Dear Drs. Dienes and Viding,

Thank you and the reviewers for the thoughtful comments on our manuscript, “The puzzling relationship between multi-lab replications and meta-analyses of the rest of the literature.” Please accept our resubmission. We have addressed your comments and the comments of the reviewers, and we believe that the manuscript is substantially improved. Please find below a point-by-point response to the comments.

Please do not hesitate to contact us if you have any questions or concerns.

Sincerely,

Molly Lewis (on behalf of all authors)

Associate Editor:

> *One theme common to both was the typical lack of heterogeneity within MLRs, which appears to challenge a number of your points. MLRs by design tend to be procedurally uniform and thereby produce pretty uniform effect size estimates. There is procedural and effect size heterogeneity in MAs of course; but how important is this in explaining the MLR MA difference? A lot of MA variability may be theory irrelevant differences like number of questions/trials, changing he population standardized effect size - a point that won't apply in MLRs. I am also reminded of the 2015 reproducibility project; while each study was not a MLR it seems relevant that the effect sizes were about halved in pre-registered replications, also found in the subsequent similar behavioural economics reproducibility project. Can you pick some MLRs that very closely followed a study and still got much smaller results? Would that undermine your points concerning material/procedural heterogeneity being a major explanatory factor?*

Thank you, we think the source of several of the reviewers’ comments may be an error in the original manuscript: the sentence reading “It therefore seems plausible that similar methodological differences could be fairly common in MLRs” should have read “...fairly common in MAs”. Nevertheless, we don’t think that heterogeneity in effect sizes in MAs can ultimately explain the discrepancy, only that it may lead to misleading estimates of the size of the discrepancy when effect size means from the two sources are compared. In the paper, we test this idea empirically by examining where each MLR mean would fall in the distribution of effects in the corresponding meta-analysis, using a method that accounts for statistical error in the point estimates. We found that, in many cases, a sizeable minority of the effects in the meta-analyses were smaller than the corresponding MLR estimate, suggesting that the mean may lead to overestimation of the effect size discrepancy with the corresponding meta-analysis.

> *The reviewers also note the importance of the high correlation you report on p 7 between MA and MLR effect sizes. It may be helpful to report the equation of your blue*

fitted line - it seems there is a slope of about 1 with an offset of about 0.3. I am not sure what that means, but maybe someone will be able to use it!

Thank you for this useful suggestion. The equation is: *MLR_ES* = -0.18 + 0.80 * *MA_ES*. We have added this information to the caption of Figure 1.

Reviewer 1:

Perhaps my most critical comment is related to the structure of the section titled "Publication bias cannot entirely explain the discrepancy". The logic of the opening paragraphs couldn't be clearer: if there is a discrepancy between meta-analysis and MLRs even after correcting for publication bias, this doesn't necessarily mean that meta-analysis / bias correction methods do not work; it could simply mean that the discrepancy is not entirely explained by publication bias. But then the logic of the rest of the section becomes more obscure. The authors conduct an alternative analysis with what is essentially a different type of bias correction method and find that even this analysis yields an (unaccounted for) discrepancy between meta-analytic estimates and MLRs. Perhaps I am missing something, but doesn't this leave is just at the same point as KSJ, only that using a different bias correction method? I missed a clearer explanation of how this analysis goes beyond (and contradict) KSJ.

Thank you for this comment -- we apologize that this was not clearer. The two methods, in fact, lead to different types of inferences. The bias correcting approach presented by KSJ aims to "correct" MA effect size estimates by estimating the degree of publication bias, and shrinking the MA estimates appropriately. The conclusion from the KSJ paper is that there is a large discrepancy between MA and MLR effect sizes due to publication bias, but statistical methods are insufficient to fully adjust for this discrepancy. The take-away is that MAs are not useful because we have inadequate statistical tools for dealing with publication bias. That is, if we had better publication bias correcting methods, then MAs would have value.

Our take-away is somewhat different. We argue that the discrepancy in effect sizes between MAs and MLRs may not be *entirely* due to publication bias (not that there is no publication bias, nor that publication bias is entirely irrelevant to explaining the discrepancy). This is important because it implies that KSJ's inability to "correct" MA effect sizes estimates using publication bias correcting methods may be due to the fact that publication bias is not the only source of the discrepancy, and not that the publication bias correcting statistical methods are entirely inadequate.

We provide evidence for this claim by conducting an analysis that assumes a worst-case model of publication bias. The goal of this analysis is not to estimate the amount of publication bias in a MA, as KSJ do, but rather to assess whether publication bias alone could account for the discrepancy between MA and MLR effect sizes. The conclusion from this analysis is that no amount of correction due to publication bias would resolve the discrepancy. It is therefore not the case, as KSJ argue, that the discrepancy they report between MA and MLR effect sizes

after publication bias correction is due to insufficient statistical methods for correcting for bias; rather, there is some additional cause of the effect size discrepancy from the two sources.

We have added some clarification of how our analysis goes beyond that of KSJ in the main text: “This suggests that KSJ's failure to eliminate the discrepancy in effect sizes derived from the two methods was not due entirely to limitations of publication bias correcting statistical methods; rather, there are additional causes of the discrepancy.”

> *Also, if the sensitivity analysis excludes all affirmatory studies, doesn’t this necessarily bias the point estimate downwards? (Pretty much for the same reason that selecting only significant results biases it upwards.)*

Yes, excluding all affirmatory studies does necessarily bias the point estimate downwards and makes the resulting estimating highly conservative. The purpose of the sensitivity analysis is not to provide an unbiased estimate but to provide a conservative worst-case estimate that invokes fewer statistical assumptions that could have compromised the performance of the methods KSJ applied. The key take away from this analysis is that, even given this *highly conservative* MA estimate, there is still a discrepancy between MLR and MA estimates. This finding provides evidence that publication bias is not the only source of the discrepancy between estimates derived from the two sources.

> *I also lost track of the logic behind the narrative on pages 5-7. The section begins saying that “Many of the meta-analyses in KSJ’s study showed considerable effect heterogeneity.” It is indeed quite intuitive that meta-analysis should contain more heterogeneity than MLRs, because the later are typically based on standardized protocols while the former include disparate procedures and samples. But somehow over the next few lines the argument changes to heterogeneity in the MLRs themselves (e.g., “it therefore seems plausible that similar methodological differences could be fairly common in MLRs”). And then at the end of the section the focus seems to fall again in heterogeneity in meta-analysis. The analysis reported at the end of the section compares the point-estimates of MLR (therefore ignoring heterogeneity in these) with the distribution of effect sizes in meta-analysis (preserving information about heterogeneity). So, it is entirely clear whether the point of this section is that meta-analyses are quite heterogeneous, or MLRs are heterogeneous or both.*

We apologize for this confusion -- there was in fact an error in the original manuscript. The sentence reading “It therefore seems plausible that similar methodological differences could be fairly common in MLRs” should have read “...fairly common in MAs”.

> *I really like the analysis of the distribution of effect sizes in meta-analyses. If I understand everything correctly the discrepancy between these distributions and the results of MLRs would look even smaller if the authors had not corrected the study-level estimates (shrinking effect sizes towards the mean). This is something worth noting, as, in the absence of this explanation, readers unfamiliar with these procedures might*

suspect that this (sophisticated) correction was introduced because the results were thus more consistent with the conclusion of the authors. But in truth it is the other way around: if the authors had naively used the observed distribution of effect sizes, then meta-analytic estimates and MLR results would have looked even more consistent.

Thank you. We have added additional explanation in the manuscript (pg. 8): "(This shrinkage is necessary because the distribution of the point estimates themselves has variability due to both statistical error and heterogeneity, so cannot be directly used to characterize heterogeneity in the population effects.)"

Incidentally, I couldn't find the figure where this is reported. There seem to be two Figure 2's mentioned in the ms (or perhaps two panels) and one is missing.

Apologies, the reference "Fig. 2 right side" referred to the text values on the right side of Figure 2 (not a different panel). We have clarified this in the Main text, and edited the figure caption to make this clearer.

Reviewer 2:

The authors perform novel analyses suggesting that publication bias alone cannot reconcile the difference in effect size estimates between meta-analyses and MLRs, and propose other factors that may play a role. To me, it remains highly likely that QRPs and publication bias are the most important factors, and the contribution of other factors is likely to be comparatively small. However, this is an open question and the authors are doing a service by imploring us to further investigate the reasons for this discrepancy.

Additionally, the authors make a strong case that meta-analysis can still be informative and plays an important role – perhaps even more-so with pre-registration of individual studies becoming more common, and considering the resource investment required for MLRs. However, for my money I would trust the effect size estimate from an MLR much more than a meta-analysis (in the presence of publication bias). Regardless of these somewhat different perspectives, I thought the paper was a thought-provoking read especially given the experience the authors have leading and participating in MLRs.

I'll make more specific comments below that also bear on that overall assessment:

P5. I think this is quite an important and overlooked observation about the high correlation between MLR and meta-analytic effect sizes (even if there's a rather large discrepancy in raw effect size).

Yes, we agree!

P6-7. Lewis & Frank, 2016 is quite compelling. We've also had some similar cases where small tweaks had large impact (i.e., exact implementation of Anchoring in Many Labs 1 doubled the effect size compared to the original implementation). I generally

agree that in many (most? All?) cases you could likely go back and tweak a (non-false-positive) finding to get a larger effect size and I'd love to see more systematic work on this.

Yet, I'm a bit skeptical that this is a substantial contributor to the MLR vs meta-analysis effect size discrepancy because in my experience these are exceptions. We've also seen quite a few instances where authors have tried to go back and tweak replications in hopes of larger effect sizes, which did not end up changing much. Two examples that come to mind are Ego Depletion (Hagger MLR followed by the Vohs MLR), and Flag Priming (failure to replicate by Many Labs 1, multiple follow-ups from original authors with similar effect sizes). Many Labs 5 also found that revised protocols reviewed by original authors produced similar effect sizes. And, I suspect in most cases we would never hear about attempts that failed to produce a larger effect size. We also often get requests for specific analyses from original authors (e.g., only US undergrad participants, or specific exclusions, or a specific mode of presentation). In the studies I've been a part of, the large majority of these intuitions about what would result in a larger effect size was not evidenced in the data. So, perhaps I would quibble with the notion that these may be "fairly common in MLRs" (P7) but of course, fine to just disagree on this.

Thank you for this thoughtful comment. We think this issue is primarily relevant for MAs (there was an error in the original manuscript "fairly common in MLRs" should have read "fairly common in MAs"). Nevertheless, we suspect it's hard to estimate the potential of small methodological tweaks to influence effect size without systematic study. Notably, one of the examples you point to--ego depletion--reports a effect size discrepancy between the first MLR (Hagger, et al., 2016) and a subsequent one with a revised method (Dang, et al., 2021) to be a factor of 4 (.04 vs. .16). This is roughly comparable to what was found by Lewis and Frank (2016), though the original effect size in that case was much larger (.71). This suggests that, while methodological tweaks may not turn a small effect size into a large one, they may substantially increase it relative to the original effect size. More generally, though, we think studying this issue systematically would be an interesting and important area for future research, and could be done using the types of data sources you reference here (e.g., comparing ESs from original vs. author-revised protocols).

P9. I'm curious where these null results are coming from – are they being reported in papers that otherwise report significant results in favor of some phenomenon? If so, I wonder if there may still be some selection bias where authors are more likely to report null results "in the right direction" than null results in the opposite direction. But, I'll rely on the other reviewers here because I'm not an expert in bias correction methods.

This is an interesting possibility. We know that at least some of these cases of non-significant *p*-values are also coming from experiments where the *p*-values were significant in the original paper, but became non-significant only after the meta-analysis recalculated effect sizes. For instance, for the case of ego depletion in KSJ, many of the effect sizes in the MA (Hagger et al.,

2010) have non-significant *p*-values whereas the effects reported in the original paper were significant (e.g., Baumeister, Bratslavsky, Muraven & Tice, 1998; Bray, Ginis, Hicks, & Woodgate, 2008; Bruyneel, Dewitte, Vohs & Warlop, 2006). This is likely due to how standard error is calculated for the effect size in the meta-analysis, which may differ from the method in the original paper. It's not clear, therefore, how often the null effects in the meta-analysis were actually null effects in the original papers (which, notably, is a problem for bias publication methods).

> *P11. In reporting the sensitivity analysis results, I think it may be informative to also show some metrics for how much the meta-analytic effect sizes are decreasing (and where it puts them in relation to the MLR estimates). In addition to the percentages currently reported, I think this may give the reader a better sense for the magnitude of the adjustment in this worst-case publication bias scenario.*

Thank you for this suggestion. We have added this information to the manuscript:
"Across the meta-analyses, the mean naïve and worst-case estimates were d = 0.39 and d = 0.17, respectively. Thus, the worst-case estimates were on average 32% as large as their corresponding naïve estimates, with a mean absolute difference of d = 0.25. "

> *~P13. The point that individual labs may be less motivated to carefully adapt materials in MLRs is well taken, especially a general lack of pilot testing/tweaking across iterations in specific locations. However, with the imagined contact example I would think that if "muslim" was (highly) variably suited across sites we would expect high heterogeneity in effect size between sites. Instead, we see quite low heterogeneity, no moderation by whether labs were in the US vs international, and the participating lab in Britain also showed a quite small effect size (d = 0.18). (Disclaimer: I'm an author on that paper and likely biased, so this is just food-for-thought and the authors should feel no pressure to make changes here).*
>
> *I would expand that line of thinking to MLRs in general, although perhaps I'm thinking about this incorrectly. But if a substantial contributor to the MLR vs meta-analysis effect size discrepancy is that individual labs in MLRs are failing to adequately adapt their materials, resulting in protocols/stimuli that vary widely in suitability at individual sites, wouldn't we expect that to be reflected in quite large heterogeneity in those effects? And so far the general theme across MLRs (as noted) has been a story of surprisingly little variation across sites. That also goes for various moderators that have been looked at, generally: testing for moderation by the order effects were presented, in-lab vs online, US vs international, time in semester, mode of presentation. So, from my view the empirical evidence thus far is supporting the idea that we generally overestimate the importance of these contextual factors.*

This is a good point, thank you. There are potentially two types of methodological variance across replications of a given phenomenon. The first is random -- labs might implement the paradigm in a way that leads to a larger or smaller effect size. For instance, in the category

learning experiment referenced in the paper (Lewis & Frank, 2016), one lab might replicate the pattern with easy to learn stimuli (e.g., geometric shapes), resulting in a large effect size, while another lab might implement it with difficult to learn stimuli (e.g., plants), resulting in a small effect size. The second source of variability is more systematic -- labs might alter the paradigm in a way that leads to higher construct validity in the local experimental context, but differs superficially from other replications (e.g. changing the identity of the outgroup in a social experiment).

Importantly, these two types of variability have different consequences on heterogeneity and aggregate effect sizes in MAs compared to MLR (see table below). We agree that the inability to tweak methodological factors that influence construct validity in a local context (like the "muslim" example referenced above) should lead to high heterogeneity in MLRs. However, this is at odds with the decrease in heterogeneity that results from fewer sources of "random" variance in methodological factors. In sum, without a more precise estimate of the magnitude of these effects and a more formal model of their interaction, it is difficult to make strong inferences about the underlying sources of these differences.

	Influence of random methodological tweaks	**Influence of construct validity methodological tweaks**
MA	Increases heterogeneity	Decreases heterogeneity, Increases aggregate ES
MLR	Decreases heterogeneity	Increases heterogeneity, Decreases aggregate ES

Finally, we'd note that there is some evidence for appreciable heterogeneity within MLRs (see, for instance, the re-analysis of Olsson-Collentine, et al. 2020, where Tau is estimated to be .21; Errington, et al, 2021, *eLife*).

> *P13. I agree original authors are more motivated to find an effect, and compared to MLRs they are certainly more likely to perform a series of studies tweaking the methods. However, this is a bit of a double-edged sword. Especially when pre-registration is uncommon, I think this leads to more motivated reasoning, (unintentional) QRPs, and file-drawering. And, I'm not sure their motivation in any one study is that much greater because they know they can discard it if it doesn't work: this is perhaps best evidenced by a willingness to run severely underpowered studies. While individual labs in MLRs may not be super motivated, I actually think replication leads have a very high incentive to "get it right" – these are high stakes affairs that generally only have one shot and authors know many eyes will be scrutinizing them.*

Thanks for this comment. This is an interesting point, and is consistent with our suggestion that something in addition to publication bias underlies the larger effect sizes observed in MAs. It is possible that original authors are more motivated in general, and this leads them both to do

more QRPs, but also to implement a paradigm in a way that is most likely to lead to a large effect size.