the appropriate analysis and differentiate between situations in which processing can occur without being consciously reportable, that is, unconsciously.

References

Barnett, A. G., van der Pols, J. C., & Dobson, A. J. (2005). Regression to the mean: What it is and how to deal with it. *International Journal of Epidemiology*, *34*(1), 215–220. https://doi.org/10.1093/ije/dyh299

Boring, E. G. (1919). Mathematical vs. scientific significance. *Psychological Bulletin*, *16*(10), 335–338. https://doi.org/10.1037/h0074554

Buchner, A., & Wippich, W. (2000). On the reliability of implicit and explicit memory measures. *Cognitive Psychology*, *40*(3), 227–259. https://doi.org/10.1006/cogp.1999.0731

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, *7*(3), 249–253. https://doi.org/10.1177/014662168300700301

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7–29. https://doi.org/10.1177/0956797613504966

Custers, R., & Aarts, H. (2010). The unconscious will: How the pursuit of goals operates outside of conscious awareness. *Science*, *329*(5987), 47–50. https://doi.org/10.1126/science.1188595

Damian, M. F. (2001). Congruity effects evoked by subliminally presented primes: Automaticity rather than semantic processing. *Journal of Experimental Psychology: Human Perception and Performance*, *27*(1), 154–165. https://doi.org/10.1037//0096-1523.27.1.154

Dehaene, S., Lau, H., & Kouider, S. (2017). What is consciousness, and could machines have it? *Science*, *358*(6362), 486–492. https://doi.org/10.1126/science.aan8871

Dehaene, S., Naccache, L., Cohen, L., Bihan, D. L., Mangin, J. F., Poline, J. B., & Riviere, D. (2001). Cerebral mechanisms of word masking and unconscious repetition priming. *Nature Neuroscience*, *4*(7), 752–758. https://doi.org/10.1038/89551

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

Dell'Acqua, R., & Grainger, J. (1999). Unconscious semantic priming from pictures. *Cognition*, *73*(1), B1–B15. https://doi.org/10.1016/S0010-0277(99)00049-9

de Zilva, D., Vu, L., Newell, B. R., & Pearson, J. (2013). Exposure is not enough: Suppressing stimuli from awareness can abolish the mere exposure effect. *PloS ONE*, *8*(10), e77726. https://doi.org/10.1371/journal.pone.0077726

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*(3), 274–290. https://doi.org/10.1177/1745691611406920

Dijksterhuis, A., Bos, M. W., Nordgren, L. F., & Van Baaren, R. B. (2006). On making the right choice: The deliberation-without-attention effect. *Science*, *311*(5763), 1005–1007.

Draine, S. C., & Greenwald, A. G. (1998). Replicable unconscious semantic priming. *Journal of Experimental Psychology: General*, *127*, 286–303.

Dulaney, D. E., & Eriksen, C. W. (1959). Accuracy of brightness discrimination as measured by concurrent verbal responses and GSRs. *Journal of Abnormal and Social Psychology*, *59*(3), 418–423. https://doi.org/10.1037/h0040134

Erdelyi, M. H. (1986). Experimental indeterminacies in the dissociation paradigm of subliminal perception. *Behavioral and Brain Sciences*, *9*(1), 30–31. https://doi.org/10.1017/S0140525X00021348

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

Finkbeiner, M. (2011). Subliminal priming with nearly perfect performance in the prime-classification task. *Attention, Perception, & Psychophysics*, *73*(4), 1255–1265. https://doi.org/10.3758/s13414-011-0088-8

Finkbeiner, M., & Palermo, R. (2009). The role of spatial attention in nonconscious processing: A comparison of face and nonface stimuli. *Psychological Science*, *20*(1), 42–51. https://doi.org/10.1111/j.1467-9280.2008.02256.x

Franz, V. H., & Gegenfurtner, K. R. (2008). Grasping visual illusions: Consistent *Cognitive Neuropsychology*, *25*(7–8), 920–950.

Franz, V. H., & von Luxburg, U. (2015). No evidence for unconscious lie detection: A significant difference does not imply accurate classification. *Psychological Science*, *26*(10), 1646–1648. https://doi.org/10.1177/0956797615597333

Green, D. M., & Swets, J. A. (1988). *Signal detection theory and psychophysics*. Peninsula.

Greenwald, A. G., Klinger, M. R., & Schuh, E. S. (1995). Activation by marginally perceptible ('subliminal') stimuli: Dissociation of unconscious from conscious cognition. *Journal of Experimental Psychology: General*, *124*, 22–42.

Hannula, D. E., Simons, D. J., & Cohen, N. J. (2005). Imaging implicit perception: promise and pitfalls. *Nature Reviews Neuroscience*, *6*(3), 247–255. https://doi.org/10.1038/nrn1630

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

Holender, D. (1986). Semantic activation without conscious identification in dichotic–listening, parafoveal vision, and visual masking—a survey and appraisal. *Behavioral and Brain Sciences*, *9*(1), 1–23. https://doi.org/10.1017/S0140525X00021269

Janssen, P. A., Thiessen, P., Klein, M. C., Whitfield, M. F., MacNab, Y. C., & Cullis-Kuhl, S. C. (2007). Standards for the measurement of birth weight, length and head circumference at term in neonates of European, Chinese and South Asian ancestry. *Open Medicine*, *1*(2), E74–E88.

Kiefer, M. (2002). The N400 is modulated by unconsciously perceived masked words: Further evidence for an automatic spreading activation account of N400 priming effects. *Cognitive Brain Research*, *13*(1), 27–39. https://doi.org/10.1016/s0926-6410(01)00085-4

Klotz, W., & Neumann, O. (1999). Motor activation without conscious discrimination in metacontrast masking. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(4), 976–992.

Kouider, S., & Dehaene, S. (2007). Levels of processing during non–conscious perception: a critical review of visual masking. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 857–875. https://doi.org/10.1098/rstb.2007.2093

Kouider, S., & Dehaene, S. (2009). Subliminal number priming within and across the visual and auditory modalities. *Experimental Psychology*, *56*(6), 418–433. https://doi.org/10.1027/1618-3169.56.6.418

Kunde, W., Kiesel, A., & Hoffmann, J. (2003). Conscious control over the content of unconscious cognition. *Cognition*, *88*(2), 223–242. https://doi.org/10.1016/S0010-0277(03)00023-4

Kunst-Wilson, W. R., & Zajonc, R. B. (1980). Affective discrimination of stimuli that cannot be recognized. *Science*, *207*(4430), 557–558. https://doi.org/10.1126/science.7352271

Mattler, U. (2003). Priming of mental operations by masked stimuli. *Perception & Psychophysics*, *65*(2), 167–187. https://doi.org/10.3758/bf03194793

Mattler, U. (2005). Inhibition and decay of motor and nonmotor priming. *Perception & Psychophysics*, *67*(2), 285–300. https://doi.org/10.3758/bf03206492

Merikle, P. M. (1992). Perception without awareness: Critical issues. *American Psychologist*, *47*(6), 792–795. https://doi.org/10.1037/0003-066X.47.6.792

Meyen, S., Zerweck, I. A., Amado, C., von Luxburg, U., & Franz, V. H. (2021). Advancing research on unconscious priming: When can scientists claim an indirect task advantage? osf.io/kp59h

Michel, M., Beck, D., Block, N., Blumenfeld, H., Brown, R., Carmel, D., Carrasco, M., Chirimuuta, M., Chun, M., Cleeremans, A., Dehaene, S., Fleming, S. M., Frith, C., Haggard, P., He, B. J., Heyes, C., Goodale, M. A., Irvine, L., Kawato, M., . . . Yoshida, M. (2019). Opportunities and challenges for a maturing science of consciousness. *Nature Human Behaviour*, *3*(2), 104–107. https://doi.org/10.1038/s41562-019-0531-8

Morris, J. S., Ohman, A., & Dolan, R. J. (1998). Conscious and unconscious emotional learning in the human amygdala. *Nature*, *393*(6684), 467–470. https://doi.org/10.1038/30976

Morris, J. S., Öhman, A., & Dolan, R. J. (1999). A subcortical pathway to the right amygdala mediating "unseen" fear. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(4), 1680–1685.

Morrison, D., & Henkel, R. (1970). *The significance test controversy*. Aldine Transaction.

Naccache, L., Blandin, E., & Dehaene, S. (2002). Unconscious masked priming depends on temporal attention. *Psychological Science*, *13*(5), 416–424. https://doi.org/10.1111/1467-9280.00474

Naccache, L., & Dehaene, S. (2001). Unconscious semantic priming extends to novel unseen stimuli. *Cognition*, *80*(3), 215–229. https://doi.org/10.1016/S0010-0277(00)00139-6

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E. J. (2011). Erroneous analyses of interactions in neuroscience: A problem of significance. *Nature Neuroscience*, *14*(9), 1105–1107. https://doi.org/10.1038/nn.2886

Palfi, B., & Dienes, Z. (2020). Why Bayesian evidence for H1 in one condition and Bayesian evidence for H0 in another condition does not mean good-enough Bayesian evidence for a difference between the conditions. *Advances in Methods and Practices in Psychological Science*, *3*(3), 300–308. https://doi.org/10.1177/2515245920913019

Peremen, Z., & Lamy, D. (2014). Do conscious perception and unconscious processing rely on independent mechanisms? A meta-contrast study. *Consciousness and Cognition*, *24*, 22–32. https://doi.org/10.1016/j.concog.2013.12.006

Pessiglione, M., Schmidt, L., Draganski, B., Kalisch, R., Lau, H., Dolan, R. J., & Frith, C. D. (2007). How the brain translates money into force: A neuroimaging study of subliminal motivation. *Science*, *316*(5826), 904–906. https://doi.org/10.1126/science.1140459

Rausch, M., Hellmann, S., & Zehetleitner, M. (2018). Confidence in masked orientation judgments is informed by both evidence and visibility. *Attention, Perception, & Psychophysics*, *80*(1), 134–154. https://doi.org/10.3758/s13414-017-1431-5

Reingold, E. M., & Merikle, P. M. (1988). Using direct and indirect measures to study perception without awareness. *Perception & Psychophysics*, *44*(6), 563–575. https://doi.org/10.3758/BF03207490

Reingold, E. M., & Merikle, P. M. (1990). On the inter-relatedness of theory and measurement in the study of unconscious processes. *Mind & Language*, *5*(1), 9–28.

Schmidt, T. (2002). The fingers in flight: Real–time control by visually masked color stimuli. *Psychological Science*, *13*(2), 112–118. https://doi.org/10.1111/1467-9280.00421

Schmidt, T. (2015). Invisible stimuli, implicit thresholds: Why invisibility judgments 903 cannot be interpreted in isolation. *Advances in Cognitive Psychology*, *11*(2), 31–41.

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

Seamon, J. G., Brody, N., & Kauff, D. M. (1983). Affective discrimination of stimuli that are not recognized: Effects of shadowing, masking, and cerebral laterality *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*(3), 544–555. https://doi.org/10.1037/0278-7393.9.3.544

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

Simons, D. J., Hannula, D. E., Warren, D. E., & Day, S. W. (2007). Behavioral, neuroimaging, and neuropsychological approaches to implicit perception. In P. D. Zelazo, M. Moscovitch, & E. Thompson (Eds.), *The Cambridge handbook of consciousness* (pp. 207–250). Cambridge University Press. https://doi.org/10.1017/CBO9780511816789.010

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

Sumner, P. (2008). Mask-induced priming and the negative compatibility effect. *Experimental Psychology*, *55*(2), 133–141. https://doi.org/10.1027/1618-3169.55.2.133

ten Brinke, L., Stimson, D., & Carney, D. R. (2014). Some evidence for unconscious lie detection. *Psychological Science*, *25*(5), 1098–1105. https://doi.org/10.1177/0956797614524421

ten Brinke, L., Vohs, K. D., & Carney, D. R. (2016). Can ordinary people detect deception after all? *Trends in Cognitive Sciences*, *20*(8), 579–588. https://doi.org/10.1016/j.tics.2016.05.012

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

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

Van den Bussche, E., Van den Noortgate, W., & Reynvoet, B. (2009). Mechanisms of masked priming: A meta–analysis. *Psychological Bulletin*, *135*(3), 452–477. https://doi.org/10.1037/a0015329

van Gaal, S., Ridderinkhof, K. R., Scholte, H. S., & Lamme, V. A. F. (2010). Unconscious activation of the prefrontal no-go network. *Journal of Neuroscience*, *30*(11), 4143–4150. https://doi.org/10.1523/JNEUROSCI.2992-09.2010

Wang, Y., Wang, Y., Liu, P., Di, M., Gong, Y., & Tan, M. (2017). The role of representation strength of the prime in subliminal visuomotor priming. *Experimental Psychology*, *64*(6), 422–431. https://doi.org/10.1027/1618-3169/a000381

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, *61*(7), 726–728. https://doi.org/10.1037/0003-066X.61.7.726

Wójcik, M. J., Nowicka, M. M., Bola, M., & Nowicka, A. (2019). Unconscious detection of ones own image. *Psychological Science*, *30*(4), 471–480. https://doi.org/10.1177/0956797618822971

Wolfe, J. M. (1999). Inattentional amnesia. In V. Coltheart (Ed.), *Cognitive psychology. Fleeting memories: Cognition of brief visual stimuli* (pp. 71–94). The MIT Press.

Zehetleitner, M., & Rausch, M. (2013). Being confident without seeing: what subjective measures of visual consciousness are about. *Attention, Perception, & Psychophysics*, *75*(7), 1406–1426. https://doi.org/10.3758/s13414-013-0505-2

Zerweck, I. A., Kao, C.-S., Meyen, S., Amado, C., von Eltz, M., Klimm, M., & Franz, V. H. (2021). No evidence for superior unconscious processing of digits. *Attention, Perception & Psychophysics*. Manuscript submitted for publication.

(*Appendix follows*)

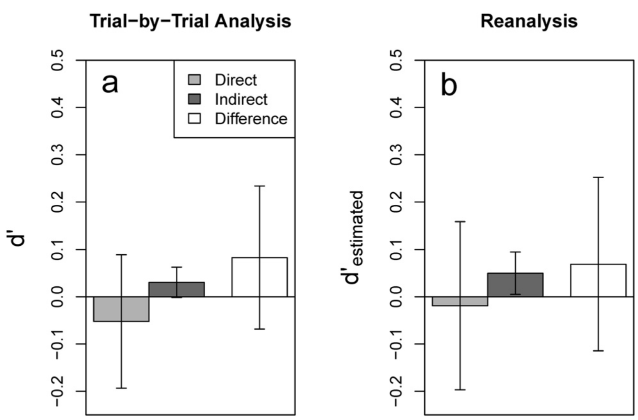
Appendix

Validation of Reanalysis Method

We demonstrate in two studies that the appropriate analysis based on the full trial-by-trial data is well approximated by our reanalysis based only on the typically reported statistics (e.g., a *t* value for the congruency effect in the indirect task). We compare the appropriate analysis using the full, trial-by-trial data on one hand and our reanalysis method based on only the reported summary statistics on the other hand. We apply both approaches to the original data from ten Brinke et al. (2014, see Figure A1) and to our replication of Dehaene et al. (1998, see Figure A2). Results from the two analyses are very comparable confirming the validity of our reanalysis.

Figure A1

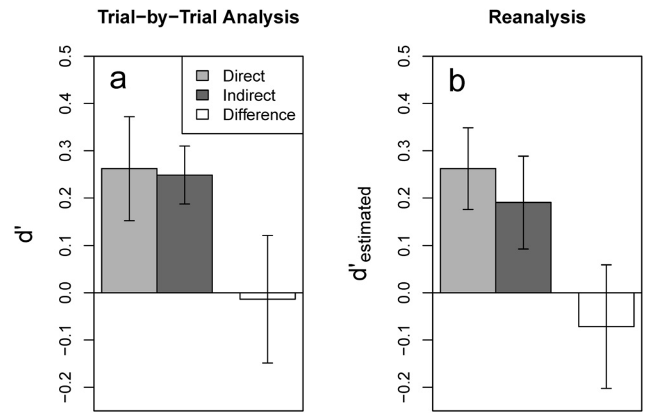
Appropriate Analysis Applied to ten Brinke et al. (2014) Using the Full, Trial-by-Trial Data in (a) and Using Our Reanalysis Method in (b)



Note. Our reanalysis using only the typically reported statistics produced approximately the same results as the trial-by-trial analysis. In both cases, there is no evidence for an indirect task advantage. Error bars indicate 95% confidence intervals.

Figure A2

Appropriate Analysis Applied to Our Replication of Dehaene et al. (1998) Using the Full, Trial-by-Trial Data in (a) and Using Our Reanalysis Method in (b)



Note. Our reanalysis using only the typically reported statistics produced approximately the same results as the trial-by-trial analysis. In both cases, there is no evidence for an indirect task advantage. Note that the indirect task sensitivity in our reanalysis is smaller than in the trial-by-trial analysis. This is not a contradiction to our claim that in expectation the indirect task sensitivity is overestimated by our reanalysis. Estimates of individual studies can vary as indicated by the error bars indicating 95% confidence intervals.

Received December 13, 2019
Revision received January 14, 2021
Accepted February 5, 2021 ■

Supplemental Materials
Advancing Research on Unconscious Priming: When can Scientists Claim an Indirect Task Advantage
by S. Meyen, I. A. Zerweck, C. Amado, U. v. Luxburg, & V. H. Franz

A Overview

B Validation of Reanalysis Method via Simulations

We conducted multiple simulations to validate that our reanalysis method appropriately controls for statistical errors (type I and type II). Each simulation was repeated $10,000$ times. In each run, we generated a trial-by-trial data set with a direct and an indirect task according to the standard repeated measures model outlined in Appendix D. We simulated N participants with sensitivities, $d'_{\text{true},i}$, independently and randomly drawn from normal distributions with expected value d'_{true} and variance q^2 (see Appendix E for why q^2 is the variance of individual true sensitivities). Note that we sampled $d'_{\text{true},i}$ for each participant independently in the direct and indirect task to avoid making additional assumptions on their correlation between tasks. Applying Signal Detection Theory, each of these individual sensitivities implies two normal distributions separated by $d'_{\text{true},i}$ standard deviations. From these normal distributions, we sampled a total of K trials for each participant, $K/2$ in each condition. We did this twice, once for each task. In the direct task, we compared each response to the true median: If the response lied on the same side as the normal distribution it was sampled from, the simulated binary decision by the participant in this trial was correct, otherwise it was wrong. In the indirect task, we simply treated the drawn responses as the indirect measures (e.g., RTs). We then applied the traditional analysis used in the standard reasoning and the appropriate analyses, first based on the full, trial-by-trial analysis and second our reanalysis based on typically reported summary statistics. We obtained similar results with log-normal distributions and only report normal distribution results for brevity.

In each simulation, we varied N, K, d'_{true} and q^2. If not declared otherwise, the same q^2 was used for data simulation and reanalysis. Only in simulations 5 and 6, we varied the true q^2 with which the data was simulated and used a different q^2 for our reanalysis in order to see how getting this parameter wrong would affect our results.

Simulations 1-3 demonstrate that the standard reasoning applied to the traditional analysis miserably fails when applied to the study of Dehaene et al. (1998). Simulation 4 shows that our replication has sufficient statistical power to find an ITA if it was there. Simulations 5 and 6 show how our reanalysis would be affected, if the true q^2 was different than what we assumed. We then summarize additional 108 simulations showing that our estimators, even though they use simplifying approximations, are approximately unbiased.

Simulations

Simulation 1: Controlling type I errors. We used the same number of participants in the direct ($N = 7$) vs. indirect ($N = 12$) task as well as the same number of trials per condition (direct $K = 112$ vs. indirect $K = 512$) as the original study of Dehaene et al. (1998). Assuming no ITA, we set sensitivities in both tasks to be equal (direct $d'_{\text{true}} = 0.25$ vs. indirect $d'_{\text{true}} = 0.25$). We assumed $q^2 = 0.0225$ for this simulation.

Even though the same sensitivity underlies both tasks, the direct task fails to reach significance half of the time (51.2%) while the indirect task is almost always significant (99.5%). This is not surprising and shows how seriously underpowered the direct task was due to fewer samples, N and K. When applying the standard reasoning, a scientist would erroneously conclude an ITA from a non-significant direct task result and a significant indirect task effect in 48.6% of the experiments. In other words: The widely used standard reasoning would infer an ITA half of the time even though there is no ITA present!

Since there is no ITA present, our reanalysis should find an ITA only as often as prespecified by the significance level $\alpha = 5\%$. Indeed, we find a difference between the two tasks only in 4.7% of the runs. This demonstrates that our reanalysis approach controls appropriately for type I errors.

Simulation 2: Controlling for type II errors with an underpowered direct task. We use the same settings as in Simulation 1 except that we now assume there exists an ITA (direct $d'_{\text{true}} = 0$ vs. indirect $d'_{\text{true}} = 0.25$). Since there is an ITA present, a high statistical power is desired to detect it and avoid type II errors. Typically, a power above $1 - \beta = 80\%$ is desired. However, our reanalysis found the ITA in only 46.2% of the runs. Using the full trial-by-trial data to test for a difference (instead of only using the reported t value from the indirect task) also produced a test power of only 45.9%. There is simply not enough data in the direct task to provide sufficient evidence for an ITA. The problem with lacking statistical power is not located in our reanalysis because the analysis based on the trial-by-trial data also has a low statistical power. Instead, the problem is the low sample size in the direct task.

Simulation 3: Controlling for type II errors with sufficient samples in the direct task. We repeated Simulation 2 but increased the number of participants and trials in the direct task to match the ones of the indirect task ($N = 12$ and $K = 512$). This is most sensible when testing for a difference because a balanced design maximizes statistical power. Here, our reanalysis method detects the ITA in 78.3% of the runs, which is close to the desired 80%. Using the full trial-by-trial data provides a power of 84.2%. This demonstrates that our reanalysis method provides sufficient power given sufficient samples.

Simulation 4: Statistical power in our replication. We repeated Simulation 3 but used the same number of participants and samples as in our replication study, $N = 24$ and $K = 256$ in both tasks. There, we have the same amount of observations as Dehaene et al. (1998) (double the participants, half the trials). Here, our reanalysis detects the ITA in 96.5% of the runs. The analysis using trial-by-trial data instead of only a t value achieves 97.0%. The increase in statistical power compared to Simulation 3 comes from sampling more participants which is more efficient than sampling more trials given a fixed total number of observations (Rouder & Haaf, 2018).

Simulation 5: Overestimating parameter q^2. We repeated Simulation 3, the balanced Dehaene et al. (1998) setting with an ITA, but generated the data with $q^2 = 0.01$. We still use $q^2 = 0.0225$ for the reanalysis, thus, we overestimate the true q^2. Our reanalysis now successfully detects the ITA in 99.6% of the runs and so does the appropriate analysis with 99.2%. We detect more ITAs here than in Simulation 3 because we make our reanalysis more liberal by choosing a larger q^2.

Simulation 6: Underestimating parameter q^2. Repeating Simulation 5, we now simulated the data with $q^2 = 0.09$ and kept the parameter of our reanalysis at $q^2 = 0.0225$, that is, we now underestimate the true q^2. Individual sensitivities vary a lot now. Even though the mean true direct task sensitivity is $d'_{\text{true}} = 0$ (50%-correct), due to a large standard deviation of $q = 0.3$, 95% of participants' true sensitivities range between -0.6 (38%-correct) and 0.6 (62%-correct). The assumption $q = 0.3$ poses a problem from a theoretical perspective because some participants can discriminate the masked stimuli relatively well (above 60%-correct). In this case, our reanalysis is more conservative and detects an ITA in only 62.2% of the runs. However, the analysis based on the trial-by-trial data also only achieves a power of 69.2% due to the large variability: Even in this case, our reanalysis would not be too conservative.

Additional Simulations. We conducted additional simulations, one for each combination of the following parameters: $N \in \{5, 10, 20\}$, $K \in \{100, 200, 400\}$, $d'_{\text{true}} = \{0, 0.1, 0.2, 0.5\}$, and $q^2 \in \{0.01, 0.0225, 0.09\}$. In all these simulations, the average, absolute deviation between true and estimated sensitivities was small, $|d'_{\text{true}} - d'_{\text{estimated}}| \leq 0.01$. A deviation of 0.01 in terms of sensitivity translates into a deviation as small as 0.2%-correct, which can be considered negligible in this setting—and deviations in simulations with $N \geq 10$ are substantially smaller.

We computed the standard deviation of $d'_{\text{estimated}}$ (denoted by $SD[d'_{\text{estimated}}]$) across the 10,000 simulations of each parameter combination. We compared this with the estimated standard error, SE. For this purpose, we squared SE of each run, averaged the values and took the square root of the average, which is the standard procedure to average standard errors. For the direct task, the difference between actual variability and our estimates was again $|SD[d'_{\text{estimated}}] - SE| \leq 0.01$. For the indirect task, the same was true when $N \geq 10$. However, for very small sample sizes ($N = 5$) our reanalysis deviated to some degree but the absolute difference between actual standard deviation and our estimates still was $|SD[d'_{\text{estimated}}] - SE| \leq 0.05$. Since all reanalyzed studies use sample sizes of $N \geq 10$ in the indirect task, our reanalysis produced approximately unbiased estimates. Overall, our reanalysis approximates the appropriate analysis sufficiently well in the context we applied it to.

C Optimality of Median Classifier

In the appropriate analysis to infer an ITA, one needs to transform continuous measurements of the indirect task (e.g., RTs) into a binary classification response. In this step it is important to use the best possible classifier, in order to achieve the highest d' or %-correct values and thereby increase the chance to establish an ITA. Depending on the type of measurement that is taken in the indirect task (e.g., RT, brain activity, grip force, etc.), this best classifier can have different forms. In many cases, the median classifier is a suitable choice. For example with RT data (as in our replication based on Dehaene et al., 1998), the classifier computes for each participant the median RT across all trials and classifies a trial as congruent if the RT is faster than the median and as incongruent if the RT is slower. Below, we prove that the median classifier is optimal in this setting. The proof requires two assumptions:

(1) The indirect measure follows a normal or log-normal distribution with an additive shift between congruent and incongruent conditions. In our case, this assumption is justified because it is well known that RT distributions are well approximated by log-normal distributions (Ulrich & Miller, 1993; Palmer, Horowitz, Torralba, & Wolfe, 2011).

(2) An equal number of observations need to be drawn in both conditions, which is satisfied by the typical experimental design.

Note, that Franz and von Luxburg (2015) also applied nonparametric machine learning classifiers with similar results.

General form of the optimal classifier. Consider a classification task where the input is a real-valued number x (e.g., a reaction time, RT), and the classifier is supposed to predict one of two labels y (e.g., 'congruent' or 'incongruent'; for simplicity we use labels 1 and 2 in the following). Following the standard setup in statistical decision theory (Bishop, 2006, section 1.5) we assume that the input data X and the output data Y are drawn according to some fixed (but unknown) probability distribution P. This distribution can be described uniquely by the class-conditional distributions $P(X \mid Y = 1)$ and $P(X \mid Y = 2)$ and the class priors $\pi_1 = P(Y = 1)$ and $\pi_2 = P(Y = 2)$. A classifier is a function $f : \boldsymbol{R} \rightarrow \{1, 2\}$ that assigns a label y to each input x. The classifier that has the smallest probability of error is called the Bayes classifier. In case the classes have equal weight, that is $\pi_1 = \pi_2$, the Bayes classifier has a particularly simple form: It classifies an input point x by the class that has the higher class-conditional density at this point. Formally, this classifier is given by

$$f_{opt}(x) := \begin{cases} 1 & \text{if } P(X = x \,|\, Y = 1) > P(X = x \,|\, Y = 2) \\ 2 & \text{otherwise.} \end{cases} \tag{1}$$

Optimal classifier for normal and log-normal distributions. We now consider the special case where the class-conditionals follow a particular distribution. Let us start with the normally distributed case. We assume that both class-conditionals are normal distributions with means μ_1, μ_2 and equal variance σ^2, and we denote their corresponding probability density functions (PDFs) by $\varphi_{\mu_1,\sigma}$ and $\varphi_{\mu_2,\sigma}$. Under the additional assumption that both classes have equal weights $\pi_1 = \pi_2 = 0.5$, the cumulative distribution function (CDF) of the input (marginal distribution of X) is given as

$$\Omega(x) := 0.5 \cdot \Big(\Phi(\frac{x - \mu_1}{\sigma}) + \Phi(\frac{x - \mu_2}{\sigma})\Big), \tag{2}$$

where Φ denotes the CDF of the standard normal distribution. For $t \in \boldsymbol{R}$, we introduce the step function classifier with threshold t by

$$f_t(x) := \begin{cases} 1 & \text{if } x \leq t \\ 2 & \text{otherwise.} \end{cases} \tag{3}$$

In the special case where the threshold t coincides with the median of the marginal distribution of X, we call the resulting step function classifier the *median classifier.*

Proposition (Median classifier is optimal for normal model) *If the input distribution is given by Equation* (2)*, then the optimal classifier f_{opt} coincides with the median classifier.*

Proof. Because both classes have the same weight of 0.5, the Bayes classifier is given by f_{opt} as in Equation (1). For any choice of μ_1, μ_2 and σ, the class-conditional PDFs $\varphi_{\mu_1,\sigma}$ and $\varphi_{\mu_2,\sigma}$ intersect exactly once, namely at $t^* = (\mu_1 + \mu_2)/2$. By definition of f_{opt}, the optimal classifier f_{opt} is then the step function classifier with threshold t^*. We now compute the value of the CDF at t^*:

$$\begin{aligned} \Omega(t^*) &= 0.5 \cdot \Big(\Phi(\frac{t^* - \mu_1}{\sigma}) + \Phi(\frac{t^* - \mu_2}{\sigma})\Big) \\ &= 0.5 \cdot \Big(\Phi(\frac{\mu_2 - \mu_1}{2\sigma}) + \Phi(\frac{\mu_1 - \mu_2}{2\sigma})\Big) \\ &= 0.5 \cdot \Big(\Phi(\frac{\mu_2 - \mu_1}{2}) + (1 - \Phi(\frac{\mu_2 - \mu_1}{2})\Big) \\ &= 0.5. \end{aligned}$$

Here, the second last equality comes from the fact that the normal distribution is symmetric about 0. This calculation shows that the optimal threshold t^* indeed coincides with the median of the input distribution, which is what we wanted to prove. □

It is easy to see that this proof can be generalized to more general types of symmetric probability distributions. It is, however, even possible to prove an analogous statement for log-normal distributions, which are not symmetric themselves. We introduce the notation $\lambda_{\mu,\sigma}$ for the PDF of a log-normal distribution, and $\Lambda_{\mu,\sigma}$ for the corresponding CDF. These functions are defined as

$$\lambda_{\mu,\sigma}(x) := \frac{1}{x\sigma\sqrt{2\pi}} \exp\Big(-\frac{(\log x - \mu)^2}{2\sigma^2}\Big) \qquad \text{and} \qquad \Lambda_{\mu,\sigma}(x) := \Phi\Big(\frac{\log x - \mu}{\sigma}\Big).$$

Consider the case where the class-conditional distributions are log-normal distributions with same scale parameter σ but different location parameters μ_1 and μ_2, and assume that both classes have the same weights $\pi_1 = \pi_2 = 0.5$. Then the PDF and

CDF of the input distribution (marginal distribution of X) are given as

$$\begin{aligned} g(x) &= 0.5 \cdot \; (\lambda_{\mu_1,\sigma}(x) + \lambda_{\mu_2,\sigma}(x) \;) \\ G(x) &= 0.5 \cdot \; (\Lambda_{\mu_1,\sigma}(x) + \Lambda_{\mu_2,\sigma}(x) \;). \end{aligned} \tag{4}$$

Proposition (Median classifier is optimal for log-normal model) *If the input distribution is given by Equation* (4)*, then the optimal classifier* f_{opt} *coincides with the median classifier.*

Proof. The proof is analogous to the previous one. For any choice of μ_1, μ_2 and σ, the densities $\lambda_{\mu_1,\sigma}$ and $\lambda_{\mu_2,\sigma}$ intersect exactly once. To see this, we solve the equation $\lambda_{\mu_1,\sigma}(t^*) = \lambda_{\mu_2,\sigma}(t^*)$, which leads to the unique solution $t^* = \exp((\mu_1 + \mu_2)/2)$. The input CDF at this value can be computed as

$$\begin{aligned} G(t^*) &= 0.5\Big(\Lambda_{\mu_1,\sigma}(t^*) + \Lambda_{\mu_2,\sigma}(t^*)\Big) \\ &= 0.5\Big(\Phi(\frac{\mu_2 - \mu_1}{2\sigma}) + \Phi(\frac{\mu_1 - \mu_2}{2\sigma})\Big) \\ &= 0.5. \end{aligned}$$

The last step follows as above by the symmetry of the normal cdf. □

D Estimating Sensitivities From Typically Reported Results

We use typically reported results from studies on unconscious priming to estimate the direct and indirect task sensitivities, $d'_{\text{estimated,direct}}$ and $d'_{\text{estimated,indirect}}$. First, we recapitulate the basic model assumptions of a standard repeated measures ANOVA and introduce the notation. We then derive estimators for the sensitivity and standard error in both tasks using only the typically reported results. Finally, we compute the difference between direct vs. indirect task sensitivities and construct a confidence interval around that difference in order to test for an ITA.

Model assumptions

Our reanalysis of both tasks is based on the standard model of repeated measures ANOVA and paired t test (Winer, Brown, & Michels, 1991; Maxwell & Delaney, 2000; Rouder & Haaf, 2018) as employed in all reanalyzed studies. In this model N participants perform M trials in each condition. In the specific setting we consider, there are only 2 conditions. In the direct task, this corresponds to trials where the masked stimulus is from either of two categories, A vs. B. In the indirect task, the two conditions are typically congruent (A-A, B-B) vs. incongruent (A-B, B-A). In each trial of a given task, Y_{ijk} denotes the response from participant i $(1, ..., N)$ in condition j (1 or 2) in trial k $(1, ..., M)$, where we assume a balanced design such that the total number of trials K is split evenly into the two conditions for $M = K/2$ trials per condition.

In the indirect task, responses Y_{ijk} are the indirect measures (e.g., RTs). In the direct task, it is plausible to assume that responses Y_{ijk} represent participants' internal evidence about the masked stimuli (some neural activity indicating whether the participant saw a masked stimulus from category A or from B). Based on this noisy internal evidence, participants make an internal classification and guess in each trial to which category the stimulus belonged.

The standard model decomposes participants' responses Y_{ijk} into five components:

$$Y_{ijk} = \mu + p_i + c_j + (p \times c)_{ij} + \epsilon_{ijk}.$$

To facilitate understanding, we now describe the model for the example of congruency effects on RTs in the indirect task; but the same notation applies to other indirect measures and to the direct task as well. RTs have a grand mean μ. Some participants have faster RTs than others which is captured in participants' effects p_i. The congruency condition has an effect c_j on RTs. While c_1 is negative leading to faster RTs in congruent trials, c_2 is positive reflecting slower RTs in the incongruent conditions. Participants differ in the extent to which the congruency conditions affect them captured in $(p \times c)_{ij}$ so that some participants have a larger congruency effect than others. The variability in the individual effects is captured by this term's variance, $\text{Var}[(p \times c)_{ij}] = \sigma^2_{p\times c}$. Additionally, there is trial-by-trial noise ϵ_{ijk} from neuromuscular noise and measurement error leading to different responses in each trial. This trial-by-trial measurement error is assumed by the standard models to have a constant variance (homogeneity) across participants and conditions, $\text{Var}[\epsilon_{ijk}] = \sigma^2_\epsilon$. The congruency effect c_j is a fixed effect while participant and interaction effects (p_i and $(p \times c)_{ij}$) are random effects because they depend on the drawn sample of participants. Random effects and trial-by-trial noise are assumed to be normally distributed with an expected value of zero and their corresponding variance.

Raw effects and sensitivities. Each participant i has an individual expected congruency effect, Δ_i, which theoretically would be obtained by sampling infinitely many trials. The expected RT difference across participants is denoted by Δ.

$$\begin{aligned}\Delta_i &= (c_2 + (p \times c)_{i2}) - (c_1 + (p \times c)_{i1})\\ \Delta &= c_2 - c_1\end{aligned}$$

In a typical experiment, the individual congruency effects are estimated by the observed mean difference between conditions. For the i-th participant, this estimate is $\hat{\Delta}_i$ and averaged across participants this is $\hat{\Delta}$.

$$\begin{aligned}\hat{\Delta}_i &= \bar{Y}_{i2\cdot} - \bar{Y}_{i1\cdot} = \frac{1}{M}\sum_{k=1}^{M} Y_{i2k} - \frac{1}{M}\sum_{k=1}^{M} Y_{i1k}\\ \hat{\Delta} &= \frac{1}{N}\sum_{i=1}^{N}\hat{\Delta}_i\end{aligned}$$

A participant's true sensitivity $d'_{\text{true},i}$ is the normalized effect—normalized by the trial-by-trial error standard deviation σ_ϵ. This quantity indicates, similar to a signal to noise ratio, how well a participant's RTs are separable and therefore to which degree the masked stimuli were processed, cf. Figure 3a. The expectation across participants is the true sensitivity d'_{true} indicating how well the RTs of a prototypical participant are separated.

$$\begin{aligned}d'_{\text{true},i} &= \frac{\Delta_i}{\sigma_\epsilon}\\ d'_{\text{true}} &= \frac{\Delta}{\sigma_\epsilon}.\end{aligned}$$

In the direct task, d'_{true} is typically measured by the sensitivity index d' averaged across participants. Participants' individual d'_i are calculated from hit rate, HR (%-correct guesses for masked stimuli from category A), and false alarm rate, FA (%-incorrect guesses for masked stimuli from category B), where Φ^{-1} is the inverse cumulative density function of the normal distribution.

$$\begin{aligned}d'_i &= \Phi^{-1}(\text{HR}) - \Phi^{-1}(\text{FA})\\ d' &= \frac{1}{N}\sum_i d'_i.\end{aligned}$$

Note that the empirical literature often uses the notation d' without a clear distinction between estimated vs. true value and individual vs. average effects. Because we need to be more precise in our derivations: We denote the true value of an individual participant by $d'_{\text{true},i}$ and the sensitivity index, which is an estimate for the true value, by d'_i. We denote the true sensitivity across participants by d'_{true}. In the direct task, this is estimated by the average across d'_i values denoted by d'. We will also label this averaged estimate $d'_{\text{estimated,indirect}}$.

Two variance sources: true effect (between-) vs. trial-by-trial error (within-subject) variance. Participants differ in their true congruency effect. The variance of these true inter-individual differences can be derived from the model variances using the standard assumptions (1) $(p \times c)_{ij} \sim \mathcal{N}(0, \sigma^2_{p\times c})$, (2) $\text{Var}[c_1] = \text{Var}[c_2] = 0$, and (3) $(p \times c)_{i1} = -(p \times c)_{i2}$. We denote this true effect variance as σ^2_{effect}:

$$\begin{aligned}\sigma^2_{\text{effect}} &= \text{Var}[\Delta_i] = \text{Var}[[c_2 + (p \times c)_{i2}] - [c_1 + (p \times c)_{i1}]]\\ &= \text{Var}[(p \times c)_{i2} - (p \times c)_{i1}] = \text{Var}[2(p \times c)_{i2}]\\ &= 4\sigma^2_{p\times c}.\end{aligned}$$

The variance of the actually observed congruency effects is conceptually different from the variance of the true effects. We denote the variance of the observed congruency effects as $\sigma^2_{\hat{\Delta}_i}$. The observed congruency effects vary more because they are not only affected by true inter-individual difference but also by trial-by-trial measurement errors:

$$\begin{aligned}
\sigma^2_{\hat{\Delta}_i} &= \mathrm{Var}[\hat{\Delta}_i] \\
&= \mathrm{Var}\left[\bar{Y}_{i2\cdot} - \bar{Y}_{i1\cdot}\right] \\
&= \mathrm{Var}\left[\frac{1}{M}\left(\sum_{k=1}^{M}\mu + p_i + c_2 + (p\times c)_{i2} + \epsilon_{i2k}\right)\right. \\
&\qquad \left. - \frac{1}{M}\left(\sum_{k=1}^{M}\mu + p_i + c_1 + (p\times c)_{i1} + \epsilon_{i1k}\right)\right] \\
&= \mathrm{Var}\left[[c_2 + (p\times c)_{i2}] - [c_1 + (p\times c)_{i1}] + \frac{1}{M}\left(\sum_{k=1}^{M}\epsilon_{i2k}\right) - \frac{1}{M}\left(\sum_{k=1}^{M}\epsilon_{i1k}\right)\right] \\
&= \mathrm{Var}\left[\Delta_i + \frac{1}{M}\left(\sum_{k=1}^{M}\epsilon_{i2k}\right) - \frac{1}{M}\left(\sum_{k=1}^{M}\epsilon_{i1k}\right)\right] \\
&= \mathrm{Var}[\Delta_i] + \mathrm{Var}\left[\frac{1}{M}\sum_{k=1}^{M}\epsilon_{i2k}\right] + \mathrm{Var}\left[\frac{1}{M}\sum_{k=1}^{M}\epsilon_{i1k}\right] \\
&= \sigma^2_{\text{effect}} + \frac{2}{M}\sigma^2_{\epsilon} \\
&= \sigma^2_{\text{effect}} + \frac{4}{K}\sigma^2_{\epsilon}.
\end{aligned}$$

This has an implication for the variance of average congruency effects, $\hat{\Delta} = \frac{1}{N}\sum_i \hat{\Delta}_i$. These observed, average congruency effects vary due to two variance sources, the true inter-individual differences and trial-by-trial measurement error.

$$\hat{\Delta} \sim \mathcal{N}\left(\Delta, \frac{\sigma^2_{\text{effect}} + \frac{4}{K}\sigma^2_{\epsilon}}{N}\right).$$

We will later have to estimate σ^2_{ϵ} from a given $\sigma^2_{\hat{\Delta}_i}$. To achieve this, we must disentangle σ^2_{effect} and σ^2_{ϵ}. We do so by defining the ratio q^2 between these two sources of variance:

$$q^2 = \frac{\sigma^2_{\text{effect}}}{\sigma^2_{\epsilon}}.$$

This parameter tells us how much of the observed variability comes from true differences vs. noise. If $q^2 = 0$ then all participants would have the same true congruency effect and observed differences are only due to trial-by-trial error. If q^2 is large then there is relatively small trial-by-trial error variance and observed differences between participants stem from reliable, true differences between participants. Crucially, note that q^2 is also the variance of true, individual sensitivities. Thus, the square root of this ratio, q, is the standard deviation of true, individual sensitivities.

$$Var[d'_{\text{true},i}] = Var\left[\frac{\Delta_i}{\sigma_\epsilon}\right] = \frac{\sigma^2_{\text{effect}}}{\sigma^2_{\epsilon}} = q^2 \quad \text{corresponding to} \quad SD[d'_{\text{true},i}] = q$$

We derive a reasonable value to use for our setting in Appendix E, which is $q^2 = 0.0225$. This means that we will assume that participants' sensitivities $d'_{\text{true},i}$ vary around some true value d'_{true} with a standard deviation of $q = 0.15$.

Relationship between sensitivity and accuracy. As we have already mentioned, some published studies report d' values, whereas other studies report %-correct values in the direct task. Because we would like to be able to work with either of them, we now discuss the relationship that can transform %-correct values into d' values and vice versa.

Recall that $d'_{\text{true},i}$ denotes the true sensitivity of participant i, and let us introduce the notation π_i for the true probability of a correct classification of a masked stimulus by participant i. We now make the assumption of a neutral criterion in the direct task, that is, participants are not inclined to guess one category of the masked stimuli (A or B) more often than the other. Under this assumption, the true relationship is $d'_{\text{true},i} = 2\Phi^{-1}(\pi_i)$ where Φ^{-1} is the inverse cumulative normal distribution (Macmillan

& Creelman, 2004). To simplify our later analysis, we now introduce the linear approximation $h(x) = 5(x - 0.5) \approx 2\Phi^{-1}(x)$. This approximation works remarkably well in the regime of sensitivities being close to zero:

$$\text{given } \pi_i \text{, we approximate } d'_{\text{true},i} \approx h(\pi_i) = 5(\pi_i - 0.5)$$

$$\text{given } d'_{\text{true},i} \text{, we approximate } \pi_i \approx h^{-1}(d'_{\text{true},i}) = \frac{1}{5} d'_{\text{true},i} + 0.5$$

For example, an accuracy of 54%-correct is approximately translated into the sensitivity $d'_{\text{true},i} \approx 5 \cdot (0.54 - 0.5) = 0.2$. This is very close to the exact translation, $d'_{\text{true},i} = 2\Phi^{-1}(\pi_i) = 0.201$. Table S1 shows that this approximation provides a very tight fit in the range of $\pi_i \in [0.4; 0.6]$ or, equivalently, $d'_{\text{true},i} \in [-0.5; 0.5]$. Larger values, that is, an accuracy above 60%-correct, would be at odds with the experimental setting in which direct task performance is assumed to be close to chance ($d'_{\text{true},i}$ close to 0 and π_i close to 0.5).

Table S1

Relation between the true accuracy (first column), the approximation of the sensitivity (second column) and the true sensitivity (third column). Note, that for π_i in the range of $[0.5, 0.6]$ *and D_i in the range of* $[0, 0.5]$ *(first six rows in the table) there is a very tight fit between $h(\pi_i)$ and $d'_{true,i}$. Negative values of $d'_{true,i}$ follow symmetrically.*

π_i	$h(\pi_i)$	$d'_{\text{true},i}$
0.50	0.000	0.000
0.52	0.100	0.100
0.54	0.200	0.201
0.56	0.300	0.302
0.58	0.400	0.404
0.60	0.500	0.507
0.62	0.600	0.611
0.64	0.700	0.717
0.66	0.800	0.825
0.68	0.900	0.935
0.70	1.000	1.049

Estimated sensitivity, $d'_{\text{estimated,direct}}$, from mean sensitivity index d'

We want to estimate the sensitivity and corresponding standard error from the typically reported direct task results. Usually, the average across individual sensitivity indices is reported as d'. This sensitivity index is already an estimate of the true sensitivity and we take it as it is (Macmillan & Creelman, 2004),

$$d'_{\text{estimated,direct}} = d'. \tag{5}$$

The standard error of d' is composed of two variances, one due to systematic variation between individuals' true sensitivities ($d'_{\text{true,i}}$) and the other due to non-systematic measurement error ($\epsilon_{d'_i}$). We use two simplifications: (a) We neglect dependencies between them because the variance of random error $\text{Var}[\epsilon_{d'_i}]$ does not change substantially for different sensitivity values in the relevant range, $D_i^{\text{dir}} \in [-0.5, 0.5]$; (b) We apply the approximation function h that relates d'_i to $\hat{\pi}_i$. This allows us to use the variance of the binomially distributed accuracies $\hat{\pi}_i$ from K trials, $\text{Var}[\epsilon_{\hat{\pi}_i}] = \pi_i(1 - \pi_i)/K$, and relate them back to the variance of d'_i, which leads to $\mathrm{Var}[\epsilon_{d'_i}] \approx 5^2\,\mathrm{Var}[\epsilon_{\hat{\pi}_i}]$.

$$
\begin{aligned}
SE_{\text{direct}} &= \sqrt{\mathrm{Var}\,[d']} \\
&= \sqrt{\mathrm{Var}\left[\frac{1}{N}\sum_i d'_i\right]} = \frac{1}{\sqrt{N}}\sqrt{\mathrm{Var}[d'_i]} = \frac{1}{\sqrt{N}}\sqrt{\mathrm{Var}[d'_{\text{true,i}} + \epsilon_{d'_i}]} \\
&\overset{(a)}{\approx} \frac{1}{\sqrt{N}}\sqrt{\mathrm{Var}[d'_{\text{true,i}}] + \mathrm{Var}[\epsilon_{d'_i}]} \overset{(b)}{\approx} \frac{1}{\sqrt{N}}\sqrt{\mathrm{Var}[d'_{\text{true,i}}] + 5^2\,\mathrm{Var}[\epsilon_{\hat{\pi}_i}]} \\
&= \underbrace{\frac{1}{\sqrt{N}}}_{\text{average}} \sqrt{\underbrace{q^2}_{\text{between subject variance}} + 5^2 \underbrace{\frac{\left(\frac{1}{5}d' + 0.5\right)\left(1 - \left(\frac{1}{5}d' + 0.5\right)\right)}{K}}_{\text{non-systematic error of } \hat{\pi}_i}}.
\end{aligned}
\tag{6}
$$

Without simplifications (a) and (b), one could construct an exact estimator. Exact calculations from Miller (1996) show that d' slightly overestimates the true sensitivity d'_{true} but that this bias is so small that the estimator can be considered approximately unbiased when typical sample sizes as in our context are used. On the other hand, our simplifications allow us to find a closed form solution that is simple to compute. Our estimators are well aligned with the true values, which we have shown by validating simulations in Appendix B.

Estimated sensitivity, $d'_{\text{estimated,direct}}$, from mean accuracy $\hat{\pi}$

Instead of d', some studies report the average classification accuracy $\hat{\pi}$ (%-correct) for the direct task. We estimate the sensitivity $d'_{\text{estimated,direct}}$ from the mean accuracy $\hat{\pi}$ by a plug-in estimator (Macmillan & Creelman, 2004),

$$
d'_{\text{estimated,direct}} = 2\Phi^{-1}(\hat{\pi}) \approx 5 \cdot (\hat{\pi} - 0.5) \tag{7}
$$

where Φ^{-1} is the inverse cumulative normal distribution. Exploiting the linearity of approximation h in (*), we can derive that this estimator is approximately unbiased:

$$
\mathrm{E}[d'_{\text{estimated,direct}}] = \mathrm{E}[2\Phi^{-1}(\hat{\pi})] \approx \mathrm{E}[h(\hat{\pi})] \overset{(*)}{=} h[\mathrm{E}(\hat{\pi})] = h(\pi) \approx 2\Phi^{-1}(\pi) = d'_{\text{true}}.
$$

Next, the standard error can be derived in the same fashion as for reported d' values so that we obtain:

$$
SE_{\text{direct}} = \frac{1}{\sqrt{N}}\sqrt{q^2 + 5^2\frac{\hat{\pi}(1-\hat{\pi})}{K}}. \tag{8}
$$

Estimated sensitivity, $d'_{\text{estimated,indirect}}$, from t and F values

Now let us move to estimating sensitivities from t values in the indirect task. We will show that an unbiased estimator is obtained from multiplying the t value by the constant c_{N,K,q^2}:

$$
d'_{\text{estimated,indirect}} = t \cdot c_{N,K,q^2} \quad \text{with} \quad c_{N,K,q^2} = \sqrt{\frac{q^2 + \frac{4}{K}}{N}}\sqrt{\frac{2}{N-1}}\,\frac{\Gamma\left(\frac{N-1}{2}\right)}{\Gamma\left(\frac{N-2}{2}\right)}, \tag{9}
$$

where Γ is the gamma distribution.

We start by considering how the t value in our setting is computed from the observed congruency effect:

$$
t = \frac{\hat{\Delta}}{\hat{\sigma}_{\hat{\Delta}_i}}\sqrt{N}
$$

We know that $\hat{\Delta} \sim \mathcal{N}\left(\Delta, (\sigma^2_{\text{effect}} + \frac{4}{K}\sigma^2_{\epsilon})/N\right)$ from above. Now we introduce independent random variables $Z \sim \mathcal{N}(0,1)$ and

$V \sim \chi^2(N-1)$ and rearrange t:

$$\begin{aligned} t &= \frac{Z\sqrt{\left(\sigma^2_{\text{effect}} + \frac{4}{K}\sigma^2_\epsilon\right)/N} + \Delta}{\sqrt{\left(\sigma^2_{\text{effect}} + \frac{4}{K}\sigma^2_\epsilon\right)}\sqrt{\frac{V}{N-1}}}\sqrt{N} \\ &= \left(Z + \Delta\sqrt{\frac{N}{\left(\sigma^2_{\text{effect}} + \frac{4}{K}\sigma^2_\epsilon\right)}}\right)\frac{\sqrt{N-1}}{\sqrt{V}}. \end{aligned}$$

We now use $\sigma^2_{\text{effect}} = q^2\sigma^2_\epsilon$ (also from above) to isolate σ^2_ϵ and obtain d'_{true}.

$$\begin{aligned} t &= \left(Z + \Delta\sqrt{\frac{N}{\left(q^2\sigma^2_\epsilon + \frac{4}{K}\sigma^2_\epsilon\right)}}\right)\frac{\sqrt{N-1}}{\sqrt{V}} \\ &= \left(Z + \frac{\Delta}{\sigma_\epsilon}\sqrt{\left(\frac{N}{q^2 + \frac{4}{K}}\right)}\right)\frac{\sqrt{N-1}}{\sqrt{V}} \\ &= \left(Z + d'_{true}\sqrt{\left(\frac{N}{q^2 + \frac{4}{K}}\right)}\right)\frac{\sqrt{N-1}}{\sqrt{V}} \end{aligned}$$

As a result, t follows a t distribution with degrees of freedom $df = N-1$ and non-centrality parameter $\delta = d'_{true}\sqrt{\frac{N}{q^2+\frac{4}{K}}}$. From Hogben, Pinkham, and Wilk (1961) and Hedges (1981), we know the expected t value to be

$$\begin{aligned} \mathrm{E}[t] &= \delta\sqrt{\frac{N-1}{2}}\frac{\Gamma(\frac{N-2}{2})}{\Gamma(\frac{N-1}{2})} \\ &= d'_{true}\cdot\sqrt{\frac{N}{q^2+\frac{4}{K}}}\sqrt{\frac{N-1}{2}}\frac{\Gamma(\frac{N-2}{2})}{\Gamma(\frac{N-1}{2})} \\ &= d'_{\text{true}}\cdot c^{-1}_{N,K,q^2}. \end{aligned}$$

In consequence, an unbiased estimator of d'_{true} is $d'_{\text{estimated, indirect}} = t\cdot c_{N,K,q^2}$.

As for the expected value, the variance of t values is also given by the properties of a non-central t distribution. Multiplying this variance by the constant c_{N,K,q^2} yields the variance of our estimated sensitivity $d'_{\text{estimated, indirect}}$. Since this depends on the non-centrality parameter, we use the plugin estimator

$$\begin{aligned} \hat{\delta} &= d'_{\text{estimated, indirect}}\sqrt{\frac{N}{q^2+\frac{4}{K}}} \\ &= t\cdot c_{N,K,q^2}\sqrt{\frac{N}{q^2+\frac{4}{K}}} \\ &= t\cdot\sqrt{\frac{2}{N-1}}\frac{\Gamma\left(\frac{N-1}{2}\right)}{\Gamma\left(\frac{N-2}{2}\right)} \end{aligned}$$

The standard error being its positive square root follows accordingly:

$$
\begin{aligned}
SE_{\text{direct}} &= \sqrt{\text{Var}[c_{N,K,q^2} \cdot t]} = c_{N,K,q^2} \sqrt{\text{Var}[t]} \\
&= c_{N,K,q^2} \sqrt{\frac{1+\hat{\delta}^2}{N-3} - \frac{\hat{\delta}^2 (N-1) \Gamma\left(\frac{N-2}{2}\right)^2}{2\Gamma\left(\frac{N-1}{2}\right)^2}} \\
&= c_{N,K,q^2} \sqrt{\left(1 + \frac{2t^2}{N-1}\left(\frac{\Gamma(\frac{N-1}{2})}{\Gamma(\frac{N-2}{2})}\right)^2\right)\left(\frac{N-1}{N-3}\right) - t^2} \qquad (10)
\end{aligned}
$$

With this, we can estimate the sensitivity and its standard error from a given t value in a repeated measures design.

Note that this approach can be applied identically to reported F values instead of t values. The reason is that in repeated measures ANOVA settings with two conditions the equality $|t| = \sqrt{F}$ holds. The main argument can be derived in the following equations using the standard definitions for the explained (SSE) and residual summed squares (SSR), see Winer et al. (1991); Maxwell and Delaney (2000):

$$
\begin{aligned}
t^2 &= \left(\frac{\hat{\Delta}}{\hat{\sigma}_{\hat{\Delta}_i}} \cdot \sqrt{N}\right)^2 = \frac{4 \cdot N \cdot (\hat{\Delta}/2)^2}{\frac{1}{N-1}\sum_{i=1}^N \left(\hat{\Delta}_i - \hat{\Delta}\right)^2} = \frac{2 \cdot N \cdot (\hat{\Delta}/2)^2}{\frac{1}{N-1}\sum_{i=1}^N \left(\frac{\hat{\Delta}_i - \hat{\Delta}}{2}\right)^2} \\
&= \frac{2 \cdot N \cdot (\hat{\Delta}/2)^2}{2\sum_{i=1}^N \left(\frac{\hat{\Delta}_i - \hat{\Delta}}{2}\right)^2 /(2N-2)} = \frac{\text{SSE}/df_{\text{E}}}{\text{SSR}/df_{\text{R}}} = F.
\end{aligned}
$$

Finally, note that this reanalysis for the indirect task can be extended to unbalanced settings in which the total number of trials K is not equally distributed to the two conditions for $M = K/2$ trials per condition but instead to M_1 and M_2 trials per condition $j = 1$ and $j = 2$, respectively. In these situations, one can analogously show that $\hat{\Delta} \sim \mathcal{N}\left(\Delta, (\sigma^2_{\text{effect}} + \frac{M_1+M_2}{M_1 M_2}\sigma^2_\epsilon)/N\right)$. Following the same steps as above, one would obtain an alternative constant that now depends on the split M_1 versus M_2 instead of only K.

$$
c_{N,M_1,M_2,q^2} = \sqrt{\frac{q^2 + \frac{M_1+M_2}{M_1 M_2}}{N}} \sqrt{\frac{2}{N-1}} \frac{\Gamma\left(\frac{N-1}{2}\right)}{\Gamma\left(\frac{N-2}{2}\right)}.
$$

As a sanity check, set $M_1 = M_2 = K/2$ and find $c_{N,M_1,M_2,q^2} = c_{N,K,q^2}$.

Confidence intervals for the difference in sensitivities

Based on the previous estimators, we now need to test for a significant difference between sensitivities in direct vs. indirect tasks. For this purpose we construct a 95% confidence interval around the difference $d'_{\text{difference}}$ while taking the standard error $SE_{\text{difference}}$ of the estimated difference into account:

$$
d'_{\text{difference}} = d'_{\text{estimated,indirect}} - d'_{\text{estimated,direct}} \qquad (11)
$$

$$
SE_{\text{difference}} = \sqrt{(SE_{\text{direct}})^2 + (SE_{\text{indirect}})^2} \qquad (12)
$$

$$
95\%\ \text{CI} = \left[d'_{\text{difference}} \pm z_{0.975} \cdot SE_{\text{difference}}\right], \qquad (13)
$$

where $z_{0.975} = 1.96$ is the 97.5% quantile of the normal distribution. If zero is included in the confidence interval, $0 \in 95\%$ CI, then there is not sufficient evidence for an ITA because the observed difference can be explained by measurement error in a situation where the true direct and indirect task sensitivities are equal. Only if the confidence interval lies above zero, that is 95% CI $= [a, b]$ and $a > 0$, there is evidence for the presence of an ITA.

Note that in this test we use quantiles z_α of the normal distribution and not quantiles of the t distribution. Using the t distribution would require to estimate the degrees of freedom, which is unnecessarily complicated for our approach. We use the quantiles of the normal distribution which leads to a more liberal test increasing the likelihood of confirming an ITA and following the benefit-of-the-doubt approach (see General Discussion).

E Estimating the Ratio q^2 of Between- vs. Within-Subject Variance

As we have seen in the reanalysis of direct and indirect task sensitivities, we need to know one parameter: q^2, a ratio of systematic vs. noise variance. This is not an artifact of our reanalysis but unavoidable.

What Does the Parameter q^2 Mean?

To see what this parameter means and why we need to estimate it, consider estimating the indirect task sensitivity $d'_{\text{estimated,indirect}}$ from t values. A t value is computed by dividing an observed effect by its standard error, $t = \bar{x}/SE$. In the indirect task, $\bar{x}$ may be the average congruency effect and SE the estimated standard error of congruency effects across participants. This standard error is influenced by two sources of variability: variance due to inter-individual differences in true congruency effects across participants (σ^2_{effect}) and variance due to trial-by-trial measurement error (σ^2_ϵ). We want to isolate the latter variance, σ^2_ϵ, because we want to estimate the underlying sensitivity $d'_{\text{true}} = \Delta/\sigma_\epsilon$ from the t value. Thus, we need to distinguish the two sources of variability. We do so by defining the ratio q^2:

$$q^2 = \frac{\sigma^2_{\text{effect}}}{\sigma^2_\epsilon}.$$

Note that this parameter is equal to the variance of individual true sensitivities, $q^2 = Var[d'_{\text{true},i}]$, see Supplement D. Therefore, it might be more intuitive to consider the un-squared parameter, which is the standard deviation of participants' true sensitivities, $q = \text{SD}[d'_{\text{true},i}]$.

Literature Review to Determine q^2 (Following Benefit-Of-The-Doubt Approach)

To estimate q^2, we consider multiple studies that either provide estimates or make explicit assumptions. All these studies yield a specific value, see our summary in Table S2, columns q^2 and q. For our reanalysis, we will use the largest plausible value, $q^2 = 0.0225$. Thus, we follow the benefit-of-the-doubt approach giving a previously established ITA the best chance to be confirmed in our reanalysis.

Table S2
We repeated our reanalysis of the indirect task sensitivity from Dehaene et al. (1998) (last column) based on the q^2 values from different studies. Larger values of q^2 increase the estimated, indirect task sensitivity. We took the largest plausible value our reanalysis method.

Study	q^2	q	Reanalysis of Dehaene et al. (1998) $d'_{\text{estimated, indirect}}$
ten Brinke et al. (2014)	**0.0020**	0.04	0.16
Our example study	**0.0074**	0.09	0.20
Rouder & Haaf (2018)	**0.0087**	0.09	0.21
Miller & Ulrich (2013)	**0.0121**	0.11	0.23
Jensen (2002)	**0.0142**	0.12	0.25
Ribeiro et al. (2016)	**0.0214**	0.15	0.28
Our assumption	**0.0225**	0.15	0.29

First, we estimated q^2 from the data of ten Brinke et al. (2014). This yielded $\hat{\sigma}^2_{\text{effect}} = (6.5\text{ ms})^2$ and $\hat{\sigma}^2_\epsilon = (144\text{ ms})^2$, which translates into an estimated ratio of $\hat{q}^2 = (6.5\text{ ms})^2/(144\text{ ms})^2 = 0.0020$.

Our replication based on Dehaene et al. (1998) produced estimates for the variances of $\hat{\sigma}^2_{\text{effect}} = (6.7\text{ ms})^2$ and $\hat{\sigma}^2_\epsilon = (78\text{ ms})^2$ translating into an estimated ratio of $\hat{q}^2 = 0.0074$.

Similarly, Rouder and Haaf (2018, p. 21) discuss the relation between the two sources of variance in psychophysics. Their formulas are identical to ours when changing the notation from σ^2_{effect} to σ^2_β and σ^2_ϵ to σ^2. They argue that reasonable values are $\sigma_{\text{effect}} = 28$ ms and $\sigma_\epsilon = 300$, which leads to $q^2 = \sigma^2_{\text{effect}}/\sigma^2_\epsilon = 0.0087$.

Other studies did not discuss the ratio between the two variances, σ^2_{effect} and σ^2_ϵ, but only the trial-by-trial error variability σ^2_ϵ. We can combine this with Dehaene et al. (1998) reporting the observed standard deviation of RT effects to be 13.5 ms.

This variability is constituted by $\hat{\sigma}^2_{\hat{\Delta}_i} = \hat{\sigma}^2_{\text{effect}} + \frac{4}{K}\hat{\sigma}^2_\epsilon = (13.5 \text{ ms})^2$. By knowing $\hat{\sigma}^2_\epsilon$ and the number of trials, K, we can rearrange the formula and estimate $\hat{\sigma}^2_{\text{effect}}$ and thereupon $\hat{q}^2$.

Miller and Ulrich (2013, p. 846, in their Table 3) suggested $\sigma_\epsilon = 96$ ms in a binary forced-choice task (without masked stimuli): Their error term E_k with variance $Var[E_k] = 91.5$ corresponds to the mean noise across 100 trials, see their Table 15. From this, we obtained $\sigma_\epsilon = \sqrt{Var[E_k] \cdot 100} = 96$ ms, as noted above. Combining this with Dehaene et al.'s results yields $\sigma_{\text{effect}} = 10.5$ ms and thereupon $q^2 = 0.0121$.

Jensen (1992, p. 877, Table 7 for task "Hick SS 2") reported an average estimate of $\hat{\sigma}_\epsilon = 91$ ms measured in $N = 863$ nine to twelve year olds yielding $q^2 = 0.014$. Ribeiro, Paiva, and Castelo-Branco (2016) report $\hat{\sigma}_\epsilon = 79$ ms in a speeded binary choice task without priming suggesting $q^2 = 0.021$. Even though the specific tasks and populations from these last two studies do not match Dehaene et al.'s setting exactly, it is plausible that variances are by and large comparable.

Given this range of parameter values, we use a value that is larger than any q^2 value reported in these studies:

$$q^2 = 0.0225 \quad \text{corresponding to} \quad q = SD[d'_{\text{true},i}] = 0.15.$$

By choosing this upper bound on q^2, we follow the benefit-of-the-doubt approach because large values of q^2 favor the ITA hypothesis in our reanalysis attributing more variance to σ^2_{effect} and less to σ^2_ϵ. This, in turn, increases our estimate of $d'_{\text{true}} = \Delta/\sigma_\epsilon$. For example, see how larger values of q^2 increase the estimated indirect task sensitivity from Dehaene et al. (1998) in the last column of Table S2. Hence, overestimating q^2 leads to an overestimation of the indirect task sensitivity increasing the chances of confirming an ITA.

How Would our Reanalysis Look Like With a Different q^2?

We have repeated our literature reanalysis from Figure 5 with different parameter values. In Figure S1, we show a more realistic reanalysis with $q^2 = 0.01$. Here, the picture resembles a null effect. In contrast, we show an overly optimistic reanalysis with $q^2 = 0.09$ in Figure S2, in which an ITA starts to emerge for many studies. However, even in this case there is no conclusive evidence for an ITA in most studies because confidence intervals for the sensitivity difference still include 0.

Note that, when assuming large q^2 values as in Figure 5 and S2, one cannot take the reanalysis result as *evidence for* an ITA. This is because large q^2-values bias our reanalysis in favor of finding an ITA. Only when we nevertheless do not find an ITA, these results can be meaningfully interpreted as *evidence against* an ITA. In order to establish evidence for an ITA, one would have to use smaller values for q^2 or, better yet, use the trial-by-trial data so that no assumption on q^2 is necessary. Otherwise, an apparent ITA result may only be due to the bias introduced by a large q^2.

We provide an online tool to perform our reanalysis with different values of q^2 at `http://www.ecogsci.cs.uni-tuebingen.de/ITAcalculator/`. There, we suggest three different values for q^2: To establish a lack of evidence for an ITA, we suggest $q^2 = 0.0225$ as in our reanalysis proper. This assumption rarely rejects evidence for an ITA if there is any. On the other hand, to establish evidence for an ITA, we instead suggest $q^2 = 0.0025$, which is more restrictive. Only when an ITA is established with a relatively small q^2 like this, we can be sure that it is a genuine ITA instead of being produced only due to a lenient assumption on q^2. Lastly, we suggest an intermediate value of $q^2 = 0.01$, which is suitable for an exploratory reanalysis. Note that depending on the exact experimental setup, different values of q^2 may be appropriate.

Overall Summary Regarding our Choice of q^2

Taken together, our replication, our simulations, and the literature review suggests that q^2 is clearly below 0.0225. We adopted this upper bound as our assumption because it increases the chances of finding an ITA, thereby, following the benefit-of-the-doubt approach. We use this assumption to show that evidence for an ITA is missing in many studies. To establish evidence for an ITA, the reanalysis would have to use smaller values to rule out the possibility that an ITA was only the product of the overestimation bias coming from a too large q^2.

Up to now, we have only discussed behavioral data (RTs) but we applied our reanalysis method also to EEG and fMRI data. The justification for this is that the relative noise level is even larger in single-trial event related potentials (ERPs) and blood-oxygen-level-dependent signals (BOLD signals) because they incorporate much more noise (Stahl, Gibbons, & Miller, 2010). Thus, the ratio between effect vs. noise variance in these measures will be even smaller, again, justifying our choice of $q^2 = 0.0225$.

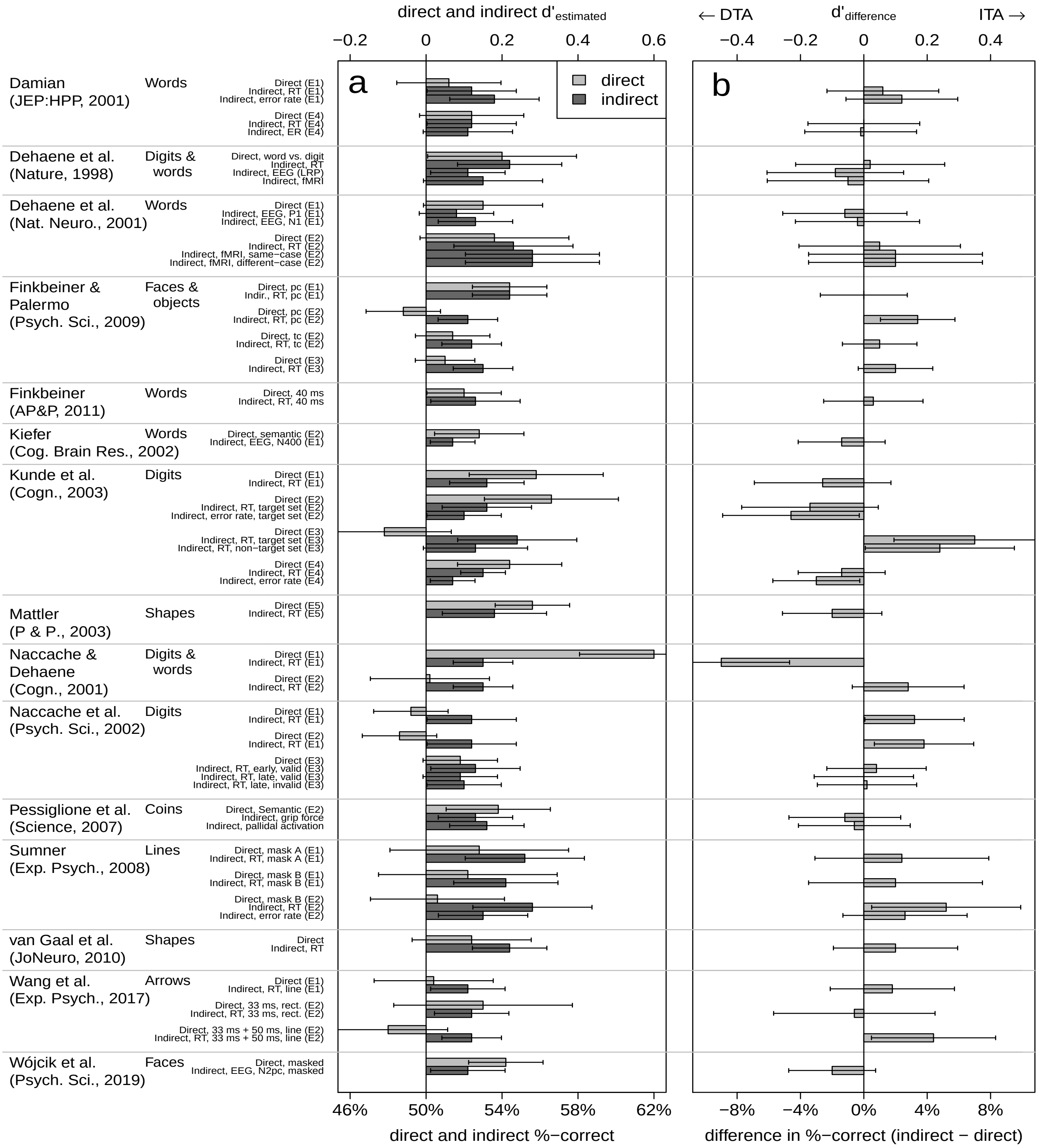


Figure S1. **Reanalysis with $q^2 = 0.01$.** Same as Figure 5 assuming that the standard deviation of true sensitivities across participants is $\text{SD}[d'_{\text{true},i}] = q = 0.1$. This assumption matches the results of our replication and is therefore more realistic but also more strict in dismissing results of an indirect task advantage (ITA). Here, only 7 reanalyzed ITAs are confirmed while 3 results yield the opposite result of a larger sensitivity in the direct task (direct task advantage [DTA]). Error bars represent 95%-confidence intervals.

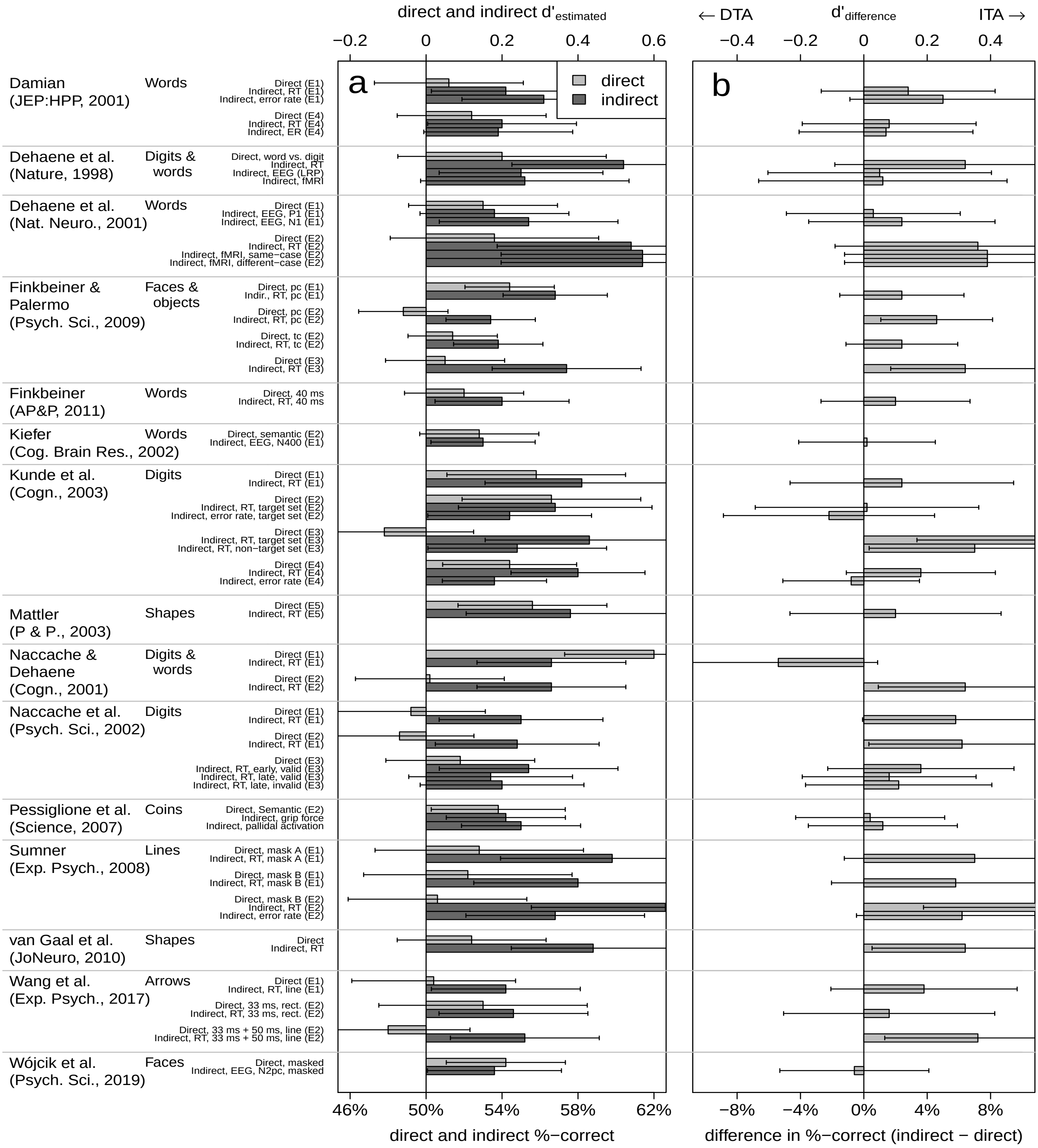


Figure S2. **Reanalysis with $q^2 = 0.09$.** Same as Figure 5 assuming that the standard deviation of true sensitivities across participants is $\mathrm{SD}[d'_{\mathrm{true},i}] = q = 0.3$. With this or even larger q^2, reanalyzed sensitivities tend to become clearly larger in the indirect compared to the direct task. However, this assumption is clearly unrealistic. First, in the direct task, this would mean that a substantial percentage of participants had a true sensitivity of $d'_{\mathrm{true},i} = 0.5$ or higher indicating that they could discriminate the masked stimuli better than 60%-correct. In the indirect task, an unrealistic implication of this assumption is that, in the study of Dehaene et al. (1998), trial-by-trial reaction times (RTs) would be estimated to vary with a standard deviation of only ±43 ms (within-subject variance $\sigma^2_\epsilon = 43^2$) even though RTs typically vary more than ±80 ms from trial to trial, see Appendix E.

F Details of Reanalyzed Studies

For each study, we give an overview of the study's structure, indicate in a table which values we extracted and explain our decisions for in- and exclusion of particular results. We only use results that follow the standard reasoning, claim an ITA and fit into our reanalysis method. We include quotes from the reanalyzed studies indicating their adherence to the standard reasoning. We use the following two abbreviations:

NR **N**ot **r**eanalyzable: Reported statistics do not match our reanalysis method. For example when the congruency factor has more than two levels (congruent, incongruent, and neutral) or when there are additional between-subject factors.

NIE **N**o **i**ndirect **e**ffect: The study attempted to find an ITA but failed due to a non-significant indirect task result. In such cases, the studies usually abort the standard reasoning, such that these cases are not relevant for us.

We report the number N of participants, the total number of trials K, and the reported statistic of the original study. Additionally, we report the sensitivities and standard errors according to our reanalysis. These are the values from Figure 5a. We then report the differences in sensitivities and their standard errors; here the difference is always taken between the current row's indirect task compared to the previously reported direct task. These results are presented in Figure 5b. We abbreviate Experiment 1 by E1, etc.

We also mark studies that excluded participants with good direct-task performance by adding the label **Regression to the mean** (see Discussion on why this is problematic). We still reanalyzed the reported results, although the exclusion introduced a bias for which our reanalysis method does not correct. This bias is liberal and favors finding an ITA. Thus, we follow the benefit-of-the-doubt approach.

15 Reanalyzed Studies

Damian (2001). The study reports four experiments but concludes an ITA only in Experiment 1 and 4. Experiments 2 and 3 were NIE.

Standard Reasoning: "Two control experiments investigated participants' ability to consciously perceive the masked primes. It was shown that performance was at chance level on both presence-absence judgments and on a number vs. random letter string discrimination task when the temporal characteristics of a trial were identical to those of the main experiment. Thus, the congruity effect described above must indeed have occurred outside of the participants' awareness" (p. 1).

	Original data			Our reanalysis (Figure 5)	
	N	K	Statistic	$d'_{\text{estimated}} \pm SE$	$d'_{\text{diff}} \pm SE_{\text{diff}}$
Direct (E1)	16	96	$d' = 0.064$	0.06 ± 0.07	
Indirect, RT (E1)	16	120	$F(1, 15) = 6.15$	0.14 ± 0.07	0.07 ± 0.10
Indirect, error rate (E1)	16	120	$F(1, 15) = 13.8$	0.21 ± 0.07	0.14 ± 0.10
Direct (E4)	16	96	$d' = 0.117$	0.12 ± 0.07	
Indirect, RT (E4)	16	120	$F(1, 15) = 5.67$	0.13 ± 0.07	0.02 ± 0.10
Indirect, ER (E4)	16	120	$F(1, 15) = 5$	0.13 ± 0.07	0.01 ± 0.10

Dehaene et al. (1998). The study reported two direct tasks and three indirect tasks. From the two direct tasks, we consider only the second direct task (word vs. digit discrimination) because it fits the neutral criterion assumption and it also shows lower sensitivity ($d' = 0.2$ in the first and $d' = 0.3$ in the second task). This way, we favor confirming the ITA hypothesis. For the first indirect measure, we computed the t value from the given estimates for the congruency effect ($M = 24$ ms and $SD = 13.5$). For the second indirect measure, the statistic ($t(11) < 3$) is taken from Figure 4, where the covert activation reflects processing of the prime as opposed processing of the target in the overt activation. For the third indirect measure, we only considered the congruency effect on fMRI the results are provided in Figure 5.

Standard Reasoning: "Under these conditions, even when subjects focused their attention on the prime, they could neither reliably report its presence or absence nor discriminate it from a nonsense string (Table 1). Nevertheless, we show here that the prime is processed to a high cognitive level [by demonstrating a priming effect]."

	Original data			Our reanalysis (Figure 5)	
	N	K	Statistic	$d'_{\text{estimated}} \pm SE$	$d'_{\text{diff}} \pm SE_{\text{diff}}$
Direct, word vs. digit	7	112	$d' = 0.2$	0.20 ± 0.11	
Indirect, RT	12	512	$t(11) = 6.16$	0.29 ± 0.09	0.09 ± 0.14
Indirect, EEG (LRP)	12	512	$t(11) = 3$	0.14 ± 0.06	-0.06 ± 0.12
Indirect, fMRI	9	128	$F(1, 8) = 6.23$	0.17 ± 0.10	-0.03 ± 0.14

Dehaene et al. (2001). The study reports two experiments. In E1, multiple measures assessed the visibility of the masked stimulus and we chose the reported binary forced-choice task (no stimulus vs. masked stimulus) because it is the most relevant result. In this experiment, the ITA refers to the absence vs. presence of the masked stimuli. The fMRI results in E1 were NR. In E2, the ITA referred to the congruency effect of repeated (congruent, either in same or in different case) vs. different words (incongruent).

Standard Reasoning: "Behaviorally, participants again denied seeing the primes and were unable to select them in a two-alternative forced-choice test [...]. However, case-independent repetition priming was observed in response times recorded during imaging [...]" (p. 755) and "As this phenomenon depends only on the identity of the masked prime, specific information about word identity must have been extracted and encoded unconsciously [...]" (p. 756).

	Original data			Our reanalysis (Figure 5)	
	N	K	Statistic	$d'_{\text{estimated}} \pm SE$	$d'_{\text{diff}} \pm SE_{\text{diff}}$
Direct (E1)	27	36	52.9%-correct	0.15 ± 0.09	
Indirect, EEG, P1 (E1)	12	300	$t(11) = 2.04$	0.10 ± 0.06	−0.04 ± 0.10
Indirect, EEG, N1 (E1)	12	300	$F(1, 11) = 9.79$	0.16 ± 0.07	0.01 ± 0.11
Direct (E2)	10	64	53.6%-correct	0.18 ± 0.11	
Indirect, RT (E2)	10	480	$F(1, 9) = 36$	0.30 ± 0.10	0.12 ± 0.15
Indirect, fMRI, same-case (E2)	10	240	$t(9) = 1.98$	0.34 ± 0.11	0.16 ± 0.16
Indirect, fMRI, different-case (E2)	10	240	$t(9) = 2.68$	0.34 ± 0.11	0.16 ± 0.16

Finkbeiner and Palermo (2009). The study reported four experiments. Prime and target stimuli were presented in different locations to the participants. In half of the trials the prime location was cued (pc) and in the other half it was the target location (tc). We excluded the target cued condition in E1 because it was NIE. In E3, multiple within-subject factors were tested but since those do not change the reported F value of the congruency effect we could nevertheless reanalyze it. E4 did not follow the standard reasoning.

	Original data			Our reanalysis (Figure 5)	
	N	K	Statistic	$d'_{\text{estimated}} \pm SE$	$d'_{\text{diff}} \pm SE_{\text{diff}}$
Direct, pc (E1)	40	80	$d' = 0.22$	0.22 ± 0.05	
Indirect, RT, pc (E1)	40	80	$F(1, 39) = 33.94$	0.24 ± 0.05	0.02 ± 0.07
Direct, pc (E2)	40	80	$d' = -0.06$	−0.06 ± 0.05	
Indirect, RT, pc (E2)	40	80	$F(1, 39) = 8.5$	0.12 ± 0.05	0.18 ± 0.07
Direct, tc (E2)	40	80	$d' = 0.07$	0.07 ± 0.05	
Indirect, RT, tc (E2)	40	80	$F(1, 39) = 10.6$	0.14 ± 0.05	0.07 ± 0.07
Direct (E3)	20	240	$d' = 0.05$	0.05 ± 0.05	
Indirect, RT (E3)	20	720	$F(1, 19) = 31.37$	0.20 ± 0.05	0.15 ± 0.07

Finkbeiner (2011, Regression to the mean). The study presented trials in two conditions, one with a short (40 ms) and one with a long (50 ms) prime presentation duration. An ITA was concluded only for the short duration and with respect to the semantic content (not color).

Standard Reasoning: "In contrast, 16 of the 21 subjects were judged to be at chance with the 40-ms primes. Following Rouder et al. (2007), the RTs for the 17 subject-by-prime-duration combinations for which subliminality was confirmed were entered into a paired-samples t test (two-tailed) to determine whether subliminal priming had occurred" (p. 1260).

	Original data			Our reanalysis (Figure 5)	
	N	K	Statistic	$d'_{\text{estimated}} \pm SE$	$d'_{\text{diff}} \pm SE_{\text{diff}}$
Direct, 40 ms	21	120	$d' = 0.098$	0.10 ± 0.06	
Indirect, RT, 40 ms	21	80	$t(20) = 2.5$	0.14 ± 0.06	0.04 ± 0.09

Kiefer (2002). The study reported two experiments. E1 reported the indirect task results and E2 reported the direct task results. In E1, indirect effects on RT, error rates and some EEG components were NR because the reported statistics combine masked and unmasked conditions (for unmasked conditions, they claimed no ITA) except for the N400 component in EEG. In E2, there were multiple direct tasks (see their Table 1). We chose the direct task on semantic judgment because the indirect task's congruency effect was an effect from semantic relatedness too.

Standard Reasoning: "Average d' measures in all tasks and context conditions did not deviate significantly from zero demonstrating that masked words were not identified" (p. 36).

	Original data			Our reanalysis (Figure 5)	
	N	K	Statistic	$d'_{\text{estimated}} \pm SE$	$d'_{\text{diff}} \pm SE_{\text{diff}}$
Direct, semantic (E2)	24	80	$d' = 0.14$	0.14 ± 0.06	
Indirect, EEG, N400 (E1)	24	320	$F(1, 23) = 5.48$	0.09 ± 0.04	−0.05 ± 0.08

Kunde et al. (2003). The study reported four experiments. In E1, there were multiple direct task measures from which we chose the one that fit our model assumptions of a neutral criterion (the identification rate is not comparable by our method). Also in E1, we chose not to consider sub-analyses of the indirect effects because they are essentially repetitions of the same comparison. In E2, we did not consider the non-target set condition and in E3 we did not consider the error rate analysis as they were NIE. In E1-E3, trials with neutral primes were not considered for calculating the priming effect.

Standard Reasoning: "The identification rate for the prime numbers was 2.2% (the chance level is 6.25% as each prime is presented four times in the 64 test trials). Thus, the primes were indeed unidentifiable, as is usually found under the experimental conditions that we adopted (Damian, 2001; Dehaene et al., 1998; Koechlin et al., 1999; Naccache & Dehaene, 2001)" (p. 230).

	Original data			Our reanalysis (Figure 5)	
	N	K	Statistic	$d'_{\text{estimated}} \pm SE$	$d'_{\text{diff}} \pm SE_{\text{diff}}$
Direct (E1)	12	64	$d' = 0.29$	0.29 ± 0.10	
Indirect, RT (E1)	12	1152	$F(1, 11) = 25.17$	0.22 ± 0.07	-0.07 ± 0.12
Direct (E2)	12	64	$d' = 0.33$	0.33 ± 0.10	
Indirect, RT, target set (E2)	12	288	$F(1, 11) = 15.24$	0.20 ± 0.07	-0.13 ± 0.12
Indirect, error rate, target set (E2)	12	288	$F(1, 11) = 6.35$	0.13 ± 0.06	-0.20 ± 0.12
Direct (E3)	12	64	$d' = -0.11$	-0.11 ± 0.10	
Indirect, RT, target set (E3)	12	144	$F(1, 11) = 21.67$	0.28 ± 0.09	0.39 ± 0.14
Indirect, RT, non-target set (E3)	12	144	$F(1, 11) = 6.58$	0.15 ± 0.08	0.26 ± 0.13
Direct (E4)	24	64	$d' = 0.22$	0.22 ± 0.07	
Indirect, RT (E4)	24	1152	$F(1, 23) = 43.2$	0.21 ± 0.05	-0.01 ± 0.08
Indirect, error rate (E4)	24	1152	$F(1, 23) = 9.17$	0.10 ± 0.04	-0.12 ± 0.08

Mattler (2003). The study reports five experiments. Only Experiments 3 and 5 are considered to be evidence for unconscious priming. Experiment 3 suffers severely from regression to the mean and is therefore not reanalyzed.

Standard Reasoning: "We might assume that performance at chance level indexes absence of all conscious information. This assumption was made in a number of studies (e.g., Dehaene et al., 1998; Klotz & Neumann, 1999; Neumann & Klotz, 1994; Vorberg et al., in press). In the present study, evidence for priming without awareness comes from Experiment 3 and Experiment 5, in which participants showed substantial non-motor priming effects although they could not discriminate primes better than chance" (p. 184)

	Original data			Our reanalysis (Figure 5)	
	N	K	Statistic	$d'_{\text{estimated}} \pm SE$	$d'_{\text{diff}} \pm SE_{\text{diff}}$
Direct (E5)	11	320	$d' = 0.28$	0.28 ± 0.06	
Indirect, RT (E5)	11	320	$F(1, 10) = 18.5$	0.22 ± 0.08	-0.06 ± 0.10

Naccache and Dehaene (2001). The study reports two experiments. For the direct tasks in both experiments, the authors additionally conducted the Greenwald method (Greenwald, Draine, & Abrams, 1996; Draine & Greenwald, 1998) which, however, has been criticized before (Dosher, 1998; Klauer, Greenwald, & Draine, 1998; Miller, 2000; Merikle & Reingold, 1998). Therefore, we only considered typical results as in all other studies. We considered only the main congruency effects on RT and no further subanalyses because the reported direct task would not have been comparable. In both experiments, an old and a new stimulus set were used. In E1, we only reanalyzed the RT effect based on the old stimulus set because the direct task sensitivity was estimated only for the old set. In E2, we reanalyzed the RT effect for the mixed, both new and old, stimulus set because the direct task sensitivity was estimated for this mixed set, too.

Standard Reasoning: "In this task, subjects performed at chance level, while priming effects were replicated. This study provides strong evidence for the unconscious nature of our semantic priming effects" (p. 227).

	Original data			Our reanalysis (Figure 5)	
	N	K	Statistic	$d'_{\text{estimated}} \pm SE$	$d'_{\text{diff}} \pm SE_{\text{diff}}$
Direct (E1)	18	32	$d' = 0.6$	0.60 ± 0.11	
Indirect, RT (E1)	18	384	$F(1, 17) = 21.99$	0.19 ± 0.06	-0.41 ± 0.12
Direct (E2)	18	64	$d' = 0.01$	0.01 ± 0.08	
Indirect, RT (E2)	18	384	$F(1, 17) = 21.62$	0.19 ± 0.06	0.18 ± 0.10

Naccache et al. (2002). The study reported three experiments. We did not consider the subanalyses for cued trials as the standard reasoning only related to the congruency effects. Note that we only counted the number of "critical" trials which were used in their analysis.

	Original data			Our reanalysis (Figure 5)	
	N	K	Statistic	$d'_{\text{estimated}} \pm SE$	$d'_{\text{diff}} \pm SE_{\text{diff}}$
Direct (E1)	12	240	$d' = -0.04$	-0.04 ± 0.06	
Indirect, RT (E1)	12	240	$F(1, 11) = 7.88$	0.15 ± 0.07	0.19 ± 0.09
Direct (E2)	12	240	$d' = -0.07$	-0.07 ± 0.06	
Indirect, RT (E1)	12	240	$F(1, 11) = 7.32$	0.14 ± 0.07	0.21 ± 0.09
Direct (E3)	12	240	$d' = 0.09$	0.09 ± 0.06	
Indirect, RT, early, valid (E3)	12	240	$F(1, 11) = 9.23$	0.16 ± 0.07	0.07 ± 0.09
Indirect, RT, late, valid (E3)	12	240	$F(1, 11) = 3.97$	0.11 ± 0.06	0.02 ± 0.09
Indirect, RT, late, invalid (E3)	12	240	$F(1, 11) = 5.34$	0.12 ± 0.07	0.03 ± 0.09

Pessiglione et al. (2007, Regression to the mean). The study deviated from the standard priming paradigm by just showing masked stimuli (in this case, coins) and no target stimuli. Presentation duration was varied in three conditions. For the separate conditions, participants were measured in one direct task and with three indirect measures. The appendix provided the required information for our reanalysis. We digitized their Figure S2 to derive the t values for the two indirect measures grip force and pallidal activation. The third indirect measure, skin conductance, was NIE. Even though these results were only reported in the appendix, the study bases their interpretation on these results. Note, that $N = 24$ relates to 24 participant $\times$ stimulus duration conditions in which the direct task was non-significant at an individual level.

Standard Reasoning: "Based on the percentage of correct responses, the analysis could then be restricted to all situations where subjects guess at chance level about stimulus identity (fig. S2) [by removing situations with significant direct task results]. Even in these situations, pallidal activation and hand-grip force were significantly higher for pounds as compared to pennies [...] " (p. 906).

	Original data			Our reanalysis (Figure 5)	
	N	K	Statistic	$d'_{\text{estimated}} \pm SE$	$d'_{\text{diff}} \pm SE_{\text{diff}}$
Direct, Semantic (E2)	24	60	$d' = 0.19$	0.19 ± 0.07	
Indirect, grip force	24	90	$t(23) = 2.92$	0.15 ± 0.06	-0.04 ± 0.09
Indirect, pallidal activation	24	90	$t(23) = 3.41$	0.17 ± 0.06	-0.02 ± 0.09

Sumner (2008, Regression to the mean). The study reported two experiments. Both, E1 and E2, had different mask conditions (A vs. B). Only E1 provided indirect task results such that we could reanalyze both conditions separately. For E2 we had to apply our reanalysis to both conditions aggregated. Therefore, we averaged over the given d' values from both conditions. We did not consider the subanalyses on the difference and interaction between the two masks but only the congruency effects as they are taken for the standard reasoning.

	Original data			Our reanalysis (Figure 5)	
	N	K	Statistic	$d'_{\text{estimated}} \pm SE$	$d'_{\text{diff}} \pm SE_{\text{diff}}$
Direct, mask A (E1)	12	40	$d' = 0.14$	0.14 ± 0.12	
Indirect, RT, mask A (E1)	12	200	$t(11) = 5.5$	0.30 ± 0.10	0.16 ± 0.15
Direct, mask B (E1)	12	40	$d' = 0.11$	0.11 ± 0.12	
Indirect, RT, mask B (E1)	12	200	$t(11) = 4.5$	0.25 ± 0.09	0.14 ± 0.15
Direct, mask B (E2)	12	80	50.5%-correct	0.03 ± 0.09	
Indirect, RT (E2)	12	400	$t(11) = 7.4$	0.36 ± 0.10	0.33 ± 0.14
Indirect, error rate (E2)	12	400	$t(11) = 4$	0.19 ± 0.07	0.17 ± 0.12

van Gaal et al. (2010, Regression to the mean). The study reported one experiment with one direct task and multiple indirect measures. However, we only considered the indirect effect on RTs as the fMRI analyses were NR.

Standard Reasoning: "[...] a, Participants were unable to discriminate between trials with a strongly masked square or diamond, as revealed by chance-level performance in a two-choice discrimination task administered after the main experiment. b, Although strongly masked no-go signals could not be perceived consciously, they still triggered inhibitory control processes, as revealed by significantly longer response times on these trials than on strongly masked go trials." (in Figure 2, p. 4145).

	Original data			Our reanalysis (Figure 5)	
	N	K	Statistic	$d'_{\text{estimated}} \pm SE$	$d'_{\text{diff}} \pm SE_{\text{diff}}$
Direct	20	48	$d' = 0.118$	0.12 ± 0.09	
Indirect, RT	20	240	$t(19) = 6.24$	0.27 ± 0.06	0.15 ± 0.11

Wang et al. (2017). The study reported two experiments. In E1, there were two outline conditions, line vs. rectangle. The line condition yielded a negative congruency effect which we treated similar to a standard (positive) priming effect. The rectangle condition was NIE. In E2, the rectangle condition with prime duration of 50 ms produced a large d' so that no ITA was claimed. Hence, we only considered the rectangle condition only for 33 ms. For the line condition, 33 ms and 50 ms trials were analyzed together since there was no interaction effect.

Standard Reasoning: "The results from the FC task indicated that similar prime visibility, equivalent to chance level, was obtained in the two preposed object type conditions. This finding confirmed that primes were processed subliminally in the primary task" (p. 425).

	Original data			Our reanalysis (Figure 5)	
	N	K	Statistic	$d'_{\text{estimated}} \pm SE$	$d'_{\text{diff}} \pm SE_{\text{diff}}$
Direct (E1)	15	64	$d' = 0.02$	0.02 ± 0.09	
Indirect, RT, line (E1)	15	208	$F(1, 14) = 6.86$	0.13 ± 0.06	0.11 ± 0.11
Direct, 33 ms, rect. (E2)	15	32	$d' = 0.15$	0.15 ± 0.12	
Indirect, RT, 33 ms, rect. (E2)	15	208	$F(1, 14) = 8.15$	0.14 ± 0.06	-0.01 ± 0.14
Direct, 33 ms + 50 ms, line (E2)	15	64	$d' = -0.103$	-0.10 ± 0.09	
Indirect, RT, 33 ms + 50 ms, line (E2)	15	416	$F(1, 14) = 11.47$	0.15 ± 0.06	0.25 ± 0.11

Wójcik et al. (2019). The study reported one experiment with masked and unmasked conditions. We only considered the masked condition for which an ITA was claimed but not the unmasked condition. In the direct task, we had to compute average d' from the openly accessible material. In the indirect task, EEG components were measured. For EEG preprocessing, some trials had to be rejected leading to an average of 131 trials. We assumed that rejection rate was approximately equal in the two indirect task conditions.

Standard Reasoning: "Analysis of the sensitivity measure d' indicated that faces were not consciously identified in the masked condition. A clear N2 posterior-contralateral (N2pc) component (a neural marker of attention shifts) was found in both the masked and unmasked conditions, revealing that one's own face automatically captures attention when processed unconsciously" (in the abstract, p. 471).

	Original data			Our reanalysis (Figure 5)	
	N	K	Statistic	$d'_{\text{estimated}} \pm SE$	$d'_{\text{diff}} \pm SE_{\text{diff}}$
Direct, masked	18	160	$d' = 0.211$	0.21 ± 0.06	
Indirect, EEG, N2pc, masked	18	131	$t(17) = 2.34$	0.12 ± 0.06	-0.09 ± 0.08

G Cost of Dichotomization in Significance Testing and Bayesian Analyses

The main fallacy of the standard reasoning persists independently of which statistical methods are chosen (significance testing or Bayesian analysis). It comes from evaluating the two tasks separately instead of using the appropriate analysis of measuring a difference between direct vs. indirect task sensitivities. To see why problems occur in both methods, consider the following simulation demonstrating the cost of dichotomization.

In multiple runs, we simulate one data set by sampling responses from $N = 12$ participants and $K = 256$ trials per participant. Thus, we sample $K/2 = 128$ observations in each of two conditions based on two normal distributions that are shifted by $d'_{\text{true}} = 0.15$ standard deviations (corresponding to a true performance of 53%-correct; log-normal distributions produce similar results). We analyze this one data set (a) as in the indirect task and (b) as in the direct task. We will show that both methods, significance testing and Bayesian analysis, produce misleading results in favor of the indirect task even though the *exact same data* is the basis for both tasks.

To mimic the RT effect from the indirect task, we tested the mean difference between two conditions against 0 (y axes in Figure S3). To mimic the direct task, we conducted a median split and tested sensitivity d' against 0 (x axes in Figure S3). The assumption here is that participants have access to the same information in both tasks and were forced to give a binary response (dichotomize) in the direct task so that the best they could do is to respond according to the optimal median split criterion. To test against 0, we used a t test (see Figure S3a) and we computed Bayes Factor (Figure S3b) using the R package provided by Morey, Rouder, Jamil, and Morey (2015)

Inspecting the results in Figure S3, we find that p values and Bayes Factors diverge from the red equality line indicating more evidence for an effect in the indirect task analysis compared to the direct task analysis. This is so because a median split dichotomization discards information (Cohen, 1983) producing larger p values and smaller Bayes Factors in the direct as compared to the indirect task.

In 23% of the simulations, there is a non-significant direct task vs. a significant indirect task result (shaded area in Figure S3a). This pattern may mislead researchers into thinking that there is an effect in the indirect task but none in the direct task. Note that this is a well–known error: One cannot take a non-significant result as evidence for the absence of an effect without a power analysis (see for example Vadillo et al., 2020).

The pattern of results from Bayes Factors is misleading in an even more severe way. In 20% of the simulations, we find Bayes Factors supporting the null hypothesis of no effect in the direct task ($BF_{10} < 1$) and simultaneously supporting the alternative hypothesis in the indirect task ($BF_{10} > 1$; on the log scale these are values below and above 0, see shaded area in

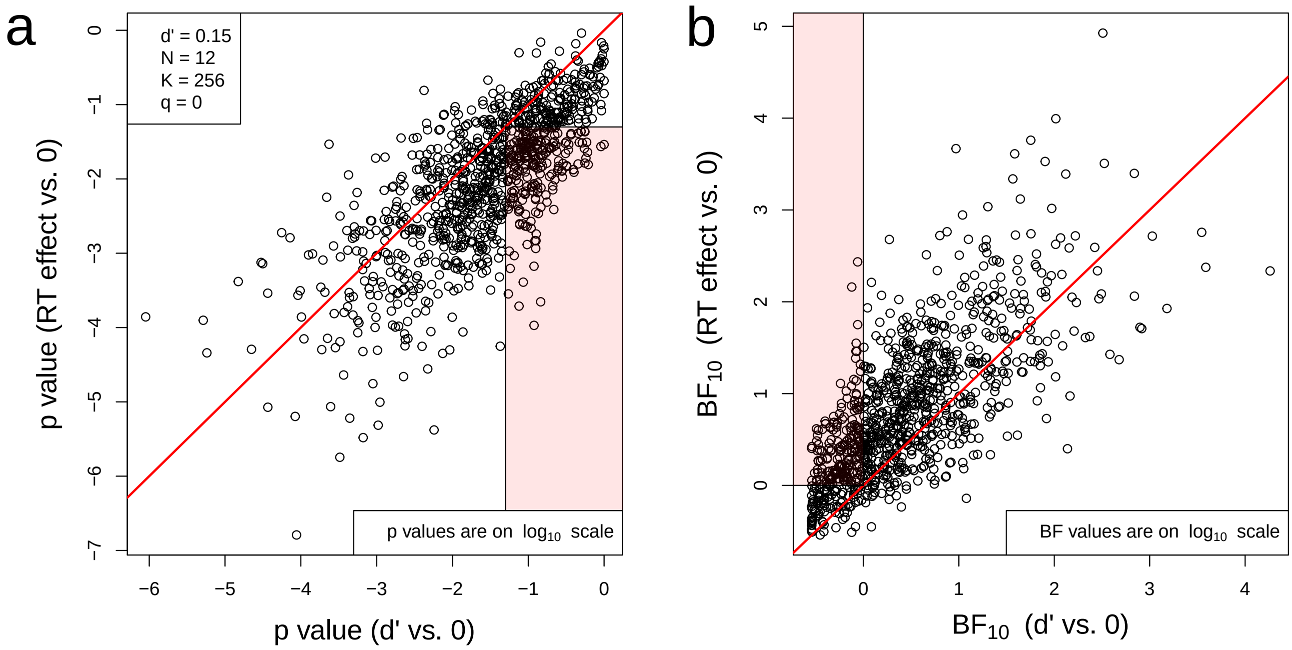


Figure S3. **Cost of Dichotomization in significance testing and Bayesian analysis.** Each point corresponds to one simulated data set. We analyzed each data set as in the direct task (x axis) and indirect task (y axes). We find that p values in (a) as well as Bayes Factors in (b) diverge from the red equality line indicating more evidence in the indirect task due to the loss of information from median splitting the data in the direct task. Shaded regions indicate a misleading pattern of result: (a) a significant indirect task vs. a non-significant direct task result; (b) a Bayes Factor supporting the null hypothesis in the direct task vs. a Bayes Factor supporting the alternative hypothesis in the indirect task.

Figure S3b). We even found some simulations, in which there is substantial evidence for the *null* hypothesis in the direct task ($BF_{10} < 1/3$) and substantial evidence for the *alternative* hypothesis in the indirect task ($BF_{10} > 3$). That is, if we ignored the main fallacy of the standard reasoning and followed the Bayesian analysis naively, we would conclude a difference in the two tasks even though the analyses in both tasks is based on the exact same data!

Analyzing the simulated data separately—computing mean difference in the indirect task and sensitivity in the direct task—produces misleading patterns of results. This problem occurs independent of the statistical methods used, significance testing or Bayes analysis, and even if the exact same data underlies both tasks. In a real experiment, direct and indirect tasks would not be based on the exact same data but on two samples, which produces additional measurement error. But in our idealized simulation here, there is no additional sampling error because both tasks are based on the same sample. Hence, no difference between the two tasks should be found. Accordingly, the appropriate analysis based on the sensitivity comparison would find exactly $d'_{\text{estimated, indirect}} - d'_{\text{estimated, direct}} = 0$ correctly identifying no difference between the two tasks and solving this problem.

H Glossary

Table S3
Description of variables.

Variable	Description
c_j	Condition effect, for example the congruent condition ($j = 1$) produces faster RTs so that $c_1 < 0$ and $c_2 = 1 - c_1 > 0$.
c_{N,K,q^2}	Constant relating t values to the estimated sensitivity in the indirect task, $d'_{\text{true}} = c_{N,K,q^2} \cdot t$. It depends on N, M and q.
d'	Observed, average sensitivity index, estimates the true sensitivity d'_{true}.
d'_i	Observed, individual sensitivity indices, estimates the true, individual sensitivities $d'_{\text{true},i}$.
d'_{true}	True sensitivity, $d'_{\text{true}} = \frac{\Delta}{\sigma_\epsilon}$.
$d'_{\text{estimated}}$	Estimated sensitivity from the reported summary statistics in the direct ($d'_{\text{estimated,direct}}$) or indirect task ($d'_{\text{estimated,indirect}}$).
$d'_{\text{true},i}$	Individual sensitivity, $d'_{\text{true},i} = \frac{\Delta_i}{\sigma_\epsilon}$.
Δ	The true difference between conditions, $\Delta = c_2 - c_1$.
$\hat{\Delta}$	The observed, mean difference between conditions.
Δ_i	True, individual effects, $\Delta_i = c_2 + (p \times c)_{i2} - (c_1 + (p \times c)_{i1})$, for example the expected congruency effect between conditions of participant i.
$\hat{\Delta}_i$	The observed difference between conditions of participant i.
ϵ_{ijk}	trial-by-trial error, noise due to measurement error or random neuronal fluctuations.
$f_{opt}(x)$	Optimal classifier taking indirect measures x (e.g., RTs) and predicting the condition (congruent/incongruent).
$f_t(x)$	Threshold classifier predicting one condition for indirect measures $x \leq t$ (e.g., RTs) and the other for $x > t$.
h	Linear approximation used to translate between sensitivities and accuracies.
i	Index for participant $i \in \{1, 2, ..., N\}$.
j	Index for condition $j \in \{1, 2\}$, for example indicator for congruent ($j = 1$) and incongruent ($j = 2$) conditions.
K	Total number of trials per particpant, $K = 2M$.
k	Index for trial $k \in \{1, 2, ..., M\}$. Since there are two conditions, the number of observed trials per participant is $2M = K$.
M	Number of trials per participant $\times$ condition. The total number of trials per participant is $2M = K$.
μ	Grand mean, for example the overall expected value of RTs.
N	Number of participants.
$\Omega(x)$	Marginal, cumulative density distribution (CDF) over indirect measures x.
p_i	Participant effect, for example participants with a faster RTs than average have a negative p_i while slower participants have a positive p_i.
$(p \times c)_{ij}$	Interaction effect, for example some participants have different reaction time effects.
π	True accuracy.
$\hat{\pi}$	Observed, mean accuracy.
π_i	True accuracy of participant i. It can be translated into a sensitivity by $d'_{\text{true},i} = 2\Phi^{-1}(\pi_i)$ where Φ is the cumulative normal distribution.
$\hat{\pi}_i$	Observed, individual accuracy.
q^2	Ratio between effect variance and trial-by-trial error variance, $q^2 = \frac{\sigma^2_{\text{effect}}}{\sigma^2_\epsilon}$. This is the variance of true sensitivities across individuals, $q^2 = \text{Var}[d'_{\text{true},i}]$. A reasonable value in our setting is $q^2 = 0.0225$ implying $\text{SD}[d'_{\text{true},i}] = 0.15$.

Table S3
(continued).

Variable	Description
SE	Estimated standard error of the estimated sensitivity.
$\sigma^2_{\Delta_i}$	Variance of true individual effects, for example, to which degree participants vary in their congruency effect.
$\sigma^2_{\hat{\Delta}_i}$	True variance of observed individual effects, for example, variance of the observable congruency effects.
$\hat{\sigma}^2_{\hat{\Delta}_i}$	Estimated variance of observed individual effects. This is what scientists get when computing the variance on the observable congruency effects across participants.
$\sigma^2_{p\times c}$	Variance of the interaction effect, $(p \times c)_{ij}$.
σ^2_{effect}	Variance of the effects Δ_i, $\sigma^2_{\text{effect}} = 4\sigma^2_{p\times c}$.
σ^2_{ϵ}	Variance of the trial-by-trial error, ϵ_{ijk}.
t	t value, in our context it comes from paired-t-tests between the two conditions of the indirect task.
Y_{ijk}	Response of participant i in condition j trial k from the direct (Y^{dir}_{ijk}) or indirect task (Y^{indir}_{ijk}). The standard repeated measures ANOVA model is $Y_{ijk} = \mu + p_i + c_j + (p \times c)_{ij} + \epsilon_{ijk}$.

I References

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, *7*(3), 249–253.

Damian, M. F. (2001). Congruity effects evoked by subliminally presented primes: Automaticity rather than semantic processing. *Journal of Experimental Psychology: Human Perception and Performance*, *27*, 154–165.

Dehaene, S., Naccache, L., Cohen, L., Bihan, D. L., Mangin, J. F., Poline, J. B., & Riviere, D. (2001). Cerebral mechanisms of word masking and unconscious repetition priming. *Nature Neuroscience*, *4*, 752–758.

Dehaene, S., Naccache, L., Le Clec'H, G., Koechlin, E., Mueller, M., Dehaene-Lambertz, G., … Le Bihan, D. (1998, October, 8th). Imaging unconscious semantic priming. *Nature*, *395*, 597–600.

Dosher, A. B. (1998). The response–window method — some problematic assumptions: Comment on Draine and Greenwald (1998). *Journal of Experimental Psychology: General*, *127*(3), 311–317.

Draine, S. C., & Greenwald, A. G. (1998). Replicable unconscious semantic priming. , *127*(3), 286–303.

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

Finkbeiner, M., & Palermo, R. (2009). The role of spatial attention in nonconscious processing: A comparison of face and nonface stimuli. *Psychological Science*, *20*, 42–51.

Franz, V. H., & von Luxburg, U. (2015). No evidence for unconscious lie detection: A significant difference does not imply accurate classification. *Psychological Science*, *26*(10), 1646–1648.

Greenwald, A. G., Draine, S. C., & Abrams, R. L. (1996). Three cognitive markers of unconscious semantic activation. *Science*, *273*, 1699–1702.

Hedges, L. V. (1981). Distribution theory for glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*(2), 107–128.

Hogben, D., Pinkham, R., & Wilk, M. (1961). The moments of the non-central t-distribution. *Biometrika*, *48*(3/4), 465–468.

Jensen, A. R. (1992). The importance of intraindividual variation in reaction time. *Personality and Individual Differences*, *13*(8), 869–881.

Kiefer, M. (2002). The N400 is modulated by unconsciously perceived masked words: Further evidence for an automatic spreading activation account of N400 priming effects. *Cognitive Brain Research*, *13*, 27–39.

Klauer, K. C., Greenwald, A. G., & Draine, S. C. (1998). Correcting for measurement error in detecting unconscious cognition: comment on Draine and Greenwald (1998). *Journal of Experimental Psychology: General*, *127*(3), 318–319.

Kunde, W., Kiesel, A., & Hoffmann, J. (2003). Conscious control over the content of unconscious cognition. *Cognition*, *88*(2), 223–242.

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide* (2nd ed.). Lawrence Erlbaum, New Jersey.

Mattler, U. (2003). Priming of mental operations by masked stimuli. *Perception & Psychophysics*, *65*(2), 167–187.

Maxwell, S. E., & Delaney, H. D. (2000). *Designing experiments and analyzing data: A model comparison perspective*. Lawrence Erlbaum: New Jersey.

Merikle, P. M., & Reingold, E. M. (1998). On demonstrating unconscious perception: comment on Draine and Greenwald (1998). *Journal of Experimental Psychology: General*, *127*(3), 304–310.

Miller, J. (1996). The sampling distribution of d'. *Perception & Psychophysics*, *58*(1), 65–72.

Miller, J. (2000). Measurement error in subliminal perception experiments: simulation analyses of two regression methods. , *26*(4), 1461–1477.

Miller, J., & Ulrich, R. (2013). Mental chronometry and individual differences: Modeling reliabilities and correlations of reaction time means and effect sizes. *Psychonomic Bulletin & Review*, *20*, 819–858.

Morey, R. D., Rouder, J. N., Jamil, T., & Morey, M. R. D. (2015). Package 'BayesFactor'. *URLh https://cran.r-project.org/web/packages/BayesFactor/BayesFactor.pdf (accessed May 21, 2021)*.

Naccache, L., Blandin, E., & Dehaene, S. (2002). Unconscious masked priming depends on temporal attention. *Psychological Science*, *13*(5), 416–424.

Naccache, L., & Dehaene, S. (2001). Unconscious semantic priming extends to novel unseen stimuli. *Cognition*, *80*, 215–229.

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

Pessiglione, M., Schmidt, L., Draganski, B., Kalisch, R., Lau, H., Dolan, R. J., & Frith, C. D. (2007). How the brain translates money into force: A neuroimaging study of subliminal motivation. *Science*, *316*, 904–906.

Ribeiro, M. J., Paiva, J. S., & Castelo-Branco, M. (2016). Spontaneous fluctuations in sensory processing predict within-subject reaction time variability. *Frontiers in Human Neuroscience*, *10*, 200.

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

Stahl, J., Gibbons, H., & Miller, J. (2010). Modeling single-trial LRP waveforms using gamma functions. *Psychophysiology*, *47*(1), 43–56.

Sumner, P. (2008). Mask-induced priming and the negative compatibility effect. *Experimental Psychology*, *55*(2), 133–141.

ten Brinke, L., Stimson, D., & Carney, D. R. (2014). Some evidence for unconscious lie detection. *Psychological Science*, *25*(5), 1098–1105.

Ulrich, R., & Miller, J. (1993). Information–processing models generating lognormally distributed reaction–times. *Journal of Mathematical Psychology*, *37*, 513–525.

Vadillo, M. A., Linssen, D., Orgaz, C., Parsons, S., & Shanks, D. R. (2020). Unconscious or underpowered? probabilistic cuing of visual attention. *Journal of Experimental Psychology: General*, *149*(1), 160.

van Gaal, S., Ridderinkhof, K. R., Scholte, H. S., & Lamme, V. A. (2010). Unconscious activation of the prefrontal no-go network. *Journal of Neuroscience*, *30*(11), 4143–4150.

Wang, Y., Wang, Y., Liu, P., Di, M., Gong, Y., & Tan, M. (2017). The role of representation strength of the prime in subliminal visuomotor priming. *Experimental Psychology*, *64*(6), 422–431.

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design* (3rd ed.). McGraw–Hill: Boston.

Wójcik, M. J., Nowicka, M. M., Bola, M., & Nowicka, A. (2019). Unconscious detection of one's own image. *Psychological Science*, *30*(4), 471–480.