XLM-2023-2723
Quantifying error in effect size estimates in attention, executive function and implicit learning
Journal of Experimental Psychology: Learning, Memory, and Cognition

Dear Prof Benjamin,

Thank you very much for the opportunity to resubmit our manuscript "Quantifying error in effect size estimates in attention, executive function and implicit learning" to the Journal of Experimental Psychology: Learning, Memory, and Cognition.We sincerely appreciate that yourself and the two reviewers liked our paper and considered it as making a valuable contribution.

We have addressed yours and the reviewers comments below, and we believe the manuscript to be much improved as a result of your collective insightful comments. New additions or amendments to the manuscript are presented in blue font.

We look forward to hearing from you regarding this revision.
Sincerely,

Kelly Garner, Christopher Nolan, Abbey Nydam, Zoie Nott, Howard Bowman & Paul E. Dux

Action Editor
E1.1 To provide practical guidance for future researchers, you need to address the question of number of repeated measures within each subject. Of course the standardized effect size increases with that number even as the unstandardized effect size stays the same. Some researchers will have access to more subjects but less time per subject (as is common with online studies) and others may have access to fewer subjects but more time with each of them (from department subject pools). I would hope that you can find a way to make your advice relevant for both parties and every option in between.

We thank you and R2 for this important point. We addressed this question by determining the extent to which features of inter- and intra-participant variability predicted error in effect size estimates. We show that skewness of inter-subject effects is a key predictor of error, whereas for very small effects, error can be predicted from skewness of intra-subject variability. This allows us to provide advice to researchers considering the trade off between *N* and repeated individual measurements in relation to error in effect size estimates. This also addresses R1's request to provide more substantive advice to researchers and editors (see point *R1.X* below). The resulting changes to the paper are as follows:

To introduce the aims of this analysis, we added the following text to the introduction on pp. 7-8, lines 144-151:
"Last, we identified data features that predict error in effect size estimates, beyond the mean and standard deviation measures of which they are a function. Such features may serve as a flag for whether data from a single experiment may be susceptible to error in effect size

estimates. We focused on the skew and kurtosis of inter- and intra-subject effects, as such measures can bias mean and variance estimates when datasets violate normality assumptions, yet remain undiscussed in simulation studies that assume normality. The results motivate guidelines for study design and interpretation, not only for future AB, MT, SRT and CC studies, but also more broadly for the investigation of cognition."

The methodological detail has been added to the Methods section on pp. 15-16, lines 329-354:

Last, we sought metrics that may inform whether an experiment has yielded an imprecise effect size estimate. Effect sizes are a function of the variability of the effect across individuals, as well as intra-individual variability over trials (Rouder & Haaf, 2018). If either of these stem from a non-normal distribution, mean and standard deviation estimates - and consequently effect size computations - may be impacted. We thus determined whether the skewness and kurtosis of this data could predict error in effect size estimates.

Error in effect sizes were defined for each task as the difference between the expected value for N_{313} and each observed effect size from N_{med}. To attain predictors for each N_{med} simulation, we calculated the key behavioural effect for each participant (in raw units) and computed the Pearson's skewness and kurtosis coefficients of the resulting distribution of effects. We also computed the variability, skew and kurtosis from each participant's performance across trials, and took the means of these measures across participants. The resulting variables (effect skewness, effect kurtosis, mean intra-individual variance, skewness, and kurtosis) served as predictors in a multiple regression analysis, using effect size error as the criterion variable. If any of the regressors themselves showed high levels of skew then a log transformation was applied. All model residuals were checked for homoscedasticity. Note that although we present the full models below, performing stepwise regression yielded the same pattern of results.

To protect against interpreting over-fitted models, we performed k-fold cross-validation for each multiple regression model, where k=10, and we report the mean r_{cv}^2 (and standard deviation) across folds. Next, we determined which regressors consistently predicted effect size error across the four tasks. We then sought to identify which values of such predictors suggest a problematic effect size error (defined as effect size errors that were less or more than the .025 and .975 quantiles for N_{313}). We achieved this using simple regression, as we sought to simulate how much variability may be accounted for when a researcher uses a single piece of information to estimate effect size imprecision.

The results have been added to the Results section on pp. 23-24, lines 528-568:

Last, we determined which aspects of the data were predictive of erroneous effect size estimates. Multiple-regression analysis showed that between ~9-40% of the variance in effect size errors were predicted by effect skewness, effect kurtosis, mean intra-individual variance, mean intra-individual skewness, and mean intra-individual kurtosis (M r_{cv}^2s (SD): AB: 0.39 (0.08), MT main effect of task: 0.09, (0.05), MT task x modality interaction: 0.11 (0.04), SRT: 0.22 (0.09), CC main effect of condition: 0.19 (0.08), all model ps < .001), apart from for the

block x condition effect from the CC task, where the model accounted for only (~1%) of effect size error ($F(5,994) = 2.35$, $p = .04$). This suggests that both inter- and intra-individual skewness and kurtosis predict variability in effect size errors.

The resulting regression equations (see appendix ii) are useful for researchers using the tasks studied here, who wish to predict the extent to which their own experiment may have yielded an imprecise effect size estimate. However, what is more widely useful is understanding which regressors significantly predict effect size imprecision across tasks. We therefore determined which regressors showed significant predictive power across tasks, applying Bonferroni correction for multiple comparisons. For the AB, MT, and SRT tasks, effect skewness and kurtosis were significant predictors of effect size error (all ps <= .005, see appendix ii). Mean intra-individual skew was a significant predictor across all effects (all ps < .008), apart from for the MT task x condition interaction (p=.08).

Having identified the regressors that suggest imprecision in effect size estimates across tasks, we next sought to determine which predictors could be used as a marker of imprecision when a researcher is unable to hold the influence of other predictors constant. Such a finding would suggest that use of a single piece of information (e.g. effect skewness) could act as a marker for whether a single experiment has yielded an imprecise effect size estimate. Simple regressions between each predictor and effect size errors showed that effect skewness tended to predict a higher proportion of the variance (Adjusted R^2s: 0.04 - 0.10, all ps < .001) than kurtosis (Adjusted R^2: -0.00 - 0.01, all ps < .7), apart from for the MT condition x task interaction (skewness: AdjustedR^2: 0.0004, p<.01, kurtosis: Adjusted R^2: 0.06, p<.001). Although mean within-participant skewness predicted higher amounts of error variance for the AB (Adjusted R^2: 0.18) and the CC's main effect of condition (Adjusted R^2: 0.16), its predictive power was poor for the remaining tasks (Adjusted R^2s: <= .02, all ps < .17). This suggests that effect skewness is the best potential general proxy of effect size imprecision, when not controlling for other influences.

As effect skewness is the best candidate for predicting variance in effect size error across tasks, we next determined which values of effect skewness predict problematic levels of effect size error (defined as values falling outside the .025 and .075 quantiles for N_{313}). Across tasks, moderate to large negative effect skewness (-0.70 - -1.29) predicted erroneous over-estimates of effect size, whereas large positive effect skewness (1.35 - 3.80) predicted erroneous under-estimates. Thus, if data from a single experiment shows moderate to large values of effect skewness, this is a signal that extra caution is warranted when interpreting effect size estimates.

We sum up with advice regarding the potential trade off between *N* and repeated measures in the Discussion on pp. 27, lines 624-640:

We also show that for the larger effect sizes studied here (ε_p^2: 0.6-0.9, d ~ 1.9), effect skewness, which is driven by inter-participant variability, shows a predictive relationship with imprecision in effect size estimates. This was not the case for the smallest effects under study, (ε_p^2: 0.02-0.31),

where intra-individual skewness and kurtosis of the data were the significant predictors of imprecision. Thus, researchers wishing to determine the likelihood of an erroneous estimate in their own data should examine different features of the data according to the expected effect size (inter- vs intra-individual skewness). This finding also carries potential consequences for the trade-off between *N* and repeated measures (number of trials) that must be decided for any given study. Specifically, when an effect size is small across participants, intra-individual variability may be the limiting factor for precisely quantifying an effect. This accords with previous observations concerning the reduction of type 2 errors (Rouder & Haaf, 2018). What the current findings show is that decision processes regarding the trade-off between *N* and repeated measures should also consider the number of each required to attain a relatively normal distribution of effects, for either inter- or intra-individual data, depending on whether the anticipated effect size is large or small respectively. Future work should use simulation approaches to verify the causal link between skewness and error in effect size estimates.

E1.2 I found your appendix unconvincing with respect to the question of whether the decreases in variability with increasing sample size reflect only increasing precision or also an artifactual increase in resampling that arises in bootstrapping. I thought the approach was reasonable but I didn't understand why you didn't use a much larger parent population (perhaps even of infinite size). Your choice of parent N = 360 with max sample N = 330 are simply to close in size to draw strong conclusions about the lack of difference between the two cases. Perhaps I've misunderstood the approach with that simulation, in which case I am sure you can correct me.

This is a good point. The sample sizes were chosen to demonstrate how the problem of decreasing variability could manifest in our experiment, in which our base-samples are of size 330/360[1]. As this query suggests, this may, though, mean that our findings are specific to these specific experiments, with the sample sizes we have for them, and, thus, are not a fully general verification of our procedure in its broadest sense.

However, whether by luck or by judgement, our inclusion of the reference condition does actually ensure the full generality of our approach. Indeed, this reference condition might be considered a case in which the base sample/parent population is infinite in size, i.e. all direct samples are completely disjoint. We obviously did not make this aspect of our simulations sufficiently clear. Consequently, we have added the following text to the end of appendix i.

The comparison of *reference* and *indirect_wi_rep* is also of note. The point of interest for this appendix is what happens with direct sampling with replacement (*direct_wi_rep*) as the sample size gets small, and whether there remains a non-trivial probability of the same data points appearing in multiple samples even when the sample size is small. This is because it is the difference of this probability of overlap between direct and indirect sampling that we are interested in. Indeed, the most liberal test of our hypothesis (i.e. giving the greatest opportunity to find a disparity between direct and indirect sampling) is one in which the probability of the same data points appearing in multiple samples with direct sampling is vanishingly small. This is

[1] Lorca-Puls et al., "The Impact of Sample Size on the Reproducibility of Voxel-Based Lesion-Deficit Mappings."

what the *reference* case gives us. With real numbers, there is (mathematically) a probability of zero of sampling the same number multiple times from a Gaussian, so all the samples generated under *reference* are disjoint by construction. (This is, of course, putting aside issues of maximum precision available on a particular computer, which no approach can overcome.) The fact that *reference* shows the same pattern as *direct_wi_rep* in figures 1 and 2, suggests that increasing the size of the *base_sample* in order to reduce sample overlap in *direct_wi_rep* will not change the basic findings, i.e. what we demonstrate here with base samples of 360 is in fact fully general.

E1.3 So, to make the case convincing to me and the reviewers, you really need to take this approach on the road with other samples. These samples should vary in age and other important demographic characteristics. Perhaps you can use existing data for this purpose. Or perhaps you can convince us that demographic individual differences are unimportant for these tasks.

We agree that quantifying whether effect sizes generalise beyond our typically studied populations is a pertinent next step across multiple domains, but we are hesitant that the influence of demographics on effect size variability should fall under the remit of the current project. The aim of our work is to assess the probability of error in effect size estimates given typical sample sizes (and demographics), and we focus on how that probability would change with increasing *N*. There are many steps that need to be taken to improve effect size estimates in psychology, and potentially the first and easiest that could be made is to increase the size of existing samples[2]. Hence, we focus on *N* in our current work.

We explored existing evidence for the impact of demographic individual differences on effect size estimates, to determine whether we should revise our aims. We probed within our own dataset whether demographic characteristics could predict raw effect sizes (computed as the key difference between conditions in either accuracy (AB) or RT (MT, CC, SRT). We assessed the influence of gender (207F), handedness (267 right), age (mean = 20.36, range 18-35) and number of languages spoken (1=191, 2=98, 3=22), using a multivariate regression model. Note that our total sample is *N* = 311 as 1 participant each spoke 4 or 5 languages, whom we excluded. The model accounted for negligible variance in effect sizes (*Adj R Sq:* -0.003, $F(5, 305) = 0.8424$, $p = 0.52$, all *beta* ps > 0.05, uncorrected). We also returned to the existing literature upon which we based our median *N* calculations. Unfortunately, so few of the existing papers provided sufficient information about both demographics and effect size (~10% of studies), we were unable to use the existing literature to assess this relationship. Thus, as our aim is to quantify effect size variability across studies as they are currently run, and given the clear influence of *N* and the scant evidence for an influence of demographic differences, we believe that the influence of demographics on effect size variability should be the subject of a separate, focused investigation, or at least, should not gain precedence within the current work. We do however acknowledge this gap in knowledge in our discussion on pp. 30-31, lines 712-729.

[2] Rouder and Haaf, “Power, Dominance, and Constraint.”

“It remains an open question whether the current findings generalize beyond the paradigms and participant pool used here. There are some suggestions of generalizability of the current observations across tasks that should be investigated in future research. Across all the ϵ_p^2 findings, the standard deviations at N_{313} were small (SDs: .01-.03), and each SD doubled or tripled as a function of moving from N_{313} to N_{med}. Therefore, it is possible that effect sizes such as ϵ_p^2 will show a comparable reduction in variability as *N* increases to the hundreds, across all paradigms. If this were found to be true, then researchers could apply the rates of change observed here to effect size estimates from their own field of study, in order to determine the *N* required to achieve a tolerable level of precision. Moreover, changes in p(hit|*N*) and qq-ratio rates were comparable across *N* for all effects, regardless of size, suggesting invariance to the measurement differences across paradigms. Future research should determine the extent to which these rates were dependent upon the current sample of N_{313}, which was arguably homogeneous with regard to population characteristics. Indeed, it is pertinent to determine the extent to which our results would hold with more heterogeneous samples. For example, estimates of effect sizes may be more variable under less constrained conditions, such as when community samples complete online studies. Future work should determine the extent to which study design choices may hamper precise effect size estimates in such groups.”

Reviewer 1

R1.1. I'm unsure how one specific study is supposed to inform these issues more than, say, a well-done meta-analysis.

We agree with R1 that a meta-analysis may be informative for quantifying effect sizes from the published literature for each task. However, our approach overcomes two limitations of the meta-analytic approach - we can ensure that there is no bias in our dataset, and by using simulations, we are able to offer a view of how the field would be if different *N*s were used as standard.

We have clarified the benefit of our approach in the introduction, on p. 6, lines 100-116):

“Meta-analytic and incomplete sampling approaches for determining an expected effect size are hampered by the quality of the existing literature (Brand, Bradley, Best, & Stoica, 2008; Friston, 2012; Gelman & Carlin, 2014; Lane & Dunlap, 1978; Lorca-Puls et al., 2018). A recent survey of effect sizes across psychology disciplines showed that effects from non-pre-registered studies were much larger than pre-registered studies (r = 0.36 vs 0.16, Schäfer and Schwarz (2019)) suggesting that prior to pre-registration, under-powered studies were contributing inflated effect size estimates to the psychology literature. Although multiple correction methods have been developed within the meta-analytic framework to account for biases due to missing literature (Schmidt & Hunter, 2015), they typically involve assumptions about the sources of missing data, which can never be fully tested. Thus even if one were to define an expected effect size using corrected meta-analyses (if available), there is much to gain from corroborating meta-analytic

results with alternate methods that can guarantee a lack of bias in the available dataset. It is also difficult to determine, on the basis of existing literature - such as when using meta-analysis - how conclusions about effect sizes would differ if a given field of study was different, e.g. how much published literature is likely to be missing if a larger N was used as standard?"

R1.2. The challenge has always been "How do we power our studies for effects for which we have little information or competing theoretical expectations?" And I don't think the present simulations are well-positioned to inform this. But, to be fair, I don't think anyone has an answer that is universally satisfying. I think what would be most appropriate is a series of large-N, registered studies that target a wide-array of problems in the field so that the mean and variance of the effects across studies could be used to inform research design.

This is a good point. We agree that such a series of studies would be of great benefit to the field. We have addressed this point in the Discussion on pp. 28-29, lines 671-676:

"It would be useful to conduct multiple large N studies aimed at characterising effect size distributions across multiple cognitive phenomena. This would not only inform tolerable precision levels, but could also help with theory development. For example, we would better understand the effect magnitude that candidate models should emulate. Further, there would exist more baseline effect magnitudes that could serve as a reference, or upper limit, when hypothesising factors that modulate the effect."

R1.3. Why not treat this as a pure simulation study?

Both pure simulations and those based on existing data are able to make unique contributions to the literature. As we state on lines xxx-xxx of the introduction:

"Typically, simulation studies generate data under some simplifying assumptions about the data generation process (e.g. Albers & Lakens, 2018; Hedges, 1982; Lane & Dunlap, 1978; Troncoso Skidmore & Thompson, 2013; Westfall et al., 2014). Although this work is necessary for informing how effect size estimates behave under varying conditions where ground truth is known, it is challenging to anticipate all the complexities of data from the repeated-measures designs used across a range of phenomena and processes, such as in the study of attention, executive function and implicit learning. Such data are often not normally distributed and carry varying levels of covariance between conditions. Thus, there remains a question mark over the extent to which the results from simulation work generalizes to real-world data. An alternative method is to simulate experimental outcomes by bootstrapping smaller samples from larger, real data-sets (e.g. Lorca-Puls et al., 2018). This approach offers the opportunity to characterize the distributional qualities of effect sizes estimated from high-dimensional data-sets, using varying levels of *N*, while maintaining ecological validity."

Indeed this maintenance of ecological validity allowed us to discover features of the data that predict imprecision in effect size estimates (see *E1.3*); yielding knowledge to form the basis of

new pure simulation studies moving forward, as we state in the discussion on p. 27, lines 635-640:

"What the current findings show is that decision processes regarding the trade-off between *N* and repeated measures should also consider the number of each required to attain a relatively normal distribution of effects, for either inter- or intra-individual data, depending on whether the anticipated effect size is large or small respectively. Future work should use simulation approaches to verify the causal link between skewness and error in effect size estimates."

R1.4. Like the authors, I think there are probably ways to improve theory building by estimating quantitative parameters and folding those back into theoretical models to generate new quantitative predictions. But, in my experience, those parameters are rarely "effect sizes" in the way they are described in this paper and elsewhere (i.e., standardized values). They are typically unstandardized values that are best interpreted in a natural metric (e.g., reaction times, accuracy) … perhaps it would be useful to explain what a study or a program of research might look like if one was using effect size estimates seriously.

It is the case that theories are often motivated in terms of differences in raw units, such as response times or accuracy. To translate this to a prediction of effect magnitude only considers that theorists consider the variability of the effect that they are predicting. We have amended the first paragraph of the introduction to reflect this (lines 48-50), and to further highlight the benefits of such an approach:

"An alternate approach is to develop theory and models that predict the magnitude of the effect. Providing predictions in terms of effect size magnitude prompts theorists to consider the variability as well as the presence of predicted effects, and is demonstrably a useful metric when considering practical relevance (Funder & Ozer, 2019)."

R1.5. I would much rather see some "rules" that exist at the editorial level for what is permissible and what is not. The field does this with alpha thresholds in significance testing. Why not do it for sample sizes and power?

Thank you for this point. In addition to the recommendations made above (see *E1.1* - Discussion section), we have amended the discussion (pp. 25-26, lines 582-606) to contain the following paragraph, which draws broader lessons from the current findings:

"Our findings have practical relevance for study planning. First, we have provided a range of effect sizes that researchers can use to inform power calculations for their own studies. Furthermore, we have shown that in the case of smaller effects (ϵ_p^2 :0.01-0.3), N_{med} was consistently smaller than is required to attain 90% power to reject the null hypothesis. This suggests that researchers should consider whether their research question concerns an effect that may be subtle or variable across participants, and if so, recruit higher *N*s than is currently standard. This would promote maintenance of appropriate type 2 error rates. For the small effects observed here, a minimum *N* of 69 participants was required. Note also that for each

task, the statistical model used was one geared at ascertaining the existence of an effect (e.g. was there an AB present?). These findings suggest that as soon as hypotheses become more nuanced, for example, referring to factors that should modulate the strength of a known effect, effect sizes are likely to be of a smaller range.

The current findings also reveal that sampling a few similar studies to determine a suitable minimum effect size for power analysis is a questionable approach, given the standard N_{med}s. For larger effects, this will lead to an inappropriately powered study 33% of the time, whereas this rate will be 50% for smaller effects. Furthermore, the current inflation bias data suggest that in the case of interactions, (and smaller effects), a comprehensive meta-analysis is likely to yield an inflated estimate when the field uses $<N_{69}$ as standard. Therefore, researchers using existing research to determine appropriate effect sizes for power analyses would be well advised to adjust (decrease) anticipated minimum effect sizes to ensure they avoid an underpowered study. However, given the suggested currently indicated state of the field, the better approach is for researchers to use theoretically motivated minimum effect size estimates, that include consideration for how likely the effect is to vary across individuals, when conducting power calculations."

Reviewer 2

R2.1. I would like to see a revision with more balanced interpretations of the results in the discussion, that allow for more uncertainty in how well they may generalise when providing recommendations.

The original paragraph to which this point referred has been removed from the Discussion. We have also added the following caveats to our discussion of the results:

Regarding the limits of our findings to the effects under study (p. 25, lines 590-594):
"Note also that for each task, the statistical model used was one geared at ascertaining the existence of an effect (e.g. was there an AB present?). These findings suggest that as soon as hypotheses become more nuanced, for example, referring to factors that should modulate the strength of a known effect, effect sizes are likely to be of a smaller range."

Regarding the limits of our findings to the current sample and task materials (p. 28, lines 668-672):
"Here we defined an acceptable level of precision as falling within the .025 and .975 quantiles of the distribution of the best estimate(N_{313}). The usefulness of our definition could potentially be limited to the current sample and task materials. It would be useful to conduct multiple large *N* studies aimed at characterising effect size distributions across multiple cognitive phenomena."

Regarding the limits owing to the demographics of the current sample (pp. 30-31, lines 725-729):

Future research should determine the extent to which these rates were dependent upon the current sample of N_{313}, which was arguably homogeneous with regard to population characteristics. Indeed, it is pertinent to determine the extent to which our results would hold with more heterogeneous samples. For example, estimates of effect sizes may be more variable under less constrained conditions, such as when community samples complete online studies. Future work should determine the extent to which study design choices may hamper precise effect size estimates in such groups.

R2.2. I think it is important to (i) make readers aware of the fact that the recommendations may depend on the number of trials per participant as well, meaning that readers can't necessarily use much fewer trials per participant than the current study, while maintaining the same sample size, and necessarily expect the same power etc., and (ii) discuss how the authors believe the number of trials per participant in the current manuscript may have influenced the results, even if this is a "not sure" type of discussion.

We have addressed this in our response to *E1.3*.