The pervasive problem of *post hoc* data selection in studies on unconscious processing – A reply to Sklar, Goldstein, & Hassin (2021)

Marcus Rothkirch[1], David R. Shanks[2+], Guido Hesselmann[3+]

1 Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Department of Psychiatry and Neurosciences, Visual Perception Laboratory, Germany

2 Division of Psychology and Language Sciences, University College London, England

3 Department of General and Biological Psychology, Psychologische Hochschule Berlin, Germany

+ = equal contribution

Acknowledgments

We thank Simone Malejka, Liad Mudrik, Ariel Sklar, and Miguel Vadillo for helpful comments. DS was supported by grant ES/S014616/1 from the United Kingdom Economic and Social Research Council. Code for all of the analyses and simulation figures reported in this article is available at https://osf.io/wxjhb/.

Abstract

Studies on unconscious mental processes typically require that participants are unaware of some information (e.g., a visual stimulus). An important methodological question in this field of research is how to deal with data from participants who become aware of the critical stimulus, according to some measure of awareness. While it has previously been argued that the *post hoc* selection of participants dependent on an awareness measure may often result in regression-to-the-mean artifacts (Shanks, 2017), a recent article (Sklar, Goldstein, & Hassin, 2021) challenged this conclusion claiming that the consideration of this statistical artifact might lead to unjustified rejections of true unconscious influences. In this reply, we explain this pervasive statistical problem with a basic and concrete example, show that Sklar et al. fundamentally mis-characterize it, and then refute the argument that the influence of the artifact has previously been overestimated. We conclude that, without safeguards, the method of *post hoc* data selection should never be employed in studies on unconscious processing.

Studies on unconscious mental processes, such as subliminal priming, typically require that participants are unaware of some information (e.g., a visual stimulus). What if, however, more than half of the participants actually notice and correctly report the critical stimulus that was supposed to remain invisible due to visual masking, and whose effects on subsequent behaviour were to be examined?

In their experiments on "unconscious arithmetic" (i.e., the presumed ability to solve simple equations without being aware of them), Sklar, Levy, Goldstein, Mandel, Maril, and Hassin (2012) decided to simply restrict the collected sample to those participants who did not see the critical priming stimulus, based on specified visibility criteria. While excluding more than half of the participants is an extreme case, the approach of selecting data *post hoc* is pervasive in studies on unconscious processing, most likely due to its compelling simplicity (albeit it does seem rather wasteful to use only half of the collected data). It has been argued before that at the heart of this approach is a sampling fallacy: Discarding some observations from the sample does not change the properties of the underlying population on which the sample is based (Schmidt et al., 2011). Another point was raised by one of the authors of this comment. Shanks (2017) showed that *post hoc* data selection can lead to false positives – namely inferring the existence of unconscious influences on behaviour when they do not truly exist – due to regression to the mean. What is more, data simulations suggested that the "unconscious arithmetic" effect reported by Sklar and colleagues (2012) is indeed compatible with the potential effect of regression to the mean.

In a recent article published in *Experimental Psychology*, Sklar, Goldstein, and Hassin (2021) provide a critical review of the analyses by Shanks (2017). To put it briefly, they argue that the dangers posed by *post hoc* selection have been exaggerated. Here, we counter their arguments, because, as we will show, the pervasive problem of *post hoc*

data selection should not be underestimated in studies on unconscious processing. Quite the contrary, we argue that without safeguards, this method should never be used, and conclude that it would be better to employ experimental manipulations to ensure that the entire group of participants is unaware of the critical stimulus, rather than selecting data *post hoc.*

Sklar et al. (2021) claim that accounting for regression to the mean artifacts would result in unjustified rejections of true unconscious influences (that is, false negatives). Importantly, this discussion revolves around the practical question of the extent to which the exclusion of trials or participants showing some degree of stimulus awareness can engender systematic biases and should thus be avoided. They caution (Sklar et al., 2021, p. 131) that the concerns voiced by Shanks (2017) "may hinder future research on consciousness and nonconscious processes". In the following, we will show that the conclusions drawn by Sklar et al. are based on several faulty lines of reasoning, including a fundamental misunderstanding of the regression problem, and that the exclusion of trials or participants can indeed lead to wrongly inferring the existence of unconscious influences.

What is the regression-to-the-mean problem?

In our view Sklar et al. (2021) do not provide a satisfactory description of the basic regression-to-the-mean problem in research on unconscious mental processes; indeed, as we argue below, their presentation includes a number of critical errors and misunderstandings. We therefore begin by explaining the problem with a concrete example. The aim of this example is two-fold: First, to illustrate, for readers not fully familiar with the problem, what it consists of in a situation stripped to its essentials; and secondly, to provide a context in which Sklar et al.'s criticisms can be evaluated.

Suppose that we acquire a mean measure of performance (henceforth *P*) on some implicit or unconscious measure such as priming, together with a measure of awareness (henceforth *A*) of the feature or regularity that underlies performance from each experimental participant. For instance *A* could be accuracy on a 2-alternative forced-choice (2AFC) measure to assess which of 2 prime stimuli was presented on each trial. The regression-to-the-mean problem relates to the awareness measure *A* in those circumstances where the researcher chooses to only analyze data from the subsample of participants whose awareness is below some cut-off (*post hoc* data selection[1]). Imagine that, on average, *A* is greater than chance: in the following simulation, the mean value of *A* is 0.6 against a chance level of 0.5, as would be appropriate for a 2AFC measure. Of course, the true value of *A* varies across participants, modelled here by assuming that *A* is sampled from a normal distribution with *M* = 0.6 and *SD* = 0.1. For each participant, the true value of *A* is measured with error *e*, modelled in this example as coming from a distribution with *M* = 0 and *SD* = 0.2 (so s_e^2 = 0.04 and the variance of measured awareness is s_A^2 = 0.05).

The left panel of Figure 1 shows data from 50 simulated participants (each depicted as a pair of connected points). For each participant, the blue triangle shows their true awareness and the orange circle their observed (i.e., measured) awareness; the difference between these is the error, shown in blue when the error is in the positive direction and orange when it is in the negative direction. The data are dispersed in a way that looks random, with about half the errors being positive and half negative, and randomly related to the smaller and larger true values, as must be the case given that the true values and errors are sampled independently. *Post hoc* selection would retain

[1] Note that "*post hoc*" refers to the fact that data selection begins when the data have been collected. Thus, even if selection rules were defined *a priori* (i.e., before data collection) and/or were preregistered, later data selection is *post hoc*.

for analysis all of the 18 participants whose measured *A* is to the left of the dashed vertical line which marks chance performance.

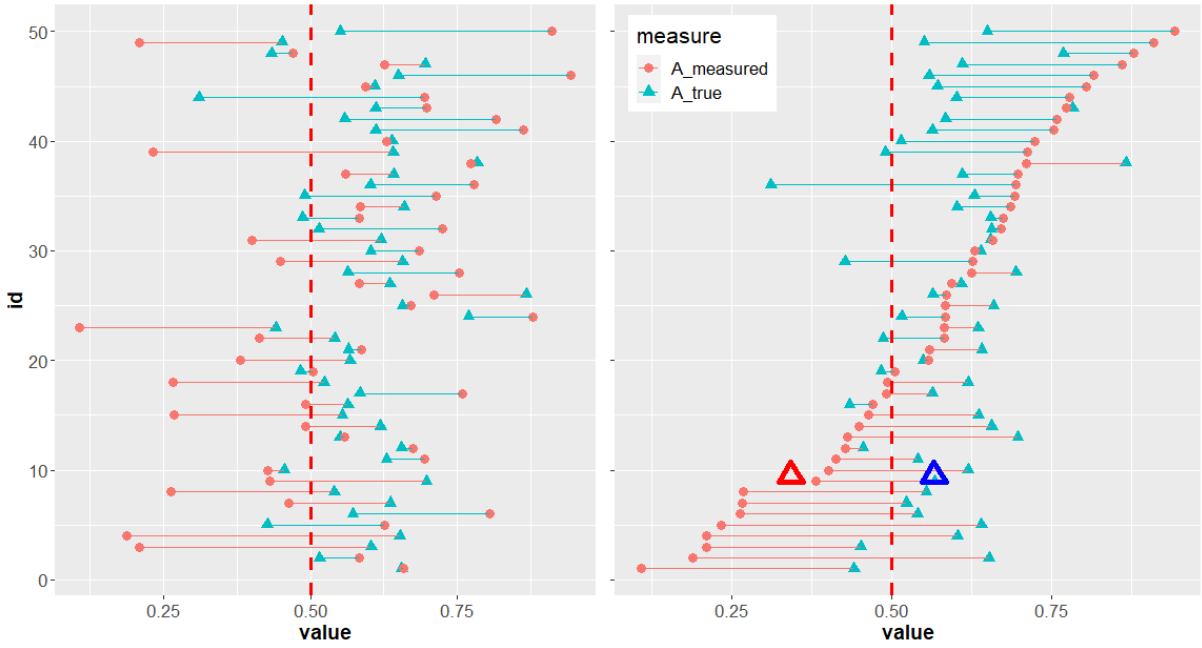

Figure 1. Left panel: Simulated data from 50 participants, each shown as a pair of connected points. The triangles (blue in the online version) mark true awareness and the circles (orange in the online version) are measured awareness once error has been added to the true values. Connecting lines indicate whether the added error is positive (blue in the online version) or negative (orange in the online version). Right panel: The same data, but ordered from largest to smallest by measured awareness. The large triangle at the lower left (red in the online version) marks the mean level of measured awareness in the subsample of participants whose measured awareness is less than chance (< 0.5), and the large triangle further right (blue in the online version) is the mean true level of awareness in these participants. The difference between these means is the regression bias: Measured awareness in participants selected *post hoc* underestimates their true awareness.

The right panel depicts exactly the same data, but now ordered from largest to smallest in terms of measured awareness. This figure makes it easy to see two things. First, the dispersion of true *A* (blue triangles) is much narrower than the dispersion of measured *A* (orange circles). While true *A* mostly ranges from about 0.4 to 0.8, the measured values range from 0.1 to 1.0. This is unsurprising: measurement error increases the range of values. The second feature is more interesting: it is that the errors tend to be

positive for the largest measured values (most of the lines at the top of the panel are blue) and negative for the smallest measured values (most of the lines at the bottom of the panel are orange). If we now look at participants selected *post hoc* for analysis, we see that for all but one of them (in this particular sample), the awareness measurement error is negative. Thus measured awareness underestimates true awareness. Indeed, as a subsample, these participants turn out to be aware, not unaware (14/18 have $A > 0.5$). While their mean measured A is below chance (red triangle), the true awareness of this sub-group is greater than chance (blue triangle). True awareness regresses towards the group mean (0.6 in this example): if we were to measure awareness a second time in this sub-group with a new set of independent error terms *now averaging zero*, their awareness would be above chance, not below. This is the essence of the problem: *post hoc* selection leads to underestimation of true awareness, an effect we henceforth refer to as regression bias.

This discussion has ignored the performance measure P. Let us suppose, for the sake of argument, that most or indeed all participants show some above-chance level of performance. The researcher carrying out *post hoc* selection and paying no heed to the regression-to-the-mean problem will conclude that P is above chance in participants lacking awareness. But this inference would be invalid, because, in reality, these participants do not lack awareness.

Shanks (2017) provided an explanation for why this regression bias occurs. It seems intuitive to believe that random error is equally distributed across the range of a measured variable. However this is false – at the positive extreme of any measured variable, errors will on average be positive and at the negative extreme they will be negative. Random error is equally distributed across the range of *true* values, but not across the range of *measured* values, because the latter include the error. Think of it

this way: if error is positive, then it is more likely that the resulting measured value will be large (because error is one of the components contributing to the observed value), whereas if error is negative, then it is more likely that the resulting measured value will be small.

It should be obvious that the extent of this regression bias depends critically on the magnitude of the error in our measurement of *A*. If error is negligible, then there will be almost no regression to the mean, as measured and true *A* in the subsample will be almost identical[2]. It is only when error is larger that it causes regression to the mean to an extent that might compromise any conclusions the researcher wishes to draw about unconscious mental processes. Indeed the relationship between the misestimation of *A* (regression bias) and the reliability of *A* (r_{AA}) is quantifiable in a simple equation (Campbell & Kenny, 1999, pp. 27, 50)[3]: the true subsample mean is equal to A_{group} - $r_{AA}(A_{group} - A_{subsample})$, where A_{group} and $A_{subsample}$ are the mean levels of measured awareness in the group and subsample, respectively. This formula has some simple graphical interpretations, as shown in Figure 2. If we denote the distance between measured *A* in the subgroup and in the complete group by *a*, and the distance between the true subgroup mean and the overall group mean by *b*, then: (1) the reliability of *A*, r_{AA}, is equal to *a*/*b*; hence the distance between the true subsample mean and the group mean (*a*), expressed as a proportion of the distance between the measured subsample mean and the group mean (*b*), is equal to r_{AA}; and (2) the distance the subsample mean needs to move upwards (*b* – *a*), relative to the maximum amount it could move (*b*), is equal to 1 - r_{AA}, which is also equal to the ratio of the variances of the error *e* and measured *A*, s_e^2/s_A^2. So if r_{AA} is 1, the estimate does not need to move

[2] Readers are invited to confirm this for themselves by running the relevant R code (available at https://osf.io/wxjhb/) with the *SD* of the error term set to 0.
[3] We thank A. Sklar for pointing out the relevance of this formula. We have validated it in simulated data of the sort shown in Figure 1.

at all, if it is 0 the estimate needs to move the whole way, and if it is 0.5 the estimate needs to move half the distance. We return to this quantification of expected bias below when we review the likely range of reliabilities in standard tests of awareness.

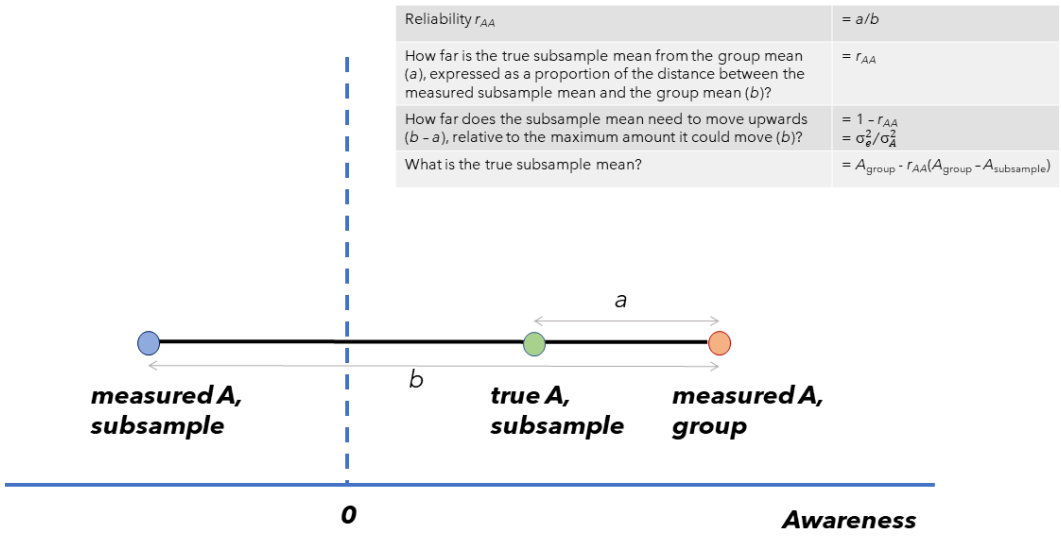

Figure 2. Graphical representation of formulae for quantifying regression bias. The chance level of awareness *A* is assumed to be zero.

As we now explain, Sklar et al. (2021) misrepresent the regression problem in two important ways: first, they state that it leads to overestimation of *P* in unaware participants (it does not); and secondly, they assert that the fundamental factor constraining the magnitude of the regression bias is the *A*-*P* correlation, when in reality it is the reliability of *A* that matters.

Sklar et al.'s misrepresentation of the regression problem

Sklar et al. (2021) repeatedly argue that the regression problem identified by Shanks (2017) and illustrated in Figure 1 is unlikely to be a major concern in reality. However

their entire argument is based on a mis-characterization of regression bias. For example, they say “In a recent article Shanks (2017) argued that this strategy, which he refers to as “post hoc data selection” *introduces an inflation of apparent nonconscious performance* […] We concur with the novel and important assertion made in Shanks (2017), that utilizing a priori selection rules can *lead to potential inflation of observed performance* due to regression to the mean” (Sklar et al., 2021, p. 131); “when there is no true correlation between awareness and performance, *there is no inflation of the nonconscious effect*” (p. 131); “Similarly, nonconscious *performance can only be underestimated* due to regression to the mean if the true correlation between awareness and performance is negative” (p. 132); “To conclude, when awareness measures are reliable and yet they do not correlate with performance, *inflation of the observed nonconscious performance* due to regression to the mean is not a major concern, and should not seriously impact our interpretation of the results” (p. 133); “Shanks (2017) also presented several additional tools […] However, none of these tools allow one to determine whether regression to the mean actually *inflates the observed unaware performance*” (p. 133, footnote 5) (emphasis added in these quotations).

These quotations indicate a profound misunderstanding. The regression bias described in the preceding section does not lead to overestimation (or indeed underestimation) of *P*: it causes no bias at all in the measurement of *P*. It causes bias in the measurement of *A* because, as explained above, the *post hoc* method selects on the basis of *A* and then uses the selected *A* values to compute mean awareness in the selected sub-group. This double-dipping (using the same measure twice) introduces bias because measurement error is not randomly sampled amongst the selected sub-group. But measurement error in *P* is not biased in the selected sub-

sample because measured *P* is not used as the basis for selection, and measurement errors in *P* and *A* are assumed to be independent. It is unclear what the origin of this crucial misunderstanding is: nowhere did Shanks (2017) state or imply that the regression bias leads to an inflation in measured performance[4].

To be clear, there is no bias in the measurement of *P* in participants selected *post hoc* on the basis of their measured awareness. This can be verified by simulating the single-process model that Shanks (2017) applied to Sklar et al.'s (2012) Experiment 6 and that was re-employed by Sklar et al. (2021, Appendix, Dataset 1). Code for this model available at https://osf.io/wxjhb/ shows that, across a range from about -50 to 100 msec, both the measured and true *P* in 'unaware' participants (*A* < 0 in this case) is approximately 2.5 msec[5].

Even when they provide a more accurate description of the bias, Sklar et al. (2021) display further misunderstanding. They state (p. 131): "[…] scientists focus on analyzing the data of participants who score low on awareness measures, that is, they analyze the data of participants who are categorized as unaware. *By definition*, the mean awareness score of this subsample is lower than that of the entire sample. As such, regression to the mean implies that these participants' true awareness scores should be, on average, higher than their observed scores" (emphasis added). But there is nothing inevitable ('*by definition*') about regression, it only occurs if there is random error in the measurement of awareness and if sampling error is negligible.

[4] If one selects a subgroup *measured* as unaware and uses (unbiased) *P* in that subgroup to estimate *P* in *truly* unaware participants, then there will be bias in this estimation. When the *A*-*P* correlation is positive the bias will lead to an over-estimation of *P* and when it is negative the bias will lead to under-estimation. However this cannot be the sense intended by Sklar et al. (2021), as indicated by their references to "observed unaware performance" in the above quotations. More importantly, the level of *P* in truly unaware participants has little relevance because – outside of simulations – we have no way of knowing which participants are truly unaware.

[5] Note that this value is lower than the one given by Shanks (2017), p. 771. The reason is that here we apply a strict awareness cut-off (*A* < 0) whereas, following Sklar et al.'s analytic approach, Shanks (2017) used a cut-off of *A* < 0.1.

In the final section of their article Sklar et al. (2021) propose an alternative method for dealing with regression bias but this method aims to solve a non-existent problem (inflation in measured performance). They say (p. 134, emphasis in original): “Having concluded that the procedure proposed by Shanks (2017) to estimate the influence of regression to the mean is not satisfactory, how *should* we estimate this influence? Briefly, our suggestion is that we rely on the multiplication we discussed in Section 1: The influence of regression to the mean on the observed performance in a given unaware group is the amount of true awareness in that unaware group multiplied by the true effect of awareness on performance.” There is no influence of regression to the mean on performance, so this method serves no purpose[6].

The awareness-performance correlation

Sklar et al. (2021) discuss at length the importance of the correlation between awareness and performance and argue that the regression bias is unlikely to be a concern when this correlation is close to zero or negative, which they claim is often the case in experiments. But the illustration provided above in Figure 1 should make it clear that the problem relates purely to the measurement of *A*. What matters is how noisy this measurement is – less noise means less regression of *A* towards the group mean and more noise means more regression.

It is true that – inevitably – the more error there is in the measurement of *A*, the less well it will correlate with another measure such as *P*. But it is the reliability of *A* that fundamentally determines the extent of regression bias, not its correlation with *P*.

[6] Equally, Sklar et al.’s example about the relationship between temperature and amount of clothing worn reflects the same misunderstanding of the regression bias problem. The issue is not miscalculation of the amount of clothing worn, it is miscalculation (because of regression bias) of the mean temperature on hot days.

Stated differently, regression bias would equally occur if *A* and *P* are uncorrelated or if they are correlated at the maximal level that their reliabilities permit. Thus there are not two forms of regression to the mean. The regression artifact occurs because error in the measurement of awareness (*A*) is biased as a result of *post hoc* selection. This means that true awareness in "unaware" participants is greater than measured awareness (we might label this the 'univariate' case). It also means that performance (*P*) in the "unaware" participants is closer to the group mean *P* than *A* in these Ss is to overall group *A* (the 'bivariate' case). These are not two distinct phenomena, they are inevitable consequences of the same underlying cause. It is true that Shanks (2017) devoted much attention to the *A*-*P* correlation, but the reason for doing so was to highlight that this correlation determines the extent to which *P* in the sub-group is regressive: it is closer to the group mean level of *P* than *A* is to group mean *A* (so long as the correlation is not 1.0 or -1.0). Being biased and being regressive are not the same thing. For the reasons explained above, measured *A* in participants selected *post hoc* is biased, but measured *P* is not. In contrast, *P* in these participants is regressive.

Their inappropriate focus on the awareness-performance correlation leads Sklar et al. to another incorrect claim about regression bias. Notably, according to their reasoning, this bias requires a positive correlation between awareness and performance. They state that in the case of negative correlations, the selection of unaware participants may even lead to an underestimation of nonconscious effects. Specifically, Sklar et al. (2021) say (p. 132) "Moreover, in cases where awareness is negatively correlated with performance, that is, when aware participants perform worse than unaware participants, the regression artifact will actually yield underestimation of nonconscious effects. […] if the true correlation between awareness and performance is negative or

nonexistent, regression to the mean cannot lead us to conclude that there is more nonconscious performance than there actually is".

This claim is incorrect in two ways: first, the regression bias does not lead to any misestimation of unconscious performance effects under any circumstances, as explained above, and secondly it is mistaken to claim that the regression bias can be safely ignored when the correlation is negative. Figure 3 shows instead that the regression bias can be very strong in such circumstances. Here, the model that Shanks (2017) applied to Sklar et al.'s (2012) Experiment 6 (Sklar et al., 2021, Appendix, Dataset 1) was again employed, but with one small modification: In the equation that generated performance scores, the latent variable *S* was replaced with (2 - *S*) to create a negative awareness-performance correlation of about $r = -.3$[7]. Just as in the example shown in Figure 1, measured awareness in participants selected *post hoc* underestimates their true awareness. The leftmost triangle (red in the online version) is the mean measured *A* and *P* in the 'unaware' group, while the triangle further right (blue in the online version) gives their true *A* and *P*. And again, this example highlights that Sklar et al. are incorrect in suggesting that regression causes misestimation of *P*: the two triangles have identical levels of *P*. But it does cause substantial underestimation of *A*: true awareness is actually above chance. Hence, the researcher would incorrectly infer that participants selected *post hoc* show an unconscious effect when in fact they are not unconscious.

[7] Note that this change means the model is no longer a single-system one, in that *P* is not zero when *A* = 0. Single-system models of the sort developed by Shanks (2017) cannot, of course, generate negative awareness-performance correlations. But this does not affect the main point here, that the regression bias is a problem regardless of the sign and magnitude of the correlation.

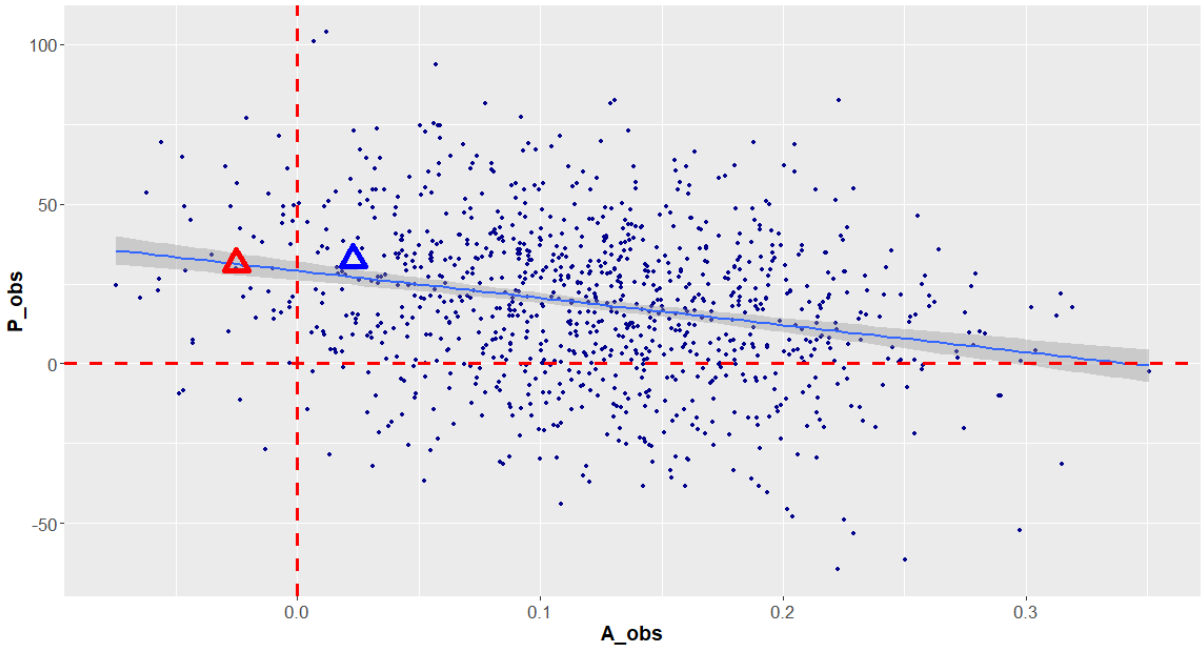

Figure 3. Simulated data from 1000 participants using a modified version of the Shanks (2017) model illustrating regression bias with a negative awareness-performance correlation. The leftmost triangle (red in the online version) marks the mean level of measured awareness in participants whose measured awareness is less than chance (< 0.0 in this case), and the triangle further right (blue in the online version) is the mean true level of awareness in these participants. The difference between these means is the regression artifact.

The reason why regression occurs in this example is the same as in the simulation shown in Figure 1: it is caused by regressive measurement errors in *A*. As such and contrary to the claims put forward by Sklar et al., regression to the mean artifacts in these contexts are not contingent on a positive correlation between participants' awareness and performance. The underestimation of participants' true awareness will also occur when this correlation is negative or absent.

How reliable are measures of awareness?

Crucial to any assessment of the likely importance of regression bias in research on unconscious processing is some knowledge about the reliability r_{AA} of standard measures of awareness. As noted above in Figure 2, the magnitude of the bias is

directly related to reliability. Sklar et al. (2021) report high reliabilities in the Sklar et al. (2012) experiments and argue (pp. 132-3) that “Thus, in many cases, including main examples used in Shanks (2017), there is direct or indirect evidence that awareness measures are reliable.” Although we disagree with Sklar et al. (2021) about the importance of the awareness-performance correlation, we agree with them that in any circumstances in which awareness is reliably measured, regression bias can safely be ignored.

But just how representative are these high reliabilities? To our knowledge, there are surprisingly few reports of the reliabilities of awareness tests in research on unconscious mental processes. Table 1 reports reliabilities (most based on split-half analysis, with Spearman-Brown correction) from an illustrative and non-exhaustive sample of studies, largely from our own research. The figures suggest a picture very different from Sklar et al.’s interpretation: Reliabilities are frequently quite modest, often falling below r_{AA} = .50.

To amplify on one example from contextual cuing, Vadillo et al. (2021) conducted a meta-analysis of reliabilities across four of their own experiments as well as three conducted by Colaguiri and Livesey (2016). In contextual cuing experiments (a very common task for studying implicit learning), awareness is measured by recognition or generation tests. Depending on exactly how the data were coded (as binary or continuous), reliability was around .53 for their own experiments (N = 470) and was even lower at around .30 for Colaguiri and Livesey’s data (N = 913). Even lower reliabilities were obtained in two masking experiments (Hesselmann et al., 2018; Stein et al., 2021), in which stimulus and/or mask contrasts were adjusted with the aim of

ensuring the invisibility of the critical stimulus. At full masking strength, when awareness approaches chance level, reliability approaches 0[8].

Table 1. Reliabilities (r_{AA}) of awareness tests.

	Test type	Reliability estimate
Buchner & Wippich (2000, Experiment 3)	Recognition	.41
Hesselmann et al. (2018, Experiment 6)	Discrimination	.11
Lee & Shanks (2021)*	Recognition confidence	.87
Malejka et al. (2021, Salvador et al. Experiment 1)†	d'	.59
Malejka et al. (2021, Salvador et al. Experiment 2)†	d'	.72
Sklar et al. (2012, Experiment 6)	Discrimination	.93
Sklar et al. (2012, Experiment 7)	Discrimination	.84
Stein et al. (2021)	Discrimination	.18
Vadillo et al. (2021, Experiments 1-4)	Recognition/generation (binary)	.46
Vadillo et al. (2021, Experiments 1-4)	Recognition/generation (continuous)	.59
Vadillo et al. (2021, Colaguiri & Livesey)‡	Recognition (binary)	.25
Vadillo et al. (2021, Colaguiri & Livesey)‡	Recognition (continuous)	.32

* Lee & Shanks (2021) reanalyzed data from Ramey, Yonelinas, and Henderson (2019). † Malejka et al. (2021) reanalyzed data from Salvador et al. (2018). ‡ Vadillo et al. (2021) reanalyzed data from Colaguiri and Livesey (2016).

The point is not to argue that awareness is measured unreliably in much research on unconscious mental processes (although this may be true). Rather, the point is that researchers cannot assume that their measures are reliable. As illustrated in the example in Figure 1, a large amount of error can lead to a truly aware sub-group being mis-classified as unaware. Even if this is an extreme case[9], the researcher risks

[8] Note that this is inevitable. Reliability is the ratio of true score variance to total measured variance. When true awareness is absent, true score variance equals 0, so reliability is 0. Therefore, regression bias is inevitable when awareness is at chance.

[9] The split-half reliability implicit in the model that generated the data in Figure 1 is about .20. The parameters of that model were chosen to make a point about the potential magnitude of regression bias, not to imply that awareness measures will typically be this noisy.

drawing invalid inferences whenever reliability is imperfect and regression bias is ignored. When reliability is 0.5, true awareness is located half way between measured awareness in the subsample and in the complete group.

It must also be borne in mind that even if a measure is reliable, that does not mean that its validity is high. Reliability is necessary but not sufficient: it places an upper bound on validity in the sense that an unreliable measure cannot be a good one, but high reliability does not guarantee that the measure is good. There is a long history of demonstrations that measures of awareness, such as the one used by Sklar et al. (2012), may have poor validity (Shanks & St. John, 1994).

Lastly, there is a logical inconsistency in Sklar et al.'s (2021) defense of the conclusions drawn by Sklar et al. (2012). On the basis of the relatively high reliability of the awareness measures employed by Sklar et al. (which Table 1 suggests are atypical), Sklar et al. (2021) conclude that regression bias is not a major concern in that study. Yet in the model Shanks (2017) constructed to simulate Sklar et al.'s (2012) data, the simulated awareness measure had high reliability (r = 0.76). Thus Sklar et al.'s (2012) data do not require a relatively large amount of error in the awareness measure – and hence a large degree of regression bias – to be explained solely by regression to the mean. In a nutshell, regression bias is a significant concern that can lead researchers to draw incorrect inferences even when awareness is measured with reasonable reliability. In reality its reliability is often quite low, in which case the bias becomes even more worrisome.

How aware were the 'unaware' participants in Sklar et al. (2012)?

Awareness is under-estimated in participants selected *post hoc* – this is the argument made by Shanks (2017) and amplified above. A paradox in Sklar et al.'s (2012) study, which Sklar et al. (2021) defend against Shanks' (2017) critique, is that their selected participants were not demonstrably unaware *even ignoring the effects of regression bias.*

Sklar et al. (2021, p. 133, footnote 8) note that Shanks' (2017) simulation of Sklar et al.'s (2012) Experiment 6 exhibits a mean awareness score in the "unaware" subgroup that is above chance level[10], and conclude that it thus "fails to replicate a crucial aspect of the experimental procedure". Sklar et al. (2021) are correct that for Exp. 6 the level of awareness in the simulated participants retained in the analysis was approximately 55% (chance = 50%), but they do not report the corresponding figure for their experiment. It is 50% with a 95% confidence interval (CI) of [47, 54]. Thus the simulated value is very close to the upper CI of the observed value, but more importantly, the observed data do not rule out an appreciable level of awareness, because the CI is so wide. To conclude that these participants were unaware is to confuse a failure to reject the null hypothesis with proof of the null. Note also that the parameters in Shanks' (2017) simulation were not chosen so as to minimize this small discrepancy.

Even more strikingly, the mean awareness score of the equivalent selected "unaware" subgroup in Sklar et al.'s (2012) Experiment 7 *was* reliably above chance (52%, 95% CI [51, 54]), and thus those results – according to Sklar et al.'s (2021) own standards – should "not be accepted as suggesting nonconscious processing" (p. 133, footnote 8). Note that both Experiments 6 and 7 are supposed to show evidence for

[10] Chance is 50% but Sklar et al. (2012) adopted a cut-off for awareness at 60%, and Shanks' (2017) simulation also adopted this more conservative (from the perspective of defining lack of awareness) cut-off.

"unconscious arithmetic" (Sklar et al., 2012). Even if the required correction for regression bias is small, the case for claiming that the "unaware" participants in Sklar et al.'s (2012) Experiments 6 and 7 were truly unaware is very weak.

Also, note that Sklar et al.'s (2021) implication that awareness in the "unaware" subgroup in Sklar et al.'s (2012) Experiment 6 was truly at chance is paradoxical: Even assuming the existence of an unconscious process in their task, the level of awareness in this subgroup *should* be above chance (i.e., 50%). This is because the cut-off for awareness was not 50% but 60%. Thus in an adequately-powered experiment, the mean level of awareness would be greater than 50%. Failing to reject the null hypothesis (as happened in Sklar et al.'s Experiment 6) is an indication of inadequate power, not a true lack of awareness.

The Shanks (2017) correction method

Sklar et al. (2021) present a simulation intended to show that a method Shanks (2017) described for detecting regression bias is invalid. In this method, bivariate *A*-*P* data are transformed into *z*-space and one then asks whether the observed level of *z(P)* in participants selected for minimal awareness is close to the level predicted from the formula $z(P) = r \times z(A)$, where r is the awareness-performance correlation and $z(A)$ is the mean *z*-standardized awareness score in the selected sub-group. When the actual level of (standardized) performance is significantly different from the level predicted by this formula, an unconscious effect can be inferred.

Like all statistical tests, this test has a non-zero false negative rate: there will be cases where it fails to detect an unconscious influence, such as the example presented by Sklar et al.. As they note, the method does not identify the increasing level of

unconscious processing in Datasets 1, 2, and 3. An initial observation about these datasets, however, is that the regression-to-the-mean problem contaminates all of them: in Datasets 2 and 3, just as in Dataset 1, whatever the level of performance for each data point, the crucial question (for all the reasons given above) is whether the participants performing below chance in awareness ($A < 0$) are actually unaware. The formulae described previously (see Figure 2) allow us to answer this. Because the error in this simulation is relatively large, the reliability of A is appreciably less than perfect ($r_{AA} \approx 0.7$) and hence measured A in simulated participants classified as unaware is biased. In fact, in all 3 Datasets (which are identical for this purpose), true $A \approx 0.02$ ($SD \approx 0.04$). Thus in these datasets simulated participants classified as unaware are in reality (slightly, and in a real experiment with adequate power, would be significantly) aware.

More fundamentally, Sklar et al. (2021) are quite wrong to say (p. 133) that "the statistical procedure proposed by Shanks (2017) appears to be inapplicable to addressing the concern of regression to the mean." Another simulation readily shows this dismissal of the method to be inappropriate. In the left panel of Figure 4 we illustrate simulated data from 1000 participants according to the following model. First, we generate A by sampling from a normal distribution with $M = 1.0$ and $SD = 1.0$. P is sampled independently from another normal distribution with $M = 10.0$ and $SD = 1.0$, except that when A drops towards zero ($A < 0.1$), P is boosted. Specifically, it is sampled from a distribution with $M = 20.0$ and $SD = 1.0$. This models a hypothetical situation in which an unconscious influence is suppressed when awareness is present. Such a conceptualization of the relationship between awareness and performance has

been proposed in many areas including masked priming (Dagenbach et al., 1989), decision making (Dijksterhuis et al., 2006), and lie detection (Ten Brinke et al., 2014)[11].

To apply the method, the data are transformed (right panel) and the formula above applied. As Figure 4 shows, the predicted and observed levels of *P* in the selected sub-group are quite divergent (to an extent that makes a statistical test superfluous). Thus at least under some circumstances the method suggested by Shanks (2017) correctly identifies unconscious influences beyond what would be expected by regression bias.

Figure 4. Simulated data from 1000 participants illustrating the application of Shanks' (2017) method for identifying true unconscious influences above and beyond what is predicted by regression bias alone. Left panel: Each point represents the mean performance (*P*) and awareness (*A*) in an individual participant. Right panel: The same data but transformed into *z*-scores with best-fitting regression line. The upper triangle (green in the online version) marks the mean level of measured awareness and performance in participants whose measured awareness is less than chance (< 0.0). The lower triangle (orange in the online version) is the mean predicted performance in these participants. The difference between these is evidence of an unconscious influence.

[11] For example, ten Brinke et al. (2014, p. 1100) interpreted their lie detection findings as support for a "mental design in which bottom-up accuracy of the unconscious is dampened by the extent to which cognitive resources are available to provide top-down interference".

In a footnote of their article (p. 131, footnote 3), Sklar et al. (2021) allude to the assumption of a linear relationship between awareness and performance (footnote 3). While it is indeed the case that Shanks (2017) demonstrated the consequences of regression to the mean based on linear regression models, postulating more complex relations between awareness and performance does not make regression to the mean artifacts dissolve, as the simulation in Figure 4 highlights. As outlined above, the critical point lies in the systematic selection of participants with negative error terms related to their awareness scores. This systematic bias persists for other types of relationships and is not strictly limited to a linear relation between awareness and performance. Furthermore, there needs to be a valid theoretical foundation for claiming higher-order relationships between these two variables.

Conclusion

In sum, Sklar et al. (2021) portray the *post hoc* selection of data as a valid approach to exclude participants that have shown some degree of awareness of a critical stimulus. Data biases caused by regression to the mean might often not be intuitively obvious. However, one can view such selection processes as a form of an extreme group approach (Preacher et al., 2005), which is especially aggravated in cases where a large portion of participants have been excluded. Shanks (2017) illustrated the consequences of regression to the mean by using the data of a previous article by Sklar et al. (2012). In the “unconscious arithmetic” experiments presented in that article, the data of more than half of the initial sample were discarded, because the awareness measure indicated some degree of awareness in these participants. Even without referring to the mathematical details of regression to the mean artifacts, it should be clear that the exclusion of such a high number of participants is problematic, as the masking procedure did not seem to have worked as intended. The problem here

is that the data were used to draw general conclusions on the scope of unconscious processes, even though they were based on a largely reduced subsample of the original sample. We would thus like to warn against the assumption that the exclusion of the majority of participants might result in a “clean” sample for which unawareness can be safely concluded. Rather, the goal of such experiments should be to exclude as few participants as possible. This can be achieved, for example, by individual or group-based adjustments of stimulus parameters (e.g., stimulus contrast; Hesselmann et al., 2018).

Sklar et al. (2021) misconstrue the influence of regression to the mean, which leads them to incorrect conclusions. The systematic bias imposed by regression to the mean does *not* lie in an over- or underestimation of the performance of the selected participants. Instead, the awareness of the selected subsample is systematically underestimated. Simply put: the selected participants are likely not truly unaware of the critical stimulus. The critical point here is that while in the whole group of participants the random measurement error of the awareness scores should be close to zero, on average, the participants with low awareness scores that have been selected *post hoc* will systematically exhibit negative errors, which thus means that their true awareness is underestimated. Since participants are selected based on their awareness scores, the bias induced by regression to the mean pertains to the estimation of participants’ true awareness, and not the estimation of participants’ true performance.

References

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

Colagiuri, B., & Livesey, E. J. (2016). Contextual cuing as a form of nonconscious learning: Theoretical and empirical analysis in large and very large samples. *Psychonomic Bulletin & Review*, *23*(6), 1996–2009. https://doi.org/10.3758/s13423-016-1063-0

Dagenbach, D., Carr, T. H., & Wilhelmsen, A. (1989). Task-induced strategies and near-threshold priming: Conscious influences on unconscious perception. *Journal of Memory and Language*, *28*(4), 412–443. https://doi.org/10.1016/0749-596X(89)90020-X

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. https://doi.org/10.1126/science.1121629

Hesselmann, G., Darcy, N., Rothkirch, M., & Sterzer, P. (2018). Investigating masked priming along the “vision-for-perception” and “vision-for-action” dimensions of unconscious processing. *Journal of Experimental Psychology. General*, *147*(11), 1641–1659. https://doi.org/10.1037/xge0000420

Lee, D. Y. H., & Shanks, D. R. (2021). *Conscious and unconscious memory and eye movements in context-guided visual search: A comment on Ramey, Yonelinas, and Henderson*.

Malejka, S., Vadillo, M. A., Dienes, Z., & Shanks, D. R. (2021). Correlation analysis to investigate unconscious mental processes: A critical appraisal and mini-

tutorial. *Cognition*, *212*, 104667. https://doi.org/10.1016/j.cognition.2021.104667

Preacher, K. J., Rucker, D. D., MacCallum, R. C., & Nicewander, W. A. (2005). Use of the Extreme Groups Approach: A Critical Reexamination and New Recommendations. *Psychological Methods*, *10*(2), 178–192. https://doi.org/10.1037/1082-989X.10.2.178

Ramey, M. M., Yonelinas, A. P., & Henderson, J. M. (2019). Conscious and unconscious memory differentially impact attention: Eye movements, visual search, and recognition processes. *Cognition*, *185*, 71–82. https://doi.org/10.1016/j.cognition.2019.01.007

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

Schmidt, F., Haberkamp, A., & Schmidt, T. (2011). Dos and don'ts in response priming research. *Advances in Cognitive Psychology*, *7*, 120–131. https://doi.org/10.2478/v10053-008-0092-2

Shanks, D. R. (2017). Regressive research: The pitfalls of post hoc data selection in the study of unconscious mental processes. *Psychonomic Bulletin & Review*, *24*, 752–775.

Sklar, A. Y., Goldstein, A., & Hassin, R. R. (2021). Regression to the mean does not explain away non-conscious processing: Critical review of Shanks (2017). *Experimental Psychology*.

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

Stein, T., Kaiser, D., Fahrenfort, J. J., & Gaal, S. van. (2021). The human visual system differentially represents subjectively and objectively invisible stimuli. *PLOS Biology*, *19*(5), e3001241. https://doi.org/10.1371/journal.pbio.3001241

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

Vadillo, M. A., Malejka, S., Lee, D. Y. H., Dienes, Z., & Shanks, D. R. (2021). Raising awareness about measurement error in research on unconscious mental processes. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-021-01923-y