

A critical reexamination of doing arithmetic nonconsciously

Pieter Moors¹ · Guido Hesselmann²

Published online: 1 May 2017
© Psychonomic Society, Inc. 2017

Abstract A recent study claimed to have obtained evidence that participants can solve invisible multistep arithmetic equations (Sklar et al., 2012). The authors used a priming paradigm in which reaction times to targets congruent with the equation's solution were responded to faster compared with incongruent ones. We critically reanalyzed the data set of Sklar et al. and show that the claims being made in the article are not fully supported by the alternative analyses that we applied. A Bayesian reanalysis of the data accounting for the random variability of the target stimuli in addition to the subjects shows that the evidence for priming effects is less strong than initially claimed. That is, although Bayes factors revealed evidence for the presence of a priming effect, it was generally weak. Second, the claim that unconscious arithmetic occurs for subtraction but not for addition is not supported when the critical interaction is tested. Third, the data do not show well-established features of numerosity priming as derived from V-shaped response time curves for prime-target distances. Fourth, we show that it is impossible to classify reaction times as resulting from congruent or incongruent prime-target relationships, which should be expected if their results imply that participants genuinely solve the equations on each trial. We conclude that the claims being made in the original article are not fully supported by the analyses that we apply. Together with a recent failure to replicate the original results

and a critique of the analysis based on regression to the mean, we argue that the current evidence for unconscious arithmetic is inconclusive. We argue that strong claims require strong evidence and stress that cumulative research strategies are needed to provide such evidence.

Keywords Interocular suppression · Unconscious processing · Reproducibility

Introduction

In their article, Sklar et al. (2012) claimed to have shown that participants can solve complex arithmetic equations nonconsciously, i.e., in the absence of consciously perceiving the equations. Specifically, they examined whether the presentation of multistep additions and subtractions with three single-digit operands (e.g., “ $9 - 3 - 4 =$ ”, “ $3 + 1 + 4 =$ ”) and without the result (2 and 8, respectively) would bias the verbal enumeration of a subsequently presented, visible number. Thus, the experiments by Sklar et al. were designed to test for “priming” effects, in which the exposure to a stimulus (or, prime) influences the response to a second stimulus (or target).¹ Following the seminal work by Meyer and Schvaneveldt (1971), priming has been widely used in the fields of cognitive psychology and neuroscience to infer the structure of semantic representations, including the representation of numerical values (Dehaene, Molko, Cohen, & Wilson, 2004; Knops, 2016). In the case of Sklar et al., target

✉ Pieter Moors
Pieter.Moors@ppw.kuleuven.be

¹ Laboratory of Experimental Psychology, Department of Brain & Cognition, University of Leuven (KU Leuven), Tiensestraat 102, box 3711, B-3000 Leuven, Belgium

² Visual Perception Laboratory, Department of Psychiatry and Psychotherapy, Charité-Universitätsmedizin, Berlin, Germany

¹ Please note that the study by Sklar et al. also included experiments on “non-conscious reading” that will not be addressed here. Instead of priming, Experiments 1-5 used a variant of interocular suppression (breaking CFS) in which the time to target detection is the dependent variable. The extent to which this paradigm can provide evidence for unconscious processing has been called into question, however (Stein & Sterzer, 2014).

numbers were either congruent or incongruent with the result of the prime equation. For example, the target number “2” is congruent with the result of the equation “ $9 - 3 - 4 =$ ”, while numbers “3” or “5” are incongruent. Repeated measures analysis of variance (rm-ANOVA) revealed significantly shorter response times (RTs) for congruent compared with incongruent priming trials. Rather surprisingly, this congruency priming effect was significant for subtractions, but not for additions.² The authors concluded from these data “that uniquely human cultural products, such as [...] solving arithmetic equations, do not require consciousness” (p. 19617).

For the following reasons, we believe that it is crucial to reexamine whether the claims made by Sklar et al. (2012) are fully supported by the available data. From a theoretical standpoint, the claim of “doing arithmetic nonconsciously” is a strong claim and, hence, demands strong evidence. Most cognitive scientists would agree that the complex nature of the underlying cognitive processes renders it implausible, rather than plausible, that effortful arithmetic operations may be performed without consciousness. Specifically, multistep additions and subtractions as used by Sklar et al. cannot be solved by declarative fact retrieval from long term memory alone. Successful performance would necessitate that arithmetic rules can be initiated and followed unconsciously, and that the unconscious intermediary results are stored in working memory. So far, there exists only a single study on unconscious addition making the former claim (Ric & Muller, 2012). Furthermore, considering the technical setup adds to the a priori implausibility of the effect reported by Sklar et al. They used continuous flash suppression (CFS) to render the prime equations invisible for up to 2 seconds. Following the introduction of this interocular suppression method (Tsuchiya & Koch, 2005), a very heterogeneous picture has emerged about the extent to which high-level unconscious visual processing is possible under CFS (Ludwig & Hesselmann, 2015; Sterzer, Stein, Ludwig, Rothkirch, & Hesselmann, 2014; Yang, Brascamp, Kang, & Blake, 2014). To the best of our knowledge, there is no evidence for Sklar and colleagues’ premise that CFS allows for more unconscious processing, because “it gives unconscious processes ample time to engage with and operate on subliminal stimuli” (p. 19614). On the contrary, it may rather be that long suppression durations (Experiment 6: 1,700 ms and 2,000 ms; Experiment 7: 1,000 ms and 1,300 ms) are associated with a particularly deep suppression of visual processing under CFS (Tsuchiya, Koch,

Gilroy, & Blake, 2006), and extended periods of invisible stimulation have been shown to lead to negative priming influences (Barbot & Kouider, 2011).

The goal of the current article is to provide a critical reexamination of the claims made by Sklar et al. (2012) based on the data they collected. We do so by approaching the original data set from different angles, and we assess whether the conclusions based on the original results still hold when taking the results of these new analyses into account. In the next sections, we provide five reanalyses of the data obtained by Sklar et al. First, we verified the repeatability of the reported analyses. This was a crucial first step that guaranteed we were analyzing the same data set as the original study. Second, we analyzed the data using a Bayesian linear mixed-effects models with crossed random effects for participants and stimuli, relying on Bayes factors to quantify how strongly the data support the predictions made by one statistical model compared to another. Indeed, throughout our reanalyses we are explicitly interested in quantifying the degree to which the data provide evidence for the claims that were made in the original study. It has been argued that classical significance testing approaches are not explicitly connected to statistical evidence, whereas the Bayes factor provides the possibility to quantify evidence in a coherent framework (Morey, Romeijn, & Rouder, 2016). This motivated us to rely on the Bayes factor throughout our reanalyses (except for the last one, see further). Third, Sklar et al. (2012) claimed that the congruency effect was observed for subtraction equations but not for addition equations. However, the interaction between congruency and operation was never tested although it is critical to ensure the congruency effect is different for subtraction compared to addition. Fourth, if the congruency effect observed for the subtraction equations was genuinely due to number processing, one would predict a distance effect to be present in the data. That is, as prime-target distance increases, response times should increase as well. Therefore, we assessed whether the data showed such distance-dependent priming effects. Fifth, Sklar et al. (2012) interpreted that the congruency effect indicated that participants unconsciously solved the equations. It has recently been argued that such a claim is only warranted if reaction times are predictive of prime-target congruency. That is, accurate classification of the prime-target relationship should be possible from the reaction time distributions.

For the sake of brevity, we report all reanalyses for Experiment 6 only in the main text of this article. We refer to the *Supplementary Materials* for the results of the reanalysis of Experiment 7, which were qualitatively the same. An extensive overview of all calculations is also available in the *Supplementary Materials*, including the code used to process the data and conduct the analyses.

² Priming effects for additions were only observed when Sklar et al. modified the experimental design and used a different dependent measure. In Experiment 9, equations with two single-digit operands (e.g., “ $8+7 =$ ”) were unconsciously presented, and participants had to report whether a subsequently presented visible addition equation with two single-digit operands and result (e.g., “ $9+6=15$ ”) was correct or not. The results showed that participants made significantly fewer mistakes in compatible trials (3.2%) than in incompatible trials (4.4%). We do not address this weaker finding in our article.

Methods

Data preparation

We obtained the data from Sklar et al. (Sklar et al., 2012). All data were processed and analyzed in R 3.3.2, a statistical programming language, and RStudio 1.0.44 (R Core Team, 2014; RStudio Team, 2015). A complete overview of this analysis can be found in the R markdown file in the *Supplementary Materials* (<https://doi.org/10.6084/m9.figshare.4888391.v2>). All data were visualized with the *yarr* package version 0.1.2 (Phillips, 2016).

Reanalysis #1: Repeatability of the reported analyses

We followed all data processing steps reported in Sklar et al. to compute mean response times for each participant – condition combination. We used the *afex* package version 0.16-1 to recalculate the repeated measures ANOVA by using the *aov_car()* function (Singmann, Bolker, Westfall, & Aust, 2016). Type III sums of squares were used, as these are default in many commercially available statistical packages and because one of these (SPSS) was used by Sklar et al. to analyze the data.

Reanalysis #2, #3, and #4: Bayesian linear mixed-effects models with crossed random effects

As linear mixed-effects models do not require a fully balanced data set, we used the trial-level (raw) data for all analyses. To account for the positive skew of the response time distributions, all analyses were performed on logarithmically transformed response times (as in Moors, Boelens, van Overwalle, & Wagemans, 2016; Moors, Wagemans, & de-Wit, 2016). All Bayes factors were calculated using the *BayesFactor* package version 0.9.12-2 (Morey, Rouder, Love, & Marwick, 2015). The Bayes factor refers to the ratio of marginal likelihoods of different statistical models under consideration (e.g., a model with a main effect of prime-target congruency versus an empty model with only random effects), quantifying the change from prior to posterior model odds:

$$\frac{p(M_1|D)}{p(M_2|D)} = \frac{p(M_1) p(D|M_1)}{p(M_2) p(D|M_2)}$$

where

$$BF_{12} = \frac{p(D|M_1)}{p(D|M_2)} = \frac{\int_{\theta_1} p(D|\theta) p(\theta) d\theta}{\int_{\theta_2} p(D|\theta) p(\theta) d\theta}$$

In itself, the Bayes Factor can be interpreted as a *relative* measure of evidence for one statistical model compared to

another. That is, the value of the Bayes Factor has no absolute meaning, and should always be interpreted relative to the statistical models under consideration. As Etz and Vandekerckhove (2016, p.4) put it: “The Bayes factor is most conveniently interpreted as the degree to which the data sway our belief from one to the other hypothesis.” Although the Bayes factor is inherently continuous, its values are sometimes partitioned into categories indicating different grades of evidence. For example, a Bayes factor of 3 or more often is associated with moderate evidence for one model, whereas Bayes factors larger than 10 are deemed strong evidence for that model. Bayes factors between 1/3 and 3 often are interpreted as providing equal support for both models, or anecdotal evidence for either model. We took these categories as guidelines, but we do not wish to fall prey to traditional accept/reject classifications, such as those that are standard in classical null hypothesis significance testing.

We used the *generalTestBF()* function to calculate the Bayes factors associated with the full model (i.e., including all fixed and random effects of interest) and most reduced versions of the full model (the *whichModels* argument was set to “withmain,” such that interaction effects were only included if the respective main effects also were included in the model) (Rouder, Engelhardt, McCabe, & Morey, 2016). With respect to the random effects, random intercepts were included for both participants and target stimuli. Initial analyses also were performed, including random slopes for participants and target stimuli for the congruency effect. However, models including random slopes were never favored over models, including random intercepts only. Therefore, we decided to drop random slopes altogether in the analyses reported. All default prior settings were used (i.e., a “medium” prior scale for the fixed effects ($r = 0.5$) and a “nuisance” prior scale for the random effects ($r = 1$)). Our general strategy of reporting the Bayes factors is as follows. We extracted the model with the highest Bayes factor compared to an empty model (i.e., an intercept-only model) and considered this to be the model that predicted the data best (in the following referred to as “best model”). We then recalculated all Bayes factors such that they are compared to this best model. In all tables, this yields an overview of the best model (Bayes factor = 1), and how strongly the data support the predictions made by this model compared to all other models. Because prior settings influence the Bayes factor, we also report on sensitivity analyses in the *Supplementary Materials* by varying the value of the prior scale of the fixed effects (which are of most interest here). We always included two models in the sensitivity analysis (i.e., yielding a single Bayes factor for each value of the prior scale). One of those was the best model, and in the other one the most important variable for the current reanalysis was included or excluded (depending on its inclusion in the best model). For example, if the best model contained only a main effect of prime-target congruency in addition to the random

effects, the sensitivity analysis would be conducted for this model and the model including random effects only.

Reanalysis #5: A significant difference does not imply accurate classification

For this analysis, we used the R code that was used in Franz and von Luxburg (Franz & von Luxburg, 2015), which is publicly available (<https://osf.io/7825t>). The classification analysis can be summarized as follows. The goal is to determine a threshold RT that can be used to classify RTs as either congruent or incongruent. In the case of the median classifier, the median RT is used as a threshold. In the case of the trained classifier, the data set is split into two halves: a training and a test set. For the training set, the threshold value is determined that leads to the fewest number of misclassifications, and this threshold is then applied to the test set. This was then repeated 10 times, and the average classification accuracy was taken as classification performance for the trained classifier.

Results

Reanalysis #1: Repeatability of the reported analyses

We first assessed the repeatability of the statistical analyses reported in Sklar et al. (2012). Repeatability entails taking the raw single-trial data, applying the same processing steps as outlined in the methods section and ending up with the same numbers reported by the authors in the original paper (Ioannidis et al., 2009). Importantly, repeatability without discrepancies implies that all following reanalyses are based on the same data set as used for the original publication. In short, this was the case (Fig. 1a). For the subtraction condition, we observed a significant priming effect ($F(1,15) = 16.79, p = 0.001$) and no interaction between prime-target congruency and presentation duration ($F(1,15) = 2.13, p = 0.17$). As reported, we did not observe a priming effect for the addition equations ($F(1,15) = 1, p = 0.33$). We refer to the *Supplementary Materials* for the repeatability of all other analyses reported by Sklar et al.

Reanalysis #2: (Bayesian) linear mixed-effects modeling with crossed random effects

Next, we applied an alternative statistical model that is generally considered more appropriate for experimental designs such as the one used in Sklar et al. (2012), linear mixed-effects modeling (LMM) with crossed random effects for participants and stimuli (Baayen, Davidson, & Bates, 2008; Clark, 1973). That is, in the experiments a sample of participants responds to a sample of stimuli (i.e., the target numbers)

in a set of different conditions. This induces variability in reaction times not only due to participants, but also due to targets. Traditionally, repeated measures ANOVAs or paired t tests are applied to data sets such as these to account for participant variability. However, the stimuli used in the experiment also induce stimulus variation, which is traditionally not accounted for and can substantially increase Type I error rates (Judd, Westfall, & Kenny, 2012). In psycholinguistics it is now standard to simultaneously model participants and items as random effects by relying on LMMs with crossed random effects (Baayen et al., 2008). Similar voices have been raised for social psychology (Judd et al., 2012; Wolsiefer, Westfall, & Judd, 2016), and these models have been applied in the CFS literature as well (Moors, Boelens, et al., 2016; Moors, Wagemans, et al., 2016; Stein, Kaiser, & Peelen, 2015). Thus, to ensure proper Type I error control, we implemented LMMs with crossed random effects throughout our reanalyses. Furthermore, in contrast to ANOVAs, LMMs do not require a fully balanced data set. As such, the raw trial-level data can be used to model RTs, rather than having to rely on mean RTs. Table 1 summarizes the results of the Bayes factor analysis applied to the subtraction condition of Experiment 6. In line with the original results, the results indicate that a model containing only a main effect of prime-target congruency is the model that predicted the data best (“best model”). The Bayes factors in Table 1 do provide some insight into the strength of this evidence, however. That is, the best model is not strongly favored over a model that does not include prime-target congruency (model 4) or one that includes no effects at all (model 3). Furthermore, the sensitivity analysis between model 1 and model 3 indicates that this pattern generalizes across all prior scales except for the smallest two considered ($r = 0.1$ and $r = 0.3$). Here, the Bayes factors indicate stronger evidence for a model including prime-target congruency.

Reanalysis #3: Analysis of the claimed “congruency x operation” interaction

Sklar et al. (2012) interpret the results of Experiments 6 and 7 as follows: “The results so far show that subtraction equations are solved nonconsciously and hence are sufficient to confirm our hypothesis that complex arithmetic can be performed unconsciously. However, why did not we find evidence for non-conscious solution of the easier-to-solve addition equations?” (p. 19616).

This interpretation of their results in terms of finding a congruency effect for subtraction but not for addition is based on comparing a difference in significance levels. However, such an interpretation is not warranted, because a difference in significance level does not imply a significant difference (Gelman & Stern, 2006; Nieuwenhuis, Forstmann, & Wagenmakers, 2011). With respect to the study of Sklar

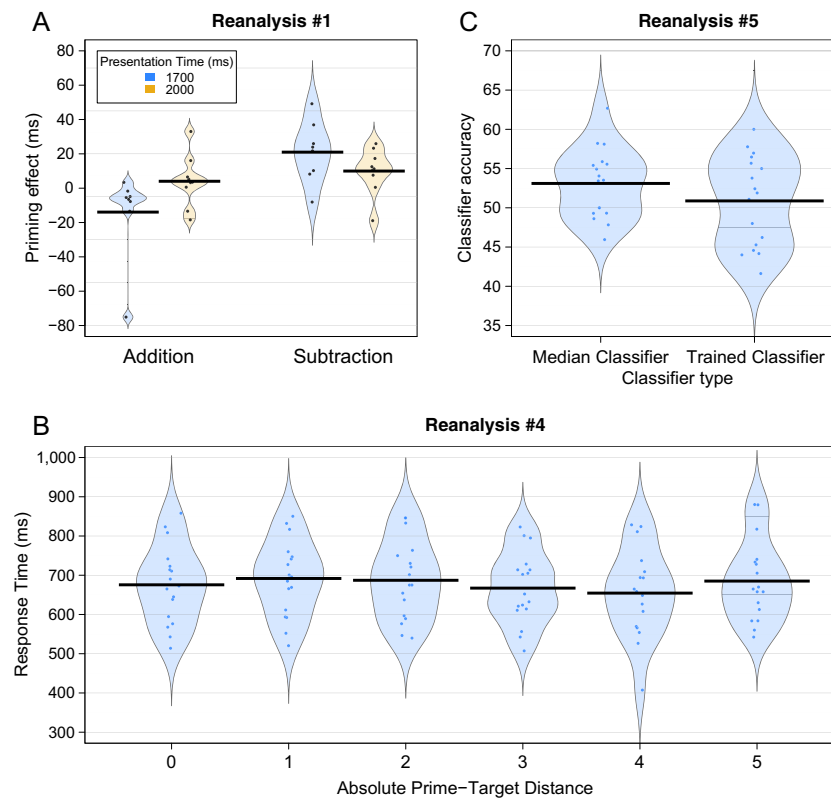


Fig. 1 Overview of the reanalyses (Experiment 6). **(a)** Reanalysis #1. Repeatability of the reported results. Priming effects (i.e., difference between incongruent and congruent condition, in ms) are depicted for the addition and subtraction equations, for both prime presentation durations. **(b)** Reanalysis #4. Effect of prime-target distance on response times. Response times (ms) are depicted in function of

absolute prime-target distance. **(c)** Reanalysis #5. Classification of response times. Classification accuracy is depicted in function of the classifier type, median or trained. In all plots, thick lines indicate the arithmetic mean, dots depict data from individual participants, and the beans indicate fitted densities (Phillips, 2016)

et al., an interpretation of the results in terms of the possibility of unconscious arithmetic for subtraction but not for addition thus requires an explicit test of the interaction between prime-target congruency and operation. In any case, for an experimental design such as the one used in Sklar et al., it would be

Table 1 Bayes factor analysis of the subtraction data (Experiment 6)

Model	BF	Error
1. Congruency + Subject + Target	1	0
2. Congruency + Presentation duration + Subject + Target	1.16	0.07
3. Subject + Target	2.15	0.03
4. Presentation duration + Subject + Target	2.65	0.03
5. Congruency * Presentation duration + Subject + Target	10.87	0.05
All other models	>100	NA

Note. All reported BFs are calculated with the model that predicted the data best (“best model”) in the numerator (the model for which BF = 1, top row). BFs > 1 therefore indicate how much more the data are consistent with this best model compared to the model under consideration in each row. A “*” indicates that both the main effects and the interaction between the variables are incorporated in the model.

most straightforward to first analyze the 2 x 2 x 2 design as a whole (prime-target congruency x prime presentation duration x operation). Depending on the presence of interactions, the analysis could then be broken up into simpler ones. We first analyzed the data using a 2 x 2 x 2 rm ANOVA. The critical interaction indeed proved to be significant ($F(1,15) = 9.88, p = 0.007$). However, as highlighted before, such an analysis does not properly control the Type 1 error rate. Therefore, we subjected the full data set to the Bayesian LMM analysis, the results of which are reported in Table 2. Two things are notable. First, in contrast with the results of the rm-ANOVA, none of the models reported in Table 2 includes an interaction between prime-target congruency and operation (or any other factor). Second, the best model does not include an effect of prime-target congruency. Although it is not strongly favored compared to models that do include an effect of prime-target congruency, the evidence for the absence of a congruency effect is notably stronger here compared to the evidence for the presence of a congruency effect in Reanalysis #2. Furthermore, a sensitivity analysis between models 1 and 4 indicates that this pattern generalizes across all prior scales considered (i.e., the Bayes factor never switches toward

Table 2 Bayes Factor analysis for the full data set (Experiment 6)

Model	BF	Error
1. Target + Subject	1	0
2. Presentation duration + Target + Subject	1.30	0.02
3. Operation + Target + Subject	4.57	0.01
4. Congruency + Target + Subject	5.28	0.02
5. Presentation duration + Operation + Target + Subject	5.83	0.07
6. Congruent + Presentation duration + Target + Subject	7.21	0.02
All other models	>100	NA

Note. All reported BFs are calculated with the model that predicted the data best (“best model”) in the numerator (the model for which BF = 1, top row). BFs > 1 therefore indicate how much more the data are consistent with this best model compared to the model under consideration in each row.

indicating evidence in favor of the model including prime-target congruency).

Reanalysis #4: The effect of numerical prime-target distance on response times

We now turn to a reanalysis that was motivated by theory-based predictions from the number processing literature. The priming model proposed by Sklar et al. (2012) can be summarized as follows: the arithmetic equation is solved, then the result (i.e., a number) is mentally represented, and this short-lived numerical representation influences the participant’s response to a subsequently presented number (i.e., the target). For symbolic primes and targets (e.g., Arabic numbers), it is commonly found that as numerical prime-target distance increases, priming effects decrease in a V-shaped manner (Reynvoet, Brysbaert, & Fias, 2002; Roggeman, Verguts, & Fias, 2007). This well-established feature of numerosity priming is generally explained by representational overlap between the prime and the target (Van Opstal, Gevers, De Moor, & Verguts, 2008). Thus, if the data reported in Sklar et al. (2012) involve genuine number processing, one should observe a V-shaped priming function for prime-target distances. To test for a V-shape, the absolute value of the prime-target distance was taken as a predictor. A positive slope, significantly different from zero, would indicate that prime-target distance influenced RTs in a V-shape manner. Table 3 shows the results of the Bayes factor analysis. As is apparent, distance was only included as a predictor in model 6, for which the evidence was considerably less strong compared to the best (empty) model (see also Fig. 1b). For the sensitivity analysis we used models 1 and 6. Here, the pattern of Bayes factors also generalized across all different prior scales.

Table 3 Bayes factor analysis of the prime-target distance (Experiment 6)

Model	BF	Error
1. Subject + Target	1	0
2. Presentation duration + Subject + Target	1.34	0.01
3. Operation + Subject + Target	4.40	0.04
4. Presentation duration + Operation + Subject + Target	6.03	0.02
5. Presentation duration*Operation + Subject + Target	9.12	0.06
6. Distance + Subject + Target	14.67	0.01
All other models	>100	NA

Note. All reported BFs are calculated with the model that predicted the data best (“best model”) in the numerator (the model for which BF = 1, top row). BFs > 1 therefore indicate how much more the data are consistent with this best model compared to the model under consideration in each row. A “*” indicates that both the main effects and in interaction between the variables are incorporated in the model.

Reanalysis #5: A significant difference does not imply accurate classification

In the abstract, Sklar et al. (Sklar et al., 2012) interpret the priming effects observed in their study as follows: “[...] multistep, effortful arithmetic equations can be solved unconsciously” (p. 19614). However, the result from which this claim was derived is the mean RT difference between the congruent and the incongruent condition (i.e., the priming effect). As recently highlighted by Franz and von Luxburg (Franz & von Luxburg, 2014, 2015), this type of analysis is not sufficient to make these claims. Specifically, the claim that participants are able to nonconsciously solve equations implies that the RTs should be predictive of the prime-target congruency with which they were associated. The idea behind the approach by Franz and von Luxburg is to leave the framework of classical null hypothesis significance testing on mean RT data behind and ask how much information about prime-target congruence is available in the RT data. This empirical question is addressed by means of trial-by-trial classification: If the significant prime-target congruency effect on RTs is supposed to serve as evidence for “good” unconscious processing, then we should be able to use the RTs to decide for each trial whether the prime (i.e., the number representing the solution of the equation) and the target (i.e., the visible target number) were congruent or incongruent. According to the underlying priming model, small RTs would indicate a congruent trial, large RTs would indicate an incongruent trial. To test how good classification performance of the prime-target relationship was for the subtraction condition, we used two different classifiers, a median and a trained one (Table 4; Fig. 1c). The median classifier assumes that the data follow a log(normal) distribution, whereas the trained classifier is a standard distribution-free classifier from the machine learning literature. As can be derived from Table 4, the trained

Table 4 Classification performance for the subtraction data (Experiment 6)

Classifier	Mean	SD	SEM	Median
Median classifier	53.10	4.43	1.07	53.52
Trained classifier	50.88	5.76	1.40	51.89

classifier performs completely at chance (50.88%). The median classifier performs only slightly above chance (53.10%). This indicates that the congruency effect for the subtraction data is only associated with poor classification of the prime-target congruency based on the RTs.

Discussion

In this article, we critically reanalyzed the data reported in Sklar et al. (2012). We first established that all analyses were repeatable without any discrepancies (Reanalysis #1). For making their data available to us and their research transparent, the authors should be applauded. Indeed, recent empirical evaluations have shown that the published biomedical literature generally lacks transparency, including public access to raw data and code (Goodman, Fanelli, & Ioannidis, 2016; Iqbal, Wallach, Khoury, Schully, & Ioannidis, 2016; Leek & Jager, 2016). Furthermore, a recent series of studies has indicated that half of the published psychology papers include at least one statistical inconsistency, and one in eight even a gross inconsistency (Nuijten, Hartgerink, van Assen, Epskamp, & Wicherts, 2015). Importantly, the full repeatability ensured that our following reanalyses were based on exactly the same data set as used for the original publication.

We then applied four different analyses to the data set, inspired by methodological, statistical, and theoretical considerations. The results of these analyses can be summarized as follows:

- When applying a statistical model that provides better control for the Type I error rate for the experimental design at hand, we showed that the evidence in favor of the presence of a congruency effect was not as strong as would be derived from the analyses that were reported in the original article, albeit the fact that the best model did include an effect of prime-target congruency (Reanalysis #2). Thus, on purely statistical grounds, this result shows that merely accounting for item variability substantially attenuates the strength of the evidence for the reported priming effect. In essence, this does not contradict the result reported by Sklar et al. Nevertheless, we argue that the strength of the evidence provides an important nuance to the interpretation of these results.
- The data do not strongly support the claim that unconscious arithmetic can happen for subtraction equations,

but not for addition equations. That is, none of the models reported in Reanalysis #3 included an interaction between prime-target congruency and operation. Moreover, all BF analyses also indicated that the data were more consistent with statistical models not including an effect of prime-target congruency. Thus, an analysis based on the full data set rather than different subsets of the data revealed that no strong evidence for main effects of or interactions with prime-target congruency was observed.

- No characteristic patterns of number processing, which have repeatedly and robustly been reported in the literature, are present in the current data set (Reanalysis #4). This indicates that one should be very cautious to invoke mechanisms related to number processing to explain these results.
- Even if the priming effect is taken at face value after the results of the three previous reanalyses, the data set does not provide evidence that participants unconsciously solved the equations that were presented subliminally. That is, the classification of the prime-target congruency based on the reaction times is nearly at chance, calling into question the assertion that people can solve equations nonconsciously (Reanalysis #5). Although the median classifier performed slightly above chance, its performance was still considerably lower than the performance that was taken to be the cutoff for establishing invisibility of the prime equations (60%).

Taken together, we conclude that the converging nature of all four reanalyses indicates that the data used for invoking the existence of unconscious arithmetic contain little evidential value for those claims (i.e., evidential value in terms of the Bayes factors obtained in the reanalyses). Within the conceptual framework proposed by Goodman et al. (2016), our reanalyses therefore suggest low *inferential reproducibility* of the study by Sklar et al.

A critical reviewer suggested that, based on the results of our reanalyses, one would expect that the original findings would not replicate easily. In this context, a direct replication of the study by Sklar et al., using the same experimental setup and exactly the same stimulus material, would be very informative. This was the goal of the recent study by Karpinski, Yale, and Briggs (2016). The authors used exactly the same materials as in Sklar et al. and aimed at replicating the original effect in a larger sample ($n = 94$). Interestingly, they obtained evidence for unconscious addition, but not subtraction (i.e., opposite findings compared to Sklar et al.). As this data set would be very informative for our reanalysis, we contacted the authors of this replication study. Upon reanalyzing the data set together with the authors, it became apparent, however, that a coding error led to an incorrect calculation of the mean RTs. A corrected analysis of the data did not reveal any priming effects for unconscious additions or subtractions (Karpinski & Briggs, personal communication). That is, the critical paired

comparisons for assessing priming effects for addition and subtraction both no longer passed the threshold for statistical significance (addition: $t(93) = 0.11$, $p = 0.92$; subtraction: $t(93) = 0.23$, $p = 0.82$). The paper has now been retracted. Thus, the single published replication study of the unconscious addition and subtraction effects reported in Sklar et al. actually failed to replicate the original pattern of results. Together with the results of our reanalyses, we argue that the results of this nonreplication calls for caution when interpreting the original results.

Exploring the scope and limits of non-conscious processing is essential for the formulation of theories of consciousness (Dehaene, Charles, King, & Marti, 2014). Since the results reported in Sklar et al. (2012) might have important implications for theories of (un)conscious processing (Dehaene et al., 2014; Koch, Massimini, Boly, & Tononi, 2016; Soto & Silvanto, 2014), we were motivated to conduct this critical reanalysis. If the reported effect is true, it can indeed be considered as an extraordinary case of subliminal perception and, as Sklar et al. argue, it might even “call for a significant update of our view of conscious and unconscious processes” (p.19614). In line with this notion, the senior author of this study recently suggested that “unconscious processes can carry out every fundamental high-level function that conscious processes can perform” (p. 195) (Hassin, 2013). Nonconscious arithmetic would be the most recent culmination point in a decades-long debate among cognitive scientists about the existence and potency of subliminal perception (Doyen, Klein, Simons, & Cleeremans, 2014). This debate has been characterized by a repeating cycle of provocative claims followed by methodological criticism, primarily aimed at the psychophysical and statistical methods used to establish the absence of conscious perception (Hesselmann & Moors, 2015). Of note, for the purpose of this reexamination, we solely relied on the data that were used to claim the existence of nonconscious arithmetic. For example, we simply took at face value that the post hoc selected sample of participants, whose data were submitted to statistical analysis, did not see the arithmetic equations; the crucial aspect of post hoc data selection and its implications has been treated elsewhere (Shanks, 2016).

The scientific study of consciousness has traditionally sought to assemble an exhaustive inventory of the psychological processes that can proceed unconsciously to isolate those that are exclusively restricted to conscious cognition (Naccache, 2009). During the course of the last decades, a large body of empirical evidence has been accumulated by applying this strategy, in particular within the domain of visual perception. Vision research provides a wide range of paradigms designed to transiently suppress visual stimuli from conscious perception, i.e., render a physically present target stimulus invisible for neurologically intact observers (Bachmann, Breitmeyer, & Ogmen, 2007). These paradigms differ with respect to what types of visual stimuli can be suppressed from

awareness, and how effective the suppression is in terms of duration and controllability of onset and offset (Kim & Blake, 2005). Along another dimension, the available paradigms may be placed within a functional hierarchy of unconscious processing, according to the extent to which features of visual stimuli are processed on an unconscious level and still induce effects on behaviour, e.g., in priming experiments (Breitmeyer, 2015). The results of our reanalysis can be framed into an emerging series of results that indicate that unconscious processing associated with CFS is not as high-level as previously thought (Hedger, Adams, & Garner, 2015; Hesselmann, Darcy, Sterzer, & Knops, 2015; Hesselmann & Knops, 2014; Moors, Boelens, et al., 2016) and that neural activity related to stimuli suppressed by CFS is considerably reduced already in early visual areas (Fogelson, Kohler, Miller, Granger, & Tse, 2014; Yuval-Greenberg & Heeger, 2013). Importantly, building such a functional hierarchy should eventually allow to formulate predictions on the level of unconscious processing that can be expected in a specific experimental setup. In the absence of prior assumptions on the depth of visual suppression associated with a specific paradigm, every new report of high-level unconscious processing seems equally plausible, and the boundaries of nonconscious processing are ultimately pushed further and further.

In sum, as extraordinary claims require extraordinary evidence, we were motivated to reanalyze the data obtained in Sklar et al. based on statistical, methodological, and theoretical considerations, and within a framework that allowed us to quantify the evidence for statistical models that reflected theoretical claims (i.e., unconscious arithmetic revealed through priming effects). Together with the recent non-replication of the original results (Karpinski et al., 2016) and a recent critique of the post-hoc selection of unaware participants that was used in the original study (Shanks, 2016), we argue that our results indicate that the evidence for the existence of unconscious arithmetic is inconclusive. This current state of affairs can only be overcome by cumulative research strategies, explicitly aimed at assessing the robustness of the findings and quantifying the strength of the evidence for the theoretical claims.

Supporting Information

All Supplementary Materials can be accessed as a HTML or R Markdown file at: <https://doi.org/10.6084/m9.figshare.4888391.v2>

Acknowledgements P.M. was supported by the Research Fund Flanders (FWO Vlaanderen) through a doctoral fellowship. GH is supported by the German Research Foundation (grant HE 6244/1-2).

Author contributions P.M. and G.H. designed research; P.M. analyzed data; P.M. and G.H. wrote the paper.

References

- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. doi:10.1016/j.jml.2007.12.005
- Bachmann, T., Breitmeyer, B., & Ogmen, H. (2007). *The experimental phenomena of consciousness: A brief dictionary*. Oxford: Oxford University Press.
- Barbot, A., & Kouider, S. (2011). Longer is not better: Nonconscious overstimulation reverses priming influences under interocular suppression. *Attention, Perception & Psychophysics*, *74*, 174–184. doi:10.3758/s13414-011-0226-3
- Breitmeyer, B. G. (2015). Psychophysical “blinding” methods reveal a functional hierarchy of unconscious visual processing. *Consciousness and Cognition*, *35*, 234–250. doi:10.1016/j.concog.2015.01.012
- Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*(4), 335–359. doi:10.1016/S0022-5371(73)80014-3
- Core Team, R. (2014). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Retrieved from <http://www.R-project.org/>
- Dehaene, S., Molko, N., Cohen, L., & Wilson, A. J. (2004). Arithmetic and the brain. *Current Opinion in Neurobiology*, *14*(2), 218–224. doi:10.1016/j.conb.2004.03.008
- Dehaene, S., Charles, L., King, J.-R., & Marti, S. (2014). Toward a computational theory of conscious processing. *Current Opinion in Neurobiology*, *25*, 76–84. doi:10.1016/j.conb.2013.12.005
- Doyen, S., Klein, O., Simons, D. J., & Cleeremans, A. (2014). On the other side of the mirror: Priming in cognitive and social psychology. *Social Cognition*, *32*(Supplement), 12–32. doi:10.1521/soco.2014.32.supp.12
- Etz, A., & Vandekerckhove, J. (2016). A Bayesian perspective on the reproducibility project: Psychology. *PLoS ONE*, *11*(2), e0149794. doi:10.1371/journal.pone.0149794
- Fogelson, S. V., Kohler, P. J., Miller, K. J., Granger, R., & Tse, P. U. (2014). Unconscious neural processing differs with method used to render stimuli invisible. *Frontiers in Psychology*, *5*, 601. doi:10.3389/fpsyg.2014.00601
- Franz, V. H., & von Luxburg, U. (2014). Unconscious lie detection as an example of a widespread fallacy in the neurosciences. *arXiv:1407.4240*. Retrieved from <http://arxiv.org/abs/1407.4240>
- Franz, V. H., & von Luxburg, U. (2015). No evidence for unconscious lie detection: A significant difference does not imply accurate classification. *Psychological Science*. doi:10.1177/0956797615597333
- Gelman, A., & Stern, H. (2006). The difference between “significant” and “not significant” is not itself statistically significant. *The American Statistician*, *60*(4), 328–331. doi:10.1198/000313006X152649
- Goodman, S. N., Fanelli, D., & Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Science Translational Medicine*, *8*(341). doi:10.1126/scitranslmed.aaf5027
- Hassin, R. R. (2013). Yes it can: On the functional abilities of the human unconscious. *Perspectives on Psychological Science*, *8*(2), 195–207. doi:10.1177/1745691612460684
- Hedger, N., Adams, W. J., & Gamer, M. (2015). Fearful faces have a sensory advantage in the competition for awareness. *Journal of Experimental Psychology: Human Perception and Performance*. doi:10.1037/xhp0000127
- Hesselmann, G., & Knops, A. (2014). No conclusive evidence for numerical priming under interocular suppression. *Psychological Science*. doi:10.1177/0956797614548876
- Hesselmann, G., & Moors, P. (2015). Definitely maybe: Can unconscious processes perform the same functions as conscious processes?. *Frontiers in Psychology*, *6*(584). doi:10.3389/fpsyg.2015.00584
- Hesselmann, G., Darcy, N., Sterzer, P., & Knops, A. (2015). Exploring the boundary conditions of unconscious numerical priming effects with continuous flash suppression. *Consciousness and Cognition*, *31*, 60–72. doi:10.1016/j.concog.2014.10.009
- Ioannidis, J. P. A., Allison, D. B., Ball, C. A., Coulibaly, I., Cui, X., Culhane, A. C., & van Noort, V. (2009). Repeatability of published microarray gene expression analyses. *Nature Genetics*, *41*(2), 149–155. doi:10.1038/ng.295
- Iqbal, S. A., Wallach, J. D., Khoury, M. J., Schully, S. D., & Ioannidis, J. P. A. (2016). Reproducible research practices and transparency across the biomedical literature. *PLoS Biology*, *14*(1), e1002333. doi:10.1371/journal.pbio.1002333
- Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*(1), 54–69. doi:10.1037/a0028347
- Karpinski, A., Yale, M., & Briggs, J. C. (2016). Unconscious arithmetic processing: A direct replication. *European Journal of Social Psychology*. doi:10.1002/ejsp.2175
- Kim, C.-Y., & Blake, R. (2005). Psychophysical magic: Rendering the visible “invisible.”. *Trends in Cognitive Sciences*, *9*(8), 381–388. doi:10.1016/j.tics.2005.06.012
- Knops, A. (2016). Probing the neural correlates of number processing. *The Neuroscientist*. doi:10.1177/1073858416650153
- Koch, C., Massimini, M., Boly, M., & Tononi, G. (2016). Neural correlates of consciousness: Progress and problems. *Nature Reviews Neuroscience*, *17*(5), 307–321. doi:10.1038/nrn.2016.22
- Leek, J. T., & Jager, L. R. (2016). Is most published research really false? *bioRxiv*, 50575. doi:10.1101/050575
- Ludwig, K., & Hesselmann, G. (2015). Weighing the evidence for a dorsal processing bias under continuous flash suppression. *Consciousness and Cognition*, *35*, 251–259. doi:10.1016/j.concog.2014.12.010
- Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*(2), 227–234.
- Moors, P., Boelens, D., van Overwalle, J., & Wagemans, J. (2016). Scene integration without awareness: No conclusive evidence for processing scene congruency during continuous flash suppression. *Psychological Science*, *27*(7), 945–956. doi:10.1177/0956797616642525
- Moors, P., Wagemans, J., & de Wit, L. (2016). Faces in commonly experienced configurations enter awareness faster due to their curvature relative to fixation. *PeerJ*, *4*. doi:10.7717/peerj.1565
- Morey, R. D., Rouder, J. N., Love, J., & Marwick, B. (2015). *BayesFactor: 0.9.12-2 CRAN* [Zenodo]. Retrieved from doi:10.5281/zenodo.31202
- Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, *72*, 6–18. doi:10.1016/j.jmp.2015.11.001
- Naccache, L. (2009). Priming. In T. Bayne, A. Cleeremans, & P. Wilken (Eds.), *The Oxford companion to consciousness* (pp. 533–536). Oxford: Oxford University Press.
- 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. doi:10.1038/nn.2886
- Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L., Epskamp, S., & Wicherts, J. M. (2015). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*. doi:10.3758/s13428-015-0664-2

- Phillips, N. (2016). *yarr: A companion to the e-book YaRrr!: The Pirate's Guide to R. R package version 0.0.5*. Retrieved from www.thepiratesguidetor.com
- Reynvoet, B., Brysbaert, M., & Fias, W. (2002). Semantic priming in number naming. *The Quarterly Journal of Experimental Psychology*, 55(4), 1127–1139. doi:10.1080/02724980244000116
- Ric, F., & Muller, D. (2012). Unconscious addition: When we unconsciously initiate and follow arithmetic rules. *Journal of Experimental Psychology: General*, 141(2), 222–226. doi:10.1037/a0024608
- Roggeman, C., Verguts, T., & Fias, W. (2007). Priming reveals differential coding of symbolic and non-symbolic quantities. *Cognition*, 105(2), 380–394. doi:10.1016/j.cognition.2006.10.004
- Rouder, J. N., Engelhardt, C. R., McCabe, S., & Morey, R. D. (2016). Model comparison in ANOVA. *Psychonomic Bulletin & Review*, 23(6), 1779–1786. doi:10.3758/s13423-016-1026-5
- RStudio Team. (2015). *RStudio: Integrated development environment for R (Version 0.99.441)*. Boston: RStudio, Inc.
- Shanks, D. R. (2016). Regressive research: The pitfalls of post hoc data selection in the study of unconscious mental processes. *Psychonomic Bulletin & Review*, in press.
- Singmann, H., Bolker, B., Westfall, J., & Aust, F. (2016). *afex: Analysis of Factorial Experiments. R package version 0.16-1*. Retrieved from <http://CRAN.R-project.org/package=afex>
- Sklar, A. Y., Levy, N., Goldstein, A., Mandel, R., Maril, A., & Hassin, R. R. (2012). Reading and doing arithmetic nonconsciously. *Proceedings of the National Academy of Sciences*, 109(48), 19614–19619. doi:10.1073/pnas.1211645109
- Soto, D., & Silvanto, J. (2014). Reappraising the relationship between working memory and conscious awareness. *Trends in Cognitive Sciences*, 18(10), 520–525. doi:10.1016/j.tics.2014.06.005
- Stein, T., & Sterzer, P. (2014). Unconscious processing under interocular suppression: Getting the right measure. *Frontiers in Psychology*, 5, 387. doi:10.3389/fpsyg.2014.00387
- Stein, T., Kaiser, D., & Peelen, M. V. (2015). Interobject grouping facilitates visual awareness. *Journal of Vision*, 15(8), 10. doi:10.1167/15.8.10
- Sterzer, P., Stein, T., Ludwig, K., Rothkirch, M., & Hesselmann, G. (2014). Neural processing of visual information under interocular suppression: A critical review. *Frontiers in Psychology*, 5, 453. doi:10.3389/fpsyg.2014.00453
- Tsuchiya, N., & Koch, C. (2005). Continuous flash suppression reduces negative afterimages. *Nature Neuroscience*, 8(8), 1096–1101. doi:10.1038/nn1500
- Tsuchiya, N., Koch, C., Gilroy, L. A., & Blake, R. (2006). Depth of interocular suppression associated with continuous flash suppression, flash suppression, and binocular rivalry. *Journal of Vision*, 6(10), 1068–1078. doi:10.1167/6.10.6
- Van Opstal, F., Gevers, W., De Moor, W., & Verguts, T. (2008). Dissecting the symbolic distance effect: Comparison and priming effects in numerical and nonnumerical orders. *Psychonomic Bulletin & Review*, 15(2), 419–425.
- Wolsiefer, K., Westfall, J., & Judd, C. M. (2016). Modeling stimulus variation in three common implicit attitude tasks. *Behavior Research Methods*
- Yang, E., Brascamp, J., Kang, M.-S., & Blake, R. (2014). On the use of continuous flash suppression for the study of visual processing outside of awareness. *Frontiers in Psychology*, 5, 724. doi:10.3389/fpsyg.2014.00724
- Yuval-Greenberg, S., & Heeger, D. J. (2013). Continuous flash suppression modulates cortical activity in early visual cortex. *The Journal of Neuroscience*, 33(23), 9635–9643. doi:10.1523/JNEUROSCI.4612-12.2013