Response to editor and referees on MS#MEE-20-02-118

Introduction

We were deeply disappointed by the Editorial decision to "Reject with an invitation to resubmit" our Comment, which documented and corrected uncontroversial and unambiguous mathematical errors in the Clark et al. (2019) paper. We were particularly surprised by this editorial decision given that **absolutely no critical issues** were identified by the two independent reviewers, both of whom confirmed our description of the **fatal mathematical and methodological errors** in the Clark et al. (2019) paper, and of our corrections of their unbiased statistical estimators (with one minor request for clarification of the math by Reviewer 3).

However, despite these positive reviews, the editors appear to have based their decision to reject our Comment entirely on a single review by *one of* the *original authors of the Clark et al. (2019) paper* (Reviewer 1), who wrote a 9 ½ page, single-spaced review of our manuscript which, despite its length, did not address a single point in our critique. Instead of dealing with our arguments, the Reviewer/author of the Clark et al. paper attempted, in a self-contradictory manner, to reframe the fatal mathematical and methodological errors we demonstrated as simply a case of *us* misunderstanding their true research "goals". They then proceeded to dismiss the mathematically correct unbiased estimators we derived as being just a "clever" derivation that was ultimately irrelevant to their paper since it was a solution to some alternative question. What 'alternative' question our *correct* estimators was supposed to address, and how that question differed from that in Clark et al. was never quite explained.

Reviewer 1's contention is simply untrue: our whole critique was based on the goals that were clearly spelled out in the original Clark et al. paper, and the unbiased estimators derived in our Comment are directly analogous to those in the Clark et al. paper. We even provided extensive examples and blockquotes in our Comment taken directly from Clark et al. (2019) to show their purported goals and applications.

Reviewer 1's attempt to couch this issue as a set of simple misunderstandings is perhaps understandable given that we showed that nothing in the Clark et al. paper is salvageable: literally every single element is both wrong and largely incoherent, aside from the infinite population estimators which they copied from a post by user "Glen_b" on an online forum. All the mathematical and methodological issues that we discovered are objectively and provably correct, and cannot be simply dismissed as us having different research goals or objectives:

1) All the invented (not derived) statistical measures described in Clark et al. (2019) are **wrong** and were at no time ever mathematically justified by the authors, let alone proven. We derived the correct statistics, and in doing so proved theirs to be incorrect (i.e., biased).

2) The stated purpose and use of the sample estimators that the authors provided will not work in the way that they claim. In particular, the shocking and incorrect claim that their sample statistics can estimate true population values *without replication* contradicts basic statistical practice and common sense.
3) The authors make a dangerously false claim that one can treat distinct experimental communities as though they were sample replicates from the larger regional pool. This inappropriate use of cross-community comparisons betrays a serious misunderstanding of what relative yields in BEF experiments represent.
4) Finally, their *particular justification* for extending the Loreau-Hector method by claiming that problems with the baseline can be controlled for, while technically true when measuring the net biodiversity effect, overlooks the *non-measurability* and hence *non-estimability*, of selection and complementarity under nonlinearity – the ultimate purpose of their paper.

Instead of addressing any of these issues, Reviewer 1's comment became a long exercise in both moving the goal posts after the fact by claiming completely new (and untenable) uses for their incorrect estimators, and distracting the reader with irrelevant non-sequiturs or outright false claims about basic statistical definitions. Below is just a *partial* list of examples:

i. Reviewer 1's comment still claims that it is "hard to dispute" that their statistics "are unbiased" despite us having **mathematically** demonstrated the contrary. Worse, their comment showed that they still do not know what the correct definition of statistical "bias" is – something they could have easily found from the first three hits of a web search.
ii. They claim our correct unbiased estimators are not so much 'correct' as simply different, and are thus for estimating different quantities than theirs – without ever explaining what those different quantities could possibly be.
iii. They go on a completely irrelevant tangent by talking about "mixed-effect models", largely when referring to infinite vs. finite statistical populations, not realizing these terms largely relate here to whether sampling occurs with/without replacement (which determines the *independence* of successive draws), not whether the underlying parameters are fixed or not.
iv. Not knowing the true statistical use of the term 'infinite population', they now claim that their statistics were always for an "infinite population" which is both untrue, and if taken seriously would make their correction factors pointless (and their result even more incorrect than before), while rendering their whole paper unnecessary from the start since the infinite population statistics were already known and even copied from a post by user "Glen_b" on an online forum.
v. They defend the shocking claim made in their original paper that sample replication is not required by insisting now that their estimators were only ever meant to gauge the "sign" of the quantity, not estimate its actual value. This surprising new goal does not appear anywhere in the main text or the appendix of Clark et al. (2019). Aside from this newly invented goal not appearing before, it is still both (i) untenable, as replication would likely still be needed even if one only wanted the sign of the quantities, and (ii) it

is flatly contradicted by explicit claims made elsewhere *in Reviewer 1's referee report* that the "purpose [of the estimators] was to generate estimates of selection and complementarity effects that are comparable across different n and N".

All of the mathematical and methodological flaws in the Clark et al. (2019) paper can be viewed as a series of nested issues ranging from the *specific* to the more *general* – that is, a move from the specific presentation of the incorrect statistics to more general issues relating to the implementation and justification of the statistics (see Figure R1). Solving an issue at one level (even if/when possible) would still lead to the failure of the authors' method at the next, more general level. Taken as a whole, there is simply no way to spin the colossal breakdown of the Clark et al. method as anything other than what it is.

Overall, we thus believe that our Comment warrants publication because it uncovers the many conceptual and mathematical errors in the Clark et al. (2019) paper. The readers of Methods in Ecology and Evolution need to be made aware of these significant issues in order to avoid making future mistakes. We now provide a point-by-point response in **bold** to each point raised by the Associate Editor and the Reviewers.

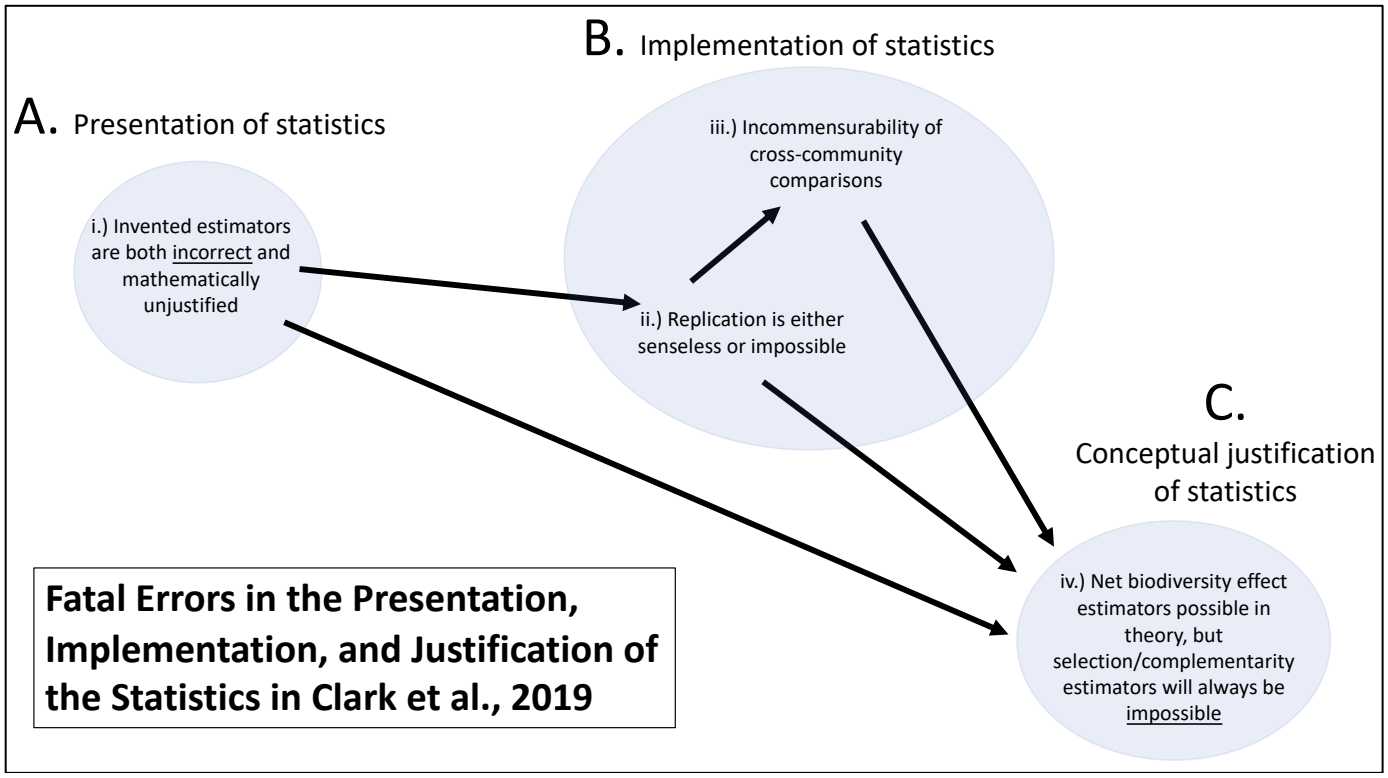


Figure R1: Fatal errors (with increasing degrees of generality) in the *presentation*, *implementation*, and *justification* of the statistics in Clark et al. Overcoming or correcting errors at a given level still will lead to the Clark et al. method failing due to the even broader mathematical/statistical issues that will be encountered at the next level. (A) Presentation: All estimators in Clark et al. were (i) presented without derivations/proofs and were shown to be incorrect (i.e., not unbiased). (B) Implementation: Even with the correct estimators one cannot use them as suggested because of (ii) the need both for large replication and for sampling from the full community, neither of which would be available to experimenters (or if available would render the need for estimators moot); and (iii)

the untenable suggestion that different experimental communities could fulfil the needs of replication, which is dangerously misguided due to the incommensurability of cross-community comparisons. (C) Justification: Even if all these issues were somehow overcome, the specific justification for statistically extending Loreau-Hector made in Clark et al. will (iv) only hold when developing estimators for the net biodiversity effect, but not estimators for the selection/complementarity effects, the focus of the paper.

Associate Editor

> **E.1)** Thank you for submitting your manuscript to MEE. I apologise for the long time we took to reach a decision. The delay was due to several reasons. First, I got two reviews with opposing views so I looked for a third reviewer. This took a lot of time because many people declined my invitation to review your manuscript. Finally, I got the third review but the covid-19 situation in Spain is not easy. I am working from home and part-time so I am doing what I can these days but, anyway, I am sorry for the delay.

We know that Spain was struck particularly hard by the pandemic, so we really appreciate the Associate Editor's efforts to manage the review process and seek a third reviewer given the initial split decision. There is no need to apologize for any delays; they are to be expected in these difficult times.

> **E.2)** As you will see, Reviewer 1 is very critical, Reviewer 2 is quite positive, and Reviewer 3 thinks you make some relevant points but also that other parts of your paper are redundant.

It is important to note that both independent reviewers (2 and 3) agreed that our Comment uncovered and corrected significant errors in the Clark et al. (2019) paper. The only dissenting opinion was that of Reviewer 1, the author of the paper being critiqued, who unsurprisingly did not approve of our critical Comment. As we explain below, we disagree that some sections of the manuscript are redundant.

> **E.3)** I think all these seemingly different views could be combined into a single way to revise your manuscript (but, of course, please address the reviewers' comments as you see fit).
> The path I am suggesting to go would be as follows. I would like you to consider if what reviewer 1 points in his section A (difference in goals) could be the frame of your contribution, instead a confrontational text saying that Clark et al contribution is flawed. I think you provide very interesting insights that need to be published and would develop our understanding of complementarity and selection. But this does not necessarily imply that previous contributions are flawed.

As we explain below, our Comment and the original Clark et al. paper had the same goals and sought to derive the same quantities. There was no "difference of goals" and our results/estimators cannot coexist with those of Clark et al. (2019) because they directly contradict each other. In other words, there is no way of reconciling these two sets of results by vaguely suggesting that they serve different purposes. In our Comment, we proved that

the estimators described in Clark et al., which involved the insertion of underived *ad hoc* (i.e., improvised and unjustified) correction factors into mathematical expressions in order to make sample estimates converge onto the desired true population parameter values, are not unbiased, and then we mathematically derived the correct unbiased estimators. Proving that our estimators are right and those derived in Clark et al. are wrong only requires taking their expected values and checking whether they match the expressions for the corresponding population parameters. Additionally, unlike Clark et al., our unbiased sample estimates of selection and complementarity always add up to the sample net biodiversity effect, which is itself an unbiased estimate of the true (population) net biodiversity effect. One would think that would be the minimum requirement for calling something a partition. Overall, there is simply no way of diplomatically writing around these fatal errors in the Clark et al. paper.

> **E.4)** I think you provide very interesting insights that need to be published and would develop our understanding of complementarity and selection. But this does not necessarily imply that previous contributions are flawed. This is directly linked with comments by reviewers 1 and 3 about the tone of your text. I absolutely agree with this view. There is no need to use such a tone to make a significant contribution.

We respectfully disagree. There was absolutely nothing in our manuscript that could even plausibly be construed as striking a problematic tone. The only way the tone of our text could be taken as inappropriate is if one were to take at face value Reviewer 1's (one of the authors of the Clark et al. paper) assertion that our approach is merely one of many possible statistical approaches: that there are really no correct or incorrect approaches, just different perspectives. This recourse to subjective relativism – that is, to statistical issues being reduced to mere preference – is simply and flatly false. The issues we brought up were not due to a difference of opinion or of differing perspectives, but about the proper and improper *use and definition* of the most basic statistical concepts. The authors of the Clark et al. paper were shown to have *invented* out of whole cloth (without any mathematical justification) incorrect unbiased estimators, all because they were unaware of the correct definition of statistical "bias", as is clear in their original paper and confirmed in Reviewer 1's referee report. Even more importantly, they completely misconstrued the nature of the sampling and replication that would have been required for their method to work.

The scale and the magnitude of the errors that we described in the Clark et al. paper are so striking that it may appear jarring when laid out bare. However, we did not go out of our way to deride them or make *ad hominem* attacks of any kind. We simply and systematically pointed out every issue with the Clark et al. paper point-by-point, and backed-up our claims with mathematical derivations and logical arguments in a dry and dispassionate manner.

If it truly were the case that the incorrect definitions/uses of basic statistical terms/tools that we described in the Clark et al. paper were not in fact 'mistakes' at all, but just different approaches to different sets of problems, and that we and the paper's authors were dealing with different but equally valid approaches, then it certainly would be confrontational for us to claim that our approach is right and all others are wrong. If this were the case, then we

would agree with the Editors that our having described the statistics in Clark et al. as being wrong would be the inappropriate tone to strike for what might merely be a difference in goals or objectives.

However, this is *not* what is happening here. What we demonstrated in our manuscript were simple and objective mathematical errors of the most basic kind: the tools/estimators that were presented in Clark et al. were flat out incorrect for the purposes that they themselves clearly described. Additionally, we showed that these estimators, even when corrected, could *never* be used in the manner they suggested because of how replication and sampling work, and because of the nature of relative yields in BEF experiments. These are simple objective facts based on uncontroversial, widely-accepted and agreed upon principles of basic statistics. We described these mathematical and methodological errors in a fair and straightforward manner, without any needless embellishment.

Additionally, in regard to Reviewer 3's comment, it is simply not true that s/he remarked on the tone of our manuscript. Reviewer 3 just made an isolated offhand (and inappropriate) remark about us having a "taste for controversy" – which aside from not being true and completely irrelevant as to whether our mathematical arguments are correct or not – was not actually a comment on the tone of the current manuscript.

> **E.5**) The three reviewers make several major technical comments that would need your attention.

As we describe below, there were no major technical issues with our Comment at all. Reviewer 3 thought s/he found a mathematical mistake in one of our derivations, but we show that s/he was incorrect. Overall, our point-by-point response to the Reviewers below makes it clear that our results are mathematically sound.

> **E.6)** Finally, I also agree with Reviewer 3 in that you should avoid redundant bits that were already published in your previous papers. These sections should be removed and the manuscript should just include the strictly necessary bits.

We respectfully disagree. There are no real 'redundant bits' aside from a few passages which we have removed under Reviewer 2's suggestion; otherwise, everything in this section relates to Clark et al.'s justification of their extension – it is all strictly necessary. The so-called redundant section is a completely different explanation (with new arguments) of how nonlinearity undermines the entire class of Loreau-Hector partitioning schemes, including Clark et al.'s extension of this method. This point was misunderstood and mischaracterized by Loreau and Hector (2019) in their Comment on our original paper in Ecology. Unfortunately, we were blocked from clarifying this point in print because the editor who was handling our paper was removed by the former Editor-in-Chief at Ecology, who then proceeded to reject our Replies for political reasons, following a mock review process during which members of the BEF community saw it fit to suppress our ability to respond for non-scientific reasons.

Including a description of the nonlinearity issue in our Comment is critical because *it directly contradicts the explicit claim made in Clark et al. (2019)*, as published in Methods in Ecology and Evolution, that the nonlinear issue is not problematic because the Loreau-Hector partitioning scheme can be used with any baseline (page 2143, right column, second paragraph). We show, using completely *new* and previously unpublished arguments and mathematical results that, even if one controls for the issues arising from the default Loreau-Hector baseline and the artificial inflation of the net biodiversity effect, the complementarity and selection effects *will still not be measurable*, particularly for the plant systems analyzed in Clark et al. (2019).

This is not a rehash of arguments published elsewhere, but a critique related specifically to the justifications in the Clark et al. (2019) paper. More importantly, despite us carefully explaining in our manuscript the problematic nature of Clark et al.'s argument for extending the Loreau-Hector method, Reviewer 1 (one of the authors of Clark et al.) *once again reiterated the mathematically flawed justifications* for their statistical extensions (see for example Response to Reviewer 1 comment R1.35). Given that the authors *are still actively making the same flawed arguments* to justify their approach, we have no choice but to include this critique in our Comment; the authors have themselves once again demonstrated its importance!

Making this point is now even more critical because, even though we derived correct unbiased estimators of the selection, complementarity and net biodiversity effects, these estimates should never be used unless the linearity of the diversity-ecosystem functioning relationship in monocultures is verified for all species. The last thing we want is for researchers to use our unbiased estimates to broaden the (mis)use of the Loreau-Hector partitioning scheme when its assumptions are violated (and its results are thus non-sensical). It is also important to remember that independent *Reviewer 2 has* also confirmed the validity, soundness and importance of this conceptual critique, and has even suggested ways of streamlining it so that it is not too repetitive.

Hence, we believe that the nonlinearity section of the text discussing the conceptual-methodological issues of the Loreau-Hector partitioning needs to be included in our Comment because our description of the issue (i) is correct, clear and has never been published in a peer-reviewed journal, (ii) is directly connected to the claims, justifications and goals of the Clark et al. paper—that is, the conceptual issues of measurability are directly connected with the methodological concerns of Clark et al. paper, and (iii) cautions against the misuse of the Loreau-Hector partitioning scheme when its assumptions are not met – even if one uses the correct unbiased estimators that we derived.

> **E.7)** In summary, I think it is clear you have some relevant points to make, but the manuscript is far from a publishable form. The way in which you revise your manuscript would be determinant to make a decision.

We have made every attempt to streamline the presentation of the results and clarify any expositional issues with our derivations. However, as we mentioned above and further explain below, there is simply no way of writing around the many fatal errors in the Clark et al. (2019) paper. And when it comes to dealing with explicit mathematical errors it is inappropriate – indeed *unethical* – to insist that one simply 'split the difference' or find a middle ground between true and false claims.

We believe that our revised Comment warrants publication because it provides critical corrections to the major mathematical and conceptual errors that went unnoticed when the Clark et al. manuscript was reviewed. We respectfully caution that rejecting our Comment for subjective reasons such as not striking a sufficiently conciliatory tone would undermine the credibility of this field. Indeed, if purely subjective concerns about an insufficiently conciliatory tone are allowed to trump fatal mathematical errors when adjudicating a Comment for publication, particularly at a journal focused on publishing technical and methodological contributions, then we fear that this will only serve to undermine the long term legitimacy of the discipline. This peculiar disposition for papering over critical issues and embarrassing contradictions with calls for 'collegiality' can only perpetuate the spread of critical mistakes, while threatening to waste a generation of research effort.

Reviewer 1 (author of original paper)

> **R1.1)** First off, let me say I am truly delighted to see that you have thought so deeply about my paper. I often feel like I am dropping theory papers into the void, and it is lovely to see someone take my work this seriously. Similarly, I want to thank you for choosing to address your points in an article rather than a reply – I find replies to be generally too short and antagonistic to be useful (and, should it not be clear, I have no intention of writing a reply to your article, wherever it might ultimately be published).

We believe there is some confusion here. Our manuscript is a Comment to which Reviewer 1 will be invited to reply. It is not a standalone contribution.

> **R1.2)** Given that your article offers such a pointed and specific critique of my article, I will break this review into two sections. In the first section, I will explain what I think are the main differences between the intention of your analysis, and the goals of our method. For clarity, I will use your notation, rather than ours. I am hopeful that these will reveal that we don't actually disagree, but rather have designed two different methods that answer two distinct questions. In the second section, I will provide a more standard review of your article. To be clear, I think that the correction that you provide totally has merit, and is a useful tool for answering some questions, even though these aren't the questions that we wanted to test.

As we explain below, we disagree. Both your original paper and our Comment sought to derive unbiased estimates of selection and complementarity for the same purposes. However, we showed that your estimates were not unbiased. More so, you claimed that the unbiased formulas you gave allowed your estimators to converge to the true value when

sample size approaches the finite population size. This is not the definition of "unbiased" – at best this is the definition of being statistically "consistent" in that the estimators might converge onto the desired population parameter value with increasing sample size. As we explained, a consistent estimator need not be unbiased and vice versa. There are an infinite number of ways of adding correction factors to the estimators of selection and complementarity in order to allow them to converge, but the definition of unbiased (and consistent) estimates is more strict and rigorous. What's more, not all the *ad hoc* correction factors you inserted into the infinite population estimators will make them converge to the true values, i.e., the estimators presented in Clark et al. (2019) are not even all statistically consistent.

Additionally, we showed that even the proper unbiased estimates that we derived cannot be used for the purposes described in your paper because of the intractable issues relating to replication and to making comparisons across experimental communities, and due to the non-measurability of complementarity and selection. Hence, these issues cannot be explained by 'differences in goals'.

> **R1.3)** Section A. Difference in goals
> Briefly, our goal was to develop a correction that allowed comparison of the classic Loreau and Hector selection and complementarity effects across gradients of species richness – both when that gradient is driven by incomplete sampling of a single larger community, and when it is driven by differences in the total diversity of several different communities. This is the test that we show in Fig. 5, which depicts relationships between the selection and complementarity effects and the diversity of several experimental and natural grassland communities.
>
> You are entirely correct that our method is based on the assumption that both subsamples of size n from a larger community, and samples of the entire community of N species, are drawn from an underlying distribution. As you suggest, one can interpret this to mean that we assume that both n and N are drawn from an infinitely large community. An alternative interpretation (which is the interpretation that I follow and, I think, one that is generally more common in ecology and hierarchical statistics) is that both n and N are drawn from an overarching distribution – i.e. that some process exists which generates values for monoculture biomass and deviation in relative yield for each species (hereafter M and dRY, respectively), and that even perfect measurement of a subset of n species, or of the total community of N species, will yield observations that differ from the "true" values of the underlying process.

Your "alternative" interpretation of an infinite population is incorrect. An 'infinite population' here merely means that sampling occurs with replacement (which could be interpreted as sampling from a distribution), and 'finite population' means that sampling occurs without replacement (so that repeated sampling depletes the population and thus affects subsequent sample estimates). 'Infinite' and 'finite' populations thus have nothing to do with whether

the population parameters being estimated are fixed or random (i.e., sampled from a distribution), which is what you seem to be suggesting.

To explain this distinction, imagine a scenario where we fill an urn with a population of 80 red marbles and 20 blue marbles. Here, the population parameter describing the proportion of red marbles is fixed at 0.8. If we drew a total of 10 marbles one at a time without replacement and obtained 9 red and 1 blue, our 'finite' sample estimate of the proportion of red marbles would be 0.9. On the other hand, if we drew a total of 10 marbles one at a time with replacement and obtained 7 red and 3 blue, our 'infinite' sample estimate of the proportion of red marbles would be 0.7. However, in both cases, the population parameter was fixed at 0.8. Hence, whether we derive estimates from an 'infinite' or a 'finite' population has no bearing on whether the population parameter is fixed or random.

> **R1.4)** I suspect that part of the confusion comes from what you understand by "population-level", vs. what is commonly used in statistical or ecological parlance. E.g. as I understand it from past conversations I've had with physics colleagues, it is typical in some fields to assume that observations at the "population"-level describe perfect observations of the complete set of values of some variable. In contrast, in statistics, and particularly in hierarchical models, estimates at the "population"-level are still assumed to contain uncertainty.

We disagree. We are not confused about what population-level statistics are because of differences in backgrounds or lack of exposure to ecological statistics (both of us have PhDs in Biology). Rather, the difference you seem to be referring to is between a frequentist and a Bayesian interpretation of a population parameter. In frequentist statistics, the assumption is that sample estimates drawn from a population vary randomly but that the underlying population parameters are fixed (that is, if our sample included all of the observations in the population, it would yield the true [fixed] population parameter value). Conversely, in Bayesian statistics, population parameters are not assumed to be fixed but rather random and characterized by some statistical distribution. However, none of this is related to whether estimates drawn from infinite or finite populations are more ecologically meaningful or whether the population parameters should be considered fixed or random.

> **R1.5)** For example, if I were to fit a mixed effects model to the abundances of N species, the species-level "random" estimates would be called the "sample"-level values, whereas the fixed effect estimate of the average abundance observed across all species would be called the "population"-level estimate, which would be expressed in terms of both a mean and a standard deviation. The underlying assumption for this interpretation is that even if the sample of N species is "complete" (i.e. we've successfully sampled all species in the community), the "population"-level statistic that we are interested in is meant to describe the overarching distribution from which the individual species-level abundances were drawn, rather than a fixed value that describes the peculiarities of the specific sample of N species that we happened to observe. I wish

that I had made this distinction clearer in the main text – I fear I hadn't anticipated statistically competent readers from outside of the general field of ecological statistics.

Again, we are not misunderstanding you because we have a background in physics or are unfamiliar with ecological statistics. We have taught mixed-effects models to undergraduate and graduate students. We are very familiar with them. What we do not understand is the purpose of this entire section of your review discussing 'mixed-effects' and 'hierarchical models'. As mentioned above, none of this has any bearing on sampling from a finite population or an infinite population, or the legitimacy and unbiasedness of your estimates.

The motivation behind this entire section appears to be to suggest that estimates from an infinite population are more ecologically meaningful than those from a finite population. However, even if we remain agnostic about this – after all we derived unbiased estimates of selection and complementarity for *both* finite and infinite populations – this interpretation cannot hold. Your estimates for selection and complementarity do not match the expressions for unbiased estimates that we derived for *either* infinite or finite populations. Instead they combine elements of both types of unbiased estimates in an *ad hoc* fashion for no legitimate reason. Hence, we are not sure why you keep harping on this.

R1.6) A practical example of how this type of "population"-level effect arises is actually available in the Jena data that we used. At the highest diversity level in the experiment, there are multiple plots that were planted with identical compositions. Thus, there is no single value that describes M and dRY at that diversity level, even if we only consider the species and conditions at Jena. Rather, there are a distribution of potential values, from which the realized observations in the experiment are drawn.

This is a confusing statement. What you are describing is random variation between replicates or samples of the same experimental treatment, which in this case consists of communities with identical compositions. Such random variation between replicates or samples says nothing about whether the population parameter should be treated as fixed, as assumed in frequentist statistics, or random, as assumed in Bayesian statistics. Furthermore, the fact that variation occurs between replicates of the same experimental treatment cannot be used to argue for the primacy of estimates based on infinite populations over those from finite populations since infinite and finite populations are defined by whether sampling occurs with vs. without replacement, and have nothing to do with whether the population parameter is fixed or random. This flawed line of reasoning, which appears to confound variation between samples with variation at the level of the population, could even be extended to argue that any random sample-to-sample variation invalidates the entire field of frequentist statistics since the latter assumes fixed population parameters! This is obviously incorrect.

R1.7) Thus, the main reason that we follow this assumption is that we wanted to find estimates that would only be correlated with community richness given a change in the underlying process that generated M and dRY. As an example, consider the following

hypothetical situation. Imagine a system in which species grow differently in monoculture than in mixture, but that the difference is independent how diverse the multi-species community is (other than the effects of differences in relative seeding density, as hypothesize in the original Loreau and Hector partition). For simplicity, let us imagine that this process leads to an average covariance between M and dRY of 1, for any community of 2 or more species. Now, imagine that we try to measure the covariance between M and dRY across several different communities that vary in diversity – e.g. communities of size N = 2, N = 3, N = 4, etc...

Despite the fact that there has been no change in the process that generates M and dRY across these communities, if we use the uncorrected formula cov(M, dRY) = mean((M-mean(M))*(dRY-mean(dRY)), then our measurements of the covariance would vary quite strongly across communities, i.e. by a factor of (N-1)/N. That is, for a community of N = 2, we would, on average, find cov(M, dRY) = 1 * (2-1)/2 = 1/2, for a community of N = 3, we would on average find a covariance of 2/3, for a community of N = 3 we would find an average covariance of 3/4, and so forth. This would, erroneously, lead many practitioners of the method to conclude that the underlying process that drives the relationship between M and dRY must vary with diversity – but really, all that this shows (as both your paper and my paper discuss at length) is that an uncorrected estimate of covariance is sample-size dependent. To avoid this problem, we apply the correction cov(M, dRY)*N/(N-1). Now, if we compare these communities of size N = 2, N = 3, etc., we will find that the expected value of the corrected covariance, regardless of the number of species in the community, is equal to 1. This would lead a practitioner of the method to correctly conclude that there was no change in the underlying process as a function of diversity.

We are not sure what you are trying to say here. However, this example seems problematic because it assumes that the numerator [i.e., the sum of the product of the deviations of M and dRY from their respective means (M-mean(M))*(dRY-mean(dRY)] is fixed at 1. Your *ad hoc* correction factor then merely cancels out the denominator to ensure that cov(M, dRY) remains constant with changes in community size. However, in reality, the sum of the product of the deviations of M and dRY from their respective means (M-mean(M))*(dRY-mean(dRY) will vary with different community compositions and sizes so the numerator will also change. Hence, there is no reason to assume that it should remain fixed.

Apologies if we misunderstood what you were stating. However, this example has no bearing on the legitimacy of your *ad hoc* correction factor and the incorrect claim that you derived unbiased estimators of selection and complementarity. Your correction factor in this example merely appears to scale cov(M, dRY) to account for differences in community size. The same result (controlling for community size) can be achieved by finding the true unbiased estimate of the population parameter of interest and taking its expected value from replicate simulations, as we have shown in Figure 1 of our Comment.

R1.8) As a brief aside, I think it is worth explaining what, exactly, this type of comparison of the partition can tell us. As you rightly point out, once this correction has been applied, the sum of "corrected" complementarity effects and selection effects no longer provides an unbiased estimate of the net biodiversity effect at the community level. Clearly, we failed to make this point sufficiently clear – although we state it several times in the supplement, and in the accompanying R package, I appear to have removed that sentence from the main text at some point in the revision process – I apologize for that, and understand why doing so could have led to confusion.

There is no confusion on our part. We were always aware that you were not primarily interested in finding an unbiased estimator for the net biodiversity effect; that was *not* the point for us talking about the biasedness of this net biodiversity effect. Our point was that if the two estimators you *were* interested in – complementarity and selection – were truly unbiased, then their sum should also be unbiased; *this is a simple mathematical consequence of summing the expected value of two random variables*. The fact that the sum of your unbiased estimators was itself not unbiased should have alerted you to the issues that we described in our Comment, and should have served as an indication that something was seriously wrong. In other words, you may have been aware of the fact that the *sum* of your two estimators of interest was not unbiased, but you did not recognize the dire implications and consequences this had for your incorrect claim that the two component estimators were themselves unbiased. That is the point we are making here. It is yet another proof that your estimates of selection and complementarity are biased.

R1.9) But, at least in my experience, it is not common practice to use the Loreau and Hector partition to calculate the net biodiversity effect. After all, if the main variable of interest is the net biodiversity effect, then it is usually much easier to simply calculate this value directly from the raw data (and indeed, it would need to be calculated in the process of deriving the partition in the first place). Rather, studies tend to use these statistics to say something about the kinds of associations that seem to be driving biodiversity effects. As we note in our paper, the "null model" of the Loreau and Hector partition asks a relatively simple question: am I better off planting N small monocultures, each covering 1/N of the total area available, or am I better off planting a single large mixture, including all N species grown together? Positive selection effects indicate that species with high monoculture biomass (i.e. high M) have, on average, higher mixture biomass than would have been expected from the single monoculture plot grown on 1/N of the total area (i.e. disproportionately high dRY). Positive complementarity effects indicate that on average, mixtures produce higher yields than expected from monocultures, but that these differences do not correlate with monoculture biomass. Thus, selection effects are generally used to summarize how species with high monoculture biomass grow in mixture, whereas complementarity effects are generally used to summarize the average change in mixture vs. monoculture biomass.

As we explained in our Comment, this scenario is still problematic. Indeed, although rescaling the area of the mixtures by N to keep density identical across mixtures and monocultures might avoid the inflation of the net biodiversity effect described in Figure 2 of our Comment, the nonlinear relationship between density and ecosystem functioning in monocultures will still lead to selection and complementarity being non measurable and thus incapable of being estimated [pgs. 19-20].

> **R1.10)** Our statistics are "unbiased" with respect to community and sample size, in the sense that they are internally consent, and will generate the same average estimate for selection and complementarity effects regardless of n or N, so long as M and dRY are driven by the same underlying processes across species (or, in your parlance, given that M and dRY are drawn from the same infinitely large sampling distribution). Thus, we can compare statistics derived from an incomplete random samples of n species, vs. those derived from a complete sample of N species, vs. those from several different communities that vary in N, without worrying that the differences in sample sizes are driving changes in our complementarity and selection effect estimates. This property is incidentally what we show in the right panels of Fig. 4. Regardless of the size of the subsample that we take (horizontal axis "N" in that figure, or n in your notation), we can always generate estimates of selection and complementarity effects that are n average comparable to those for the full community (i.e. "Q" in that figure, or N in your notation). Thus, even if we only observe a random sample of n species, we can still generate, on average, the same estimate of the statistic that we would have calculated had we measured all N species.

This is a far cry from showing that your estimates of complementarity and selection are unbiased (which would also achieve the objective that you described), which is why both we and the independent reviewers are objecting to your use of the term 'unbiased'. This is just an *ad hoc* correction with no mathematical justification. Additionally, the use case for these 'unbiased' estimators that you are describing does not make sense because the same species taken from different communities and compositions will be characterized by different dRY values because the latter are not properties of the species but are structural properties of the community in which they are embedded. Hence, the *ad hoc* correction you created cannot facilitate comparisons between completely incommensurable quantities.

> **R1.11)** You can certainly argue that you don't agree with the sample size scaling that we use at the level of the full community N, and you are entirely correct that the correction that we use yields a biased estimate of the net biodiversity effect. But again, as stated above, the purpose of this study was not to generate an unbiased estimate of the net biodiversity effect. The purpose was to generate estimates of selection and complementarity effects that are comparable across different n and N.

The problem is that you *explicitly and repeatedly* claimed in your paper to have found unbiased estimates of complementarity and selection (even in the abstract), when in fact, you had not. You claimed that your "unbiased" formulas would allow your estimators to

converge to the true finite population value, but that is not the definition of what an unbiased estimator represents; at best it would make them "consistent" estimators (as Reviewer 2 pointed out). However, not all of your estimators will even converge as sample size increases – in other words, not all the estimators presented were even statistically consistent. At this point we are not sure what other possible use they could have.

As for the net biodiversity effect, because this is the sum of complementarity and selection, it being biased is of critical importance as it indicates that one or both of your estimators for selection and complementarity are themselves biased. The fact that you personally were not interested in estimating the net biodiversity effect itself is completely irrelevant. Furthermore, despite your stated lack of interest in the net biodiversity effect estimator, you nevertheless still attempted to force it converge to the true population value by inserting yet another *ad hoc* (unjustified) correction factor into the expression for the sum of the complementarity and selection estimators – a sum, which the authors of Clark et al. themselves admitted may be used in cases "where it is vital that the sum of the selection effects and complementarity effects equals the change in relative yield". No additional *ad hoc* or made-up correction factor would have been required by the authors if the correct unbiased estimators for selection and complementarity had been used from the beginning.

> **R1.12)** And, while I am perfectly receptive to critiques of the Loreau and Hector partition, the goal of this paper was not to invent a new partition or a new null model. It was to show the enormous community of practitioners who are currently using this method how they might generate estimates of the statistics that are more comparable across communities and samples of different sizes.

We were not suggesting that you develop a new partition or null model. Our Comment simply showed that (i) your stated goal of allowing researchers to make comparisons across experiments characterized by different community sizes was flawed because dRY values are not properties unique to species but of the communities they are embedded in, (ii) the correction factors you used were *ad hoc* and unjustified, and did not yield unbiased estimates of either selection or complementarity as you claimed, and (iii) the Loreau-Hector method should not be used let alone extended unless the assumption of linearity in monocultures can be verified for all species.

> **R1.13)** Importantly, I am relatively certain that there is no way to generate estimates that simultaneously are uncorrelated with sample size given a fixed underlying process, and that provide an unbiased estimate of the realized net biodiversity effect. One can only do one, or the other. This is simply because, as both of our papers state, the properties of the sample of N individuals will never perfectly reflect the properties of the distribution that generated M and dRY. I was actually initially quite excited as I read your paper, as I hoped you had somehow found a solution that does both – but alas, no. As with the standard "uncorrected" formula, your method of calculating, e.g., selection effects simplifies to mean((M-mean(M))*(dRY-mean(dRY)) for n = N, which, as discussed above, leads to intrinsic correlations between the selection effect and community size

> N, when we compare the metric across multiple communities (again, under the assumption that both n and N are drawn from a common distribution).

This a confusing statement. It suggests that the goal of your *ad hoc* corrections was simply to delete any term that scaled the sample estimator. In the case of the selection effect for example, the goal would be to compute the sum of the product of (i) the deviations of the monocultures from their mean and (ii) the deviations of the dRY from their mean without dividing by N because doing so introduces a correlation between N and the estimator for the selection effect. However, that still does not make these measures comparable across experiments because dRY is a property of the community in which a species is embedded and not just its size. Hence, correcting for community size using your *ad hoc* correction factors is not sufficient to make the goal of your paper feasible.

> **R1.14)** To be clear, I think the fact that your statistic is intrinsically correlated with sample size N by no means implies that your method is "wrong" or "flawed". It simply demonstrates that your method is designed to answer a different question than ours is, and that it is therefore not suitable for the analyses that we wanted to be able to address with our method (i.e. comparing selection and complementarity effects across communities of different sizes).

You never quite explain what "different question" our method was "designed to answer". More importantly, your goal of "comparing selection and complementarity effects across communities of different sizes" is not possible. Even though what you sought to do is not feasible (i.e., because dRY is a function of community composition), at least using the correct unbiased statistics we provided would allo you to estimate the true population parameters you wanted because their expected values (across replicate simulations) are comparable across different sample sizes.

> **R1.15)** And, indeed, I think that the derivation is quite clever, and arguably produces a statistic that has great value. Specifically, if a practitioner for some reason needs estimates of the selection and complementarity effect that do add together to perfectly produce the observed net biodiversity effect, [...]

For the record, ours is not a "clever" derivation, it is a *correct* derivation of the unbiased estimators. More importantly, it is *an actual derivation* – that is, the unbiased estimators were actually derived and justified with proofs/mathematical arguments – as opposed to simply being *made up*, as was done with the incorrect estimators presented in Clark et al.

Additionally, what does it mean when Reviewer 1 states:

> **"if a practitioner for *some reason* needs estimates of the selection and complementarity effect that do add together to perfectly produce the observed net biodiversity effect"? [emphasis added]**

Since complementarity and selection are the only components that arise from partitioning the net biodiversity effect, shouldn't the *minimum requirement* be that they sum-up to the net biodiversity effect? Otherwise you have converted a proper partition into an improper and 'leaky' one.

And just to repeat: we did not derive the estimators with the explicit goal that they would "add together to perfectly produce the net biodiversity effect"; we derived them in order to show how one should properly derive the *correct* unbiased estimators for selection and complementarity. This is in contrast to the incorrect and mathematically unjustified estimators provided by Clark et al. Obtaining an unbiased estimator for the net biodiversity effect will be the inevitable consequence of having properly derived true unbiased estimators for selection and complementarity, whether one is interested or wishes to make use of the total net biodiversity estimator or not.

> **R1.16)** [...] and cares less about intrinsic correlations between community size and the values of the selection and complementarity effects, then I totally agree that your method would be superior to our method. I can imagine this applying in at least two situations: (1) studies that assume that they have perfectly measured the value of M and dRY (i.e. no observation error or process noise); or (2) studies that are interested in explaining the specific outcomes observed within a single plot at a single moment in time. That said, I suspect that in most cases, ecologists will be more interested in using these partitions to study the general processes that produce M and dRY, rather than the peculiarities of individual realizations of M and dRY derived from a single measurement. So, I would strongly suggest that you make the implications of using finite vs. infinite sample corrections very clear (and of course, to do a better job at making the distinction clear than we did in our paper).

Again, we are confused by the claim that our estimates are problematic because of their correlation with/dependence on community size. Taking their expected value (average) following multiple replicate simulations would yield estimates that are unbiased and unaffected by community size (see Figure 1 of our Comment). That is exactly the role of an unbiased estimator.

> **R1.17)** Note – I'm sure that you could try to brush aside this concern by saying that it would be trivial to derive the infinite sample covariance etc. using your statistic, with just a few corrective factors that are proportional to functions of n and N. But, of course, the same could be said about the metrics in our paper with respect to sample-level statistics. I fear that on some level you might assume too much mathematical competence in the average practitioner of these methods – In general, I think it is unreasonable to think that anything other than the precise equations outlined in a paper (and, more commonly, the functions supplied in an appendix or statistical package) will be used.

This is wholly unnecessary since taking the expected value of our unbiased estimators would yield the desired result: an estimate of complementarity and selection that does not depend on community size.

> **R1.18)** I hope that this helps explain how the points that you raise are separate from the one that we tried to address in our article. I am therefore a bit sad that you choose to call this difference a "flaw" and mathematically "incorrect". To me, this would imply that we claimed to be doing something that we failed to do (e.g. if our estimates at the level of n actually failed to return the same expected value as our measurements at the scale N). I hope that perhaps in retrospect, you can agree that what you propose is simply a solution to a different problem.

We disagree. The terms "incorrect" and "flawed" are accurate because you claimed that your estimators were unbiased, a term that has a specific statistical definition. However, we proved that your *ad hoc* (mathematically unjustified) estimators were not unbiased and then derived unbiased estimators that could (but should not) be used to achieve your intended goal of comparing selection and complementarity across communities with different sizes/compositions. This could be accomplished by taking the expected value (mean) of the unbiased estimates of selection or complementarity from a large number of replicate simulations. Hence, our unbiased estimates are the correct solution to the problem you posed. In other words, we proved that you failed to accomplish what you specifically set out to do. Ours is not "a solution to a different problem": it is the correct solution to the problem that you incorrectly claimed to have addressed.

> **R1.19)** One final point before moving on to the more formal review. One issue that both of our papers punt on, but that is obviously deeply important for any empirically tractable metric, is observation error. I think that a formal treatment (and perhaps correction) of these statistics that helps address the effects of observation error would be exceedingly valuable and novel – and probably would add a lot more to the literature than would either of our papers on sampling frequencies. For the most part, the biases that we discuss here will only shift estimates of selection and complementarity effects by a few percent (and, for most n and N, our metrics will both return values that are virtually identical).

We would agree that accounting for observation error would be important if we thought the comparisons that you propose to make across different communities and experiments were not fatally flawed and untenable. However, as we mentioned above, those comparisons cannot be made even if community size is accounted for, so accounting for observation error is not necessary.

Additionally, we agree that our unbiased estimates might differ only slightly from your biased estimators in some cases. However, we hope that you are not trying to claim that this is a 'distinction without a difference' or that the difference is not meaningful. That would be analogous to arguing that the sample formulas for the standard deviation and covariance are

only slightly different than their population counterparts, especially as sample size increases, so why bother using or teaching both sample and population formulas to students or including them in textbooks? However, we continue to teach (and software typically defaults to computing) sample estimates of these measures because they are unbiased. In other words, the small size of the discrepancy that may occur in some cases between the numerical values produced by our unbiased estimators and those produced by your biased estimators cannot be used to justify the latter.

> **R1.20)** But, given that observation error in most ecological studies is around 30-40%, these sample-size driven differences are dwarfed by observation error. I fear that there can be no general, analytically tractable solution to this problem – observation error occurs at the level of individual species monoculture and mixture biomasses, meaning that, e.g., dRY actually follows a ratio distribution. As we note in our paper, there are some limited circumstances under which analytically tractable corrections apply to such a distribution, but for the most part, solutions will need to be ad-hoc, and mostly numerically and simulation driven. If this is a question that interests you, I would be delighted to discuss the point with you sometime in the future.

We do not think such comparisons across experiments or different communities are feasible or justified. This has nothing to do with the statistical properties of these estimates but everything to do with their context-dependency. We hope that the serious issues we have uncovered here will dissuade you or anyone else from pursuing this line of inquiry in the future.

> **R1.21)** Section B: Primary review of manuscript
> 1. General comments
> As I note above, I think that your derivation is technically quite interesting, and that it addresses questions that are potentially of great interest to some ecologists. I have three general comments that I think might help your paper be more broadly accessible to readers within the fields of ecology and evolution, as well as several more specific comments.

There seems to be some confusion. Our paper is a Comment on Clark et al. (2019), not a standalone contribution. Its goal is to correct the record and address the issues with the *ad hoc* and biased nature of the estimates of selection and complementarity described in your paper.

> **R1.22)** 1. Guide for applications and description of scope:
> I think that in order for this paper to have the highest impact possible, it will be necessary to specifically outline a step-by-step procedure that practitioners should follow, as well as a discussion of the circumstances under which the metric is likely to give sensible answers, vs. when it might produce misleading answers. Ideally, I think this would require two steps.

As we explain repeatedly in the text, there is *no circumstance* under which these unbiased estimators should be used in the manner described in Clark et al. (2019). This is because the paper encourages comparisons between incommensurable experiments/communities.

> **R1.23)** First, I would suggest including some kind of a computer function along with your method, so that even readers who are not particularly numerically competent can still apply it. If you'd like, you could achieve this by releasing your own R package – or if you would prefer, I would be more than happy to include it in the existing package that we released with our paper (to be clear – how you would like to release the function is obviously entirely up to you, and I won't feel snubbed if you decide to release your own).

Again, because we do not want to encourage anyone to use these unbiased estimators in the manner described in Clark et al. (2019), we chose not to include any computer code to allow ecologists to compute these quantities without good reason.

> **R1.24)** Second, I would suggest including a few ecological examples that outline the kinds of comparisons where these methods would provide useful metrics. For example, in what kinds of systems would the assumptions of zero observation error and zero process noise hold? E.g. I can imagine this is more or less the case in some microcosm studies. Alternatively, in what kinds of systems does the precise net biodiversity effect observed in individual plots matter more than attempts to understand the general underlying processes that drive these effects? Again, I am sure that these sorts of conditions exist in some systems and studies, and I think it would be good to explore some examples – ideally even including analyses with empirical data.

None of that is necessary since we explicitly explained that these comparisons should never be made.

> **R1.25)** 2. Discussion of the differences between finite- and infinite-sample statistics
> Although you clearly take some pains to discuss the problems with infinite-sample approximations, I believe that there are two points that are generally missing from the discussion and analyses. First, as I outline above, there are both costs and benefits that come from utilizing the infinite-sample approximation, and arguably in many, if not a majority, of cases studied by ecologists, the infinite-sample approximation actually provides something that better reflects the questions that ecologist would like to have answered. I would suggest exploring these cases a bit more carefully in your main text. I would also take it as a personal favour if you could stress in your paper that the intention of the infinite-sample approximation was never to generate unbiased estimates of the net biodiversity effect (given that I missed saying this in the main text, I think that this would be a very helpful point of clarification).

Neither the infinite nor the finite estimates of selection and complementarity should be used in the way that you described in Clark et al. (2019). This is why we never compared the two in

terms of their applicability. Our point is that neither your biased estimators nor our unbiased estimators should be used in this way.

> **R1.26)** Second, for a wide range of n and N values, both the finite sample approach that you present, and the infinite sample approach that we use, provide almost identical answers – down to a few hundredths of a percent in many cases. I would suggest including some analyses (ideally, again, of empirical data) in which you can show the size of the difference between your estimates and our estimates. Although both you and I spend a substantial amount of time discussing the differences between these approaches, I fear we will find that it is a distinction without difference in the vast majority of cases that are likely to come up in the real world.

According to this logic, we should stop using the sample formula for the standard deviation and covariance because they only differ slightly from their population counterparts, especially as sample size increases. This is analogous to saying that a defective (biased) calculator that always adds the value 5 to the sum of two user-specified numbers *x* and *y* should continue to be used because the totals it generates are very similar to those computed by a non-defective (unbiased) calculator that simply sums the numbers *x* and *y*, especially as *x* and *y* become large. Why not just use the non-defective calculator and dispose of the defective one?

> **R1.27)**
> 3. Tone and content of the paper
> I obviously leave this entirely up to you and to the editors. But, I suspect from the structure and tone of your paper, that your backgrounds are primarily in physics or applied math (apologies if this is incorrect). In my experience, it is quite common to write papers in an intentionally confrontational manner, as it is a good way to make distinctions clear, and because those disciplines often have very specific definitions of terminology which are consistent across wide swaths of the literature.

As mentioned above, we are biologists. The tone is not confrontational, whatsoever. We are simply correcting the errors in the paper point-by-point in a dry and dispassionate manner. The claim about "tone" is only plausible here if one were to take at face value Reviewer 1's claim that the incorrect definitions and uses of basic statistical terms and tools that we described in Clark et al. are not really mistakes at all, but just different approaches to different experimental problems. If it truly were the case that we were dealing with different but equally valid approaches, then it certainly would be confrontational for us to claim that ours is right and all others are wrong.

However, this is not what is happening here. What we pointed out in our paper were simple and straight forward mathematical errors of the most basic kind: the tools/estimators that were presented in Clark et al. were flat out incorrect for the purposes that they themselves clearly spelled out. Moreover, their statistical estimators cannot be used in the manner they themselves claimed because the authors misunderstood the nature and scale of replication

and sampling that would be required, and more importantly, incorrectly treated distinct experimental communities as though they were different samples taken from a larger species pool.

These are unambiguous flaws, not differences in opinions or research agendas; they represent straightforward errors in statistics, they evince a serious misunderstanding of how basic sampling and replication work, and they involve the inappropriate use of BEF and ecological concepts, such as relative yields, when making statistical inferences. These are clear and objective *flaws* in the Clark et al. method, as has been confirmed by the two independent reviewers.

It is disingenuous to spin a perfunctory listing and correction of mathematical errors as representing a problem with the 'tone' by creatively reframing our comment as though it were (as you said elsewhere) "largely based on a misinterpretation of the purpose of [your] paper", since it clearly is not.

We (and the other two referees) have not misinterpreted the purpose of your paper. It is your misinterpretation of our mathematical arguments that have allowed you to claim that our critique boils down to nothing more than a difference in research goals. This, and this alone, is the basis for the unwarranted claim that there is "tone" issue with our manuscript.

> **R1.28)** My advice would be that in ecology, styles tend to be very different. Because ecology is so interdisciplinary, many pieces of terminology are used in different ways by different sub-fields (see, for example, my discussion of "population"-level statistics above).

We used and assumed the proper statistical definitions of populations and unbiased estimators. Unfortunately, it was your paper that apparently used an *ad hoc* definition (i.e., improvised, made-up) of what unbiasedness meant in order to justify the use of *ad hoc* correction factors that confounded the concepts of statistical consistency and bias. In other words, we are not 'talking past each other'. We are all referring to the same statistical phenomena and the differences in our estimates cannot be ascribed to different definitions of "population statistics" – your estimators just happen to be, <u>objectively speaking,</u> wrong.

> **R1.29)** Moreover, intentionally confrontational papers tend to give people the reputation of being a bully, which reduces future opportunities for collaboration and builds ill will in the discipline. I think that this is especially true in biodiversity ecosystem functioning literature, where there are so many angry schisms among different groups that I think it would be a shame to contribute to new ones. To be clear, I have spent most of my career working with physicists and applied mathematicians, so I in no way think that your intention was to be antagonistic or rude. But, in my opinion, I fear that you will find that papers in this style will generally have much less impact in the long-run.

In order for this portrayal of us as bullies to hold, we would need to have some kind of power over the field. However, neither of us holds any kind of editorial position at a scientific journal that would confer such power, nor are we "big names" in this or any other subdiscipline. Additionally, the fact that we were not able to publish Replies to Comments on our paper in Ecology because of editorial interference and malfeasance by its former Editor-in-Chief, who allowed hostile referees to preemptively dismiss our Replies without actually reviewing them, suggests that we wield no such power.

> **R1. 30)** 2. Specific points
> 2.3-13: I hope that some of the points discussed above help explain the difference between your approach and ours, in terms other than “flaws” and “errors”.

Unfortunately, as we explained above, the terms "flaws" and "errors" apply to the estimates that you described in Clark et al. (2019). We are not unnecessarily or unfairly characterizing your estimators as flawed or incorrect. Instead, we proved that they were biased and thus incorrect, and then derived the proper unbiased estimators. Additionally, both independent reviewers agreed with us. Hence, your estimators and ours cannot both be right or coexist.

> **R1. 31)** 2.15-19: Here, I think we may have more agreement than you might think. Especially in the presence of observation error, it seems unlikely that precise estimates of selection and complementarity effects will be possible in individual systems. This is one of the main reasons that we wanted to generate a statistic that could compare multiple communities, as this would allow pooling information from a larger number of samples.

Again, our contention is that such comparisons are flawed and should not be made.

> **R1. 32)**
> 2.20-3.2: Again, I very much agree with you that M and dRY are highly context specific, and can’t be treated as fixed values. This is, again, a reason that we choose to treat them as samples from an overall distribution, even when measuring all N species in the community (and, I would argue, is a reason not to apply a correction that treats the population-level statistic as a fixed number). Note, also, that this is one of the reasons that we give for only applying this method in cases where natural communities are near experimental monocultures – i.e. if conditions between monocultures and mixtures vary, then the resulting partitions are not particularly useful.

This is problematic because the correction that you provide only accounts for differences in community size not composition. Hence, even if one attempts to control for differences in environmental conditions by stipulating that only communities that are in close geographical proximity should be compared (as you stated in your paper), differences in community composition will lead to aberrant results. This is because such comparisons implicitly assume that the dRY values of each species come from the same statistical population, which is

untrue since that statistical population will shift with community composition. Hence, this issue has nothing to do with whether one is sampling from a finite or an infinite population or whether the population parameter is assumed to be fixed or random.

> **R1. 33)** 3.3-9: I'm not convinced that this example falls within the scope of this article.

We disagree. Your paper presents an augmentation of the Loreau-Hector partitioning scheme and thus suffers from the same issues with nonlinearity. These issues need to be disclosed to practitioners. And above all, the nonlinearity issue renders the entire set of statistical tools you attempted to develop moot from the beginning.

More so, you *specifically justified* your extension/augmentation of the Loreau-Hector partitioning by making claims that were problematic, specifically that the choice of baselines would not affect the use of the Loreau-Hector partitioning. We needed to explain how, even if/when you are able to control for the problems related to the baseline, there would still be an issue with the estimation of selection and complementarity – the very focus of your paper. The arguments that we presented are new and have not been published elsewhere, and more importantly these arguments are tailored to the specific claims made in your paper regarding the validity of the Loreau-Hector method for augmentation – they are not a general critique or listing-off of all the problems with Loreau-Hector method.

There are currently several other papers being written/published by researchers using the Loreau-Hector method to analyze data where nonlinearity is likely to invalidate the results, yet we have no particular intention or interest in responding or commenting on each of these. Yours, however, was a *methods* paper – and in particular, *one that made very specific methodological claims* about the suitability of the Loreau-Hector method for being statistically extended or augmented in experimental situations. The record needs to be clear in order to allow researchers to be aware of the issues particular to your paper, and to prevent them from using other similar justifications in the future, even if the problems relating to unbiased estimators and replication were to be overcome.

> **R1. 34)** As I said above, I am very receptive to alternate null models that fall beyond the null model used by Loreau and Hector. But, designing such a metric wasn't the purpose of our methods paper. The alternate null model that you present in your Ecology paper is, in its own right, interesting – though again, it is only one of an infinite number of null models that could be used. When you simulate data that are drawn from your own null model and then test your null model, then it is hardly surprising that your null model performs better than the Loreau and Hector null. But, that doesn't at all mean that the null model is "better" than any other null model.

This is a mischaracterization of what we did in our Ecology paper. We did not simulate data from our model to show that our baseline was better than that of Loreau-Hector (doing so would be circular and we are not sure how you got that impression). Instead we used the data from the BIODEPTH experiment to calculate pairwise effects, which served as our

baseline. This allowed us to show that the difference between the degree of ecosystem functioning observed in 3+ species communities in the BIODEPTH experiment and that observed in the relevant pairwise mixtures became more negative as diversity increased. We never claimed that our approach "performed" better (how can one even compare different partitions that sum-up to the same observable quantity in order to assess their relative performance?). We merely showed that our approach and baseline allowed us to identify the novel and non-redundant effects of biodiversity on ecosystem functioning, which in this case happened to be negative at all diversity levels. While it is true that our null model is just one of many possible baselines, it does have several advantages including the fact that it does not suffer from the kinds of nonsensical paradoxes that emerge with the default baseline expectation of the Loreau-Hector partitioning scheme.

And we are well aware that "designing [an alternate null model] wasn't the purpose of [your] methods paper". Our purpose for bringing up the null model was due to the explicit claims in your paper that were used as a justification for statistically extending the Loreau-Hector method, specifically that the Loreau-Hector method will work with any null baseline or model. We pointed out that such a justification might work if one *only* wished to develop an estimator for the net biodiversity effect, but it cannot be used to justify developing statistics for selection or complementarity – the real goal of your paper. It needs to be clear to readers that your justifications based on the Loreau-Hector method working with any null model will not overcome the critical issue of the non-measurability (and hence non-estimability) of selection and complementarity.

> **R1. 35)** In particular, if we stick to the strict interpretation of the Loreau and Hector model that I discuss above (i.e. monocultures in 1/N of the area, vs. a mixture in the full area), then I am afraid that your critique doesn't hold anymore (since we aren't trying to pretend that we can interpolate between mixture and monoculture arrangements). Again – it's perfectly fine to think that this isn't a very useful null model, and I am inclined to agree that it isn't really what we want in many ecological applications. But, this seems like a curious place to make that argument again.

Although it is true that rescaling the area of the mixtures by N may in some cases avoid the inflation of the net biodiversity effect by keeping the densities in monocultures and mixtures identical, the nonlinear relationship between density and ecosystem functioning observed in monocultures will still lead to spurious estimates of complementarity and selection as we described in the text of our Comment [pgs. 19-20]. Hence, you are incorrect when you state:

> **"I am afraid that your critique doesn't hold anymore (since we aren't trying to pretend that we can interpolate between mixture and monoculture arrangements)".**

This statement explains precisely why we need to include this section of text in our Comment: in order to avoid having practitioners incorrectly believe that as long as the baseline that they use in the Loreau-Hector partitioning scheme avoids inflating the net biodiversity effect (e.g., by scaling the area of mixtures by the number of species being

planted in order to keep the densities equal between mixtures and monocultures), they will bypass the nonlinearity issues that yield spurious estimates of complementarity and selection. This is simply not true, as we explain in our Comment [pgs. 19-20]. If someone who is familiar enough with the Loreau-Hector partitioning scheme to publish an extension of it in the literature does not appreciate this fact, then most practitioners will also likely make the same mistake. Hence, it is *absolutely* imperative that this section of the text be included in our Comment.

> **R1. 36)**
> 4.8-14: I would suggest including the full formula for your expected values and covariance terms here, given that you use a wide range of similar looking notation later in the paper.

Agreed. We have clarified the notation throughout the text.

> **R1. 37)** 4.18: Can you explain what you mean by "scaled" in this case? It should be clear to people who use the metrics a lot, but to other readers this may be a bit mystifying.

We defined it several times early on, specifically it is the per species complementarity or selection effect (the estimated full effect size divided by the full community size N). We have expanded the introductory paragraphs and clarified this further in the text [p4: lines 23-24 to p5: line 1] .

> **R1. 38)** 4.12-15: I hope that the discussion above will help you understand the justifications a bit better. That said, it is certainly true that our "correction" for the alternate formulation for the complementarity effect term that forces the population and sample level statistics to converge is ad-hoc – but, we state that very explicitly in the paper. All other terms arise directly from the distributional assumptions in the discussion above. If you believe that the rest of the derivation that we use is arbitrary or incorrect, could you please point to the specific lines in the derivation that you think are unsupported, and explain why?

Admitting that your correction was *ad hoc* (i.e., improvised and unjustified) in the paper does not excuse it, nor does it shield you against mathematically motivated criticisms. If one is claims to have produced unbiased estimators, one still has to prove they are unbiased. We described the issue with the correction explicitly in our Comment:

> **"Instead of doing this, Clark et al. (2019) took the unbiased estimators for an infinite population and inserted, without any mathematical justification, arbitrary correction factors that would allow the expressions for the property of interest to converge in value to that observed in the N-species experimental system as the sample size n approached N. In fact, Clark et al. (2019) inserted not one but two different ad hoc correction factors into two different expressions: one to allow the estimator for the complementarity effect to converge to the true population value (Eq. 3c in Clark et al.**

2019), and another to allow the sample net biodiversity effect (sum of the estimators for complementarity and selection) to converge to its true population value (Appendix B.II in Clark et al. 2019)."

R1. 39) 5.18-20: Again, especially for the equation in the appendix, I think that it would be polite to point out that we repeatedly told users that we did not suggest that they use this correction, and that only added it under protest to allow calculation of selection and complementarity effects that matched the exact observed net biodiversity effect. And, as discussed above, I freely admit that your method is a much better way to derive a partition that exactly matches the net biodiversity effect, but also caution that this wasn't the main goal of our analyses.

6.eq6: Note that you use the sigma-hat_{x,y}* symbol here, but don't define it until pg. 9. I'd suggest either moving things around in your paper, or given the definition earlier. Or is this a typo here?

This was a typo, which we have fixed. As to earlier comment about the correction for net biodiversity effect, see our response to Reviewer 1, R1.41 and R1.46 below.

R1. 40) 6.4-7: Yes, as discussed above, this was our intention. This might be a good place for you to explain what is gained by using such a correction, and what is lost when one instead chooses a finite sample approximation as you do. Obviously, the rest of your paper can still focus on cases where the finite sample approximation performs better (e.g. especially cases where the net biodiversity effect needs to be calculated from the partition).

This section merely highlighted the problematic nature of adding an arbitrary correction to the estimates for an infinite population. We are not so much advocating for the use of estimates based on finite vs. infinite populations, but against the use of *ad hoc* and unnecessary correction factors. As we mentioned in our Comment, you combined elements from unbiased estimators for both finite and infinite populations to create your estimators of selection and complementarity for no justifiable reason. Your estimators are thus both incorrect (biased) and incoherent.

R1. 41) 6.7-9: As noted in the general text above, I would take it as a personal kindness if you could be very explicit in noting that we never presented your Eq. (6) in our paper, specifically because we did not intent our partition to provide an unbiased estimate of the net biodiversity effect. Especially since we only explicitly say this in the appendix of our paper, I think it would be very helpful to prevent confusion among readers who might otherwise try to incorrectly apply our methods.

We have included text explaining elsewhere that the net biodiversity effect was presented in the Appendix as an additional result [p7: 9]. However, the corrected net biodiversity effect was still reported as an actual *published result*. Furthermore, it was represented by the

authors as an "augmented correction" that could be used in special "cases where it is vital that the sum of the selection effects and complementarity effects equals the change in relative yield". No "augmented correction" would have been necessary if the correct estimators for selection/complementarity had been used to calculate the net biodiversity effect. If Reviewer 1 wishes to elaborate further on the ultimate purpose/motives behind this net biodiversity statistic, they can do so themselves in their Reply.

> **R1. 42)** 6.13-17: As discussed above, I hope you now understand that, although our statistics are not unbiased estimators of the quantities that you thought we were trying to calculate, they are unbiased relative to sample size – i.e. regardless of whether we sample n or N individuals, we find the same expected value of the statistics, so long as M and dRY for both the n and N pools are drawn from the same distribution.

We disagree. You are using standard statistical jargon about unbiasedness with a specific definition to refer to an *ad hoc* correction whose sole objective appears to be to correct for differences in community size. This is simply incorrect and should have been caught by the reviewers. True unbiased estimators will have the same property you desire when computing their expected value from multiple replicate simulations. Again, for a sample estimate to be unbiased, the expression for its expected value must match that of the corresponding population parameter. However, the expected values of neither of your sample estimates matches the corresponding population parameters so your sample estimates are *not unbiased*.

> **R1. 43)**
> 7.3-6: Yikes! Can you please highlight the exact text that makes you think we said this? My intention, in both cases, was to say that the properties of the overlying distribution could be derived from a finite sample using this correction (i.e. identical to what you say in 4.23-5.1 in your paper). Going through the text, I still don't understand exactly what about these sentences makes you think we were claiming that this allowed an estimate of the finite sample statistic for N individuals based on a finite sample of n individuals. Can you highlight this unclarity and – ideally – make it clear that this isn't what we intended to communicate?

See Response to Reviewer 1, (R1.45) below where we cite examples. Clark et al. did not use the terminology of finite/infinite population in their paper (likely because they were unaware of the issues associated with these terms, as is clear from their referee report), but their estimators for both selection and complementarity were nevertheless treated as a statistic for a *finite* population as seen in the (otherwise confusing) Figure 3 (see also figure caption where this is confirmed).

This is part of the incoherence of the sample estimates presented in Clark et al. (2019): it's a hodge-podge of statistics where the sample complementarity and net biodiversity effects are estimated using (incorrect) statistics for a finite population, while the sample selection effect is estimated using statistics for an infinite population.

> **R1. 44)** 7.6-12: This is a place where I believe your text to be factually incorrect. We say that our statistics are unbiased with respect to sample and community size – meaning that they provide consistent estimates of the quantities shown in our derivations. Note that this is subtly different from the claim that you say that we made (i.e. that we calculated unbiased estimates of the sample-level statistics for N individuals). The fact that our estimates are unbiased with respect to sample size is relatively hard to dispute. That is, our figures 4b and d demonstrate that we can take information about, e.g., samples of 2/8 species in a community, and calculate an estimate of the statistic that is centred around what we would have calculated had we measured all 8 species. That is, with incomplete information, we can reconstruct the expected value of the state that we would have gotten with full information. That is the definition of an unbiased statistic – regardless of whether or not it an unbiased estimate of the quantity that you would like to approximate.

This is <u>not</u> the definition of an unbiased statistic. *We* derived the correct unbiased statistics for selection and complementarity, whereas you did not. You simply attempted to gauge whether your made-up estimators were unbiased impressionistically by visual inspection of a plot. We derived the correct unbiased statistic for the exact same quantities you claimed yours were estimating – our statistics were not meant for different quantities.

And we are still not sure how to parse the meaning of this phrase:

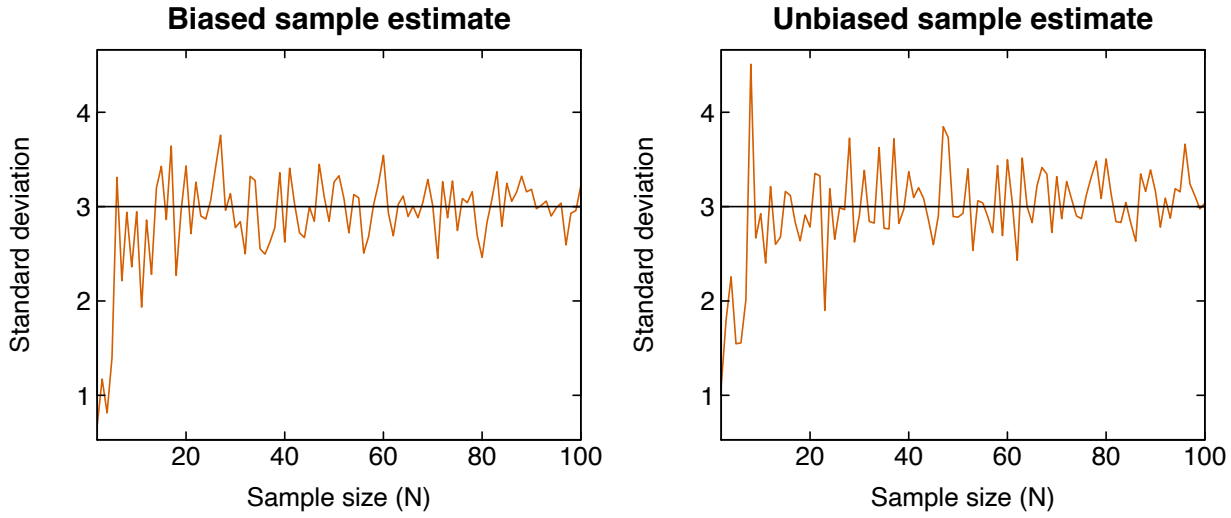
> **"our statistics are unbiased with *<u>respect to sample size and community size</u>* -- meaning that they provide *<u>consistent estimates</u> of the quantities shown in our derivations*" [emphasis added].**

No part of this phrase actually makes sense or relates to what "unbiased" means in statistics. What does it mean to say something is "unbiased with respect to sample size and community size"? And what does providing "*consistent estimates* of the quantities" mean? It's unlikely given the phrasing, but is Reviewer 1 trying to say their statistic is a 'consistent estimator' in the technical sense?

If so, Reviewer 1 seems to be confusing statistical consistency and statistical bias, as helpfully pointed out by Reviewer 2. As sample size increases, both (i) a consistent but biased estimator and (ii) a consistent and unbiased estimator will converge onto the true population parameter value (that is what you designed your correction to do). However, when sample size is relatively small, the unbiased and consistent estimator will be centered on the true population parameter value whereas the consistent but biased estimator will not. The fact that your estimate *appears* to be centered around the true population parameter value in your figures is that they were plotted incorrectly as pointed out in our Comment.

More importantly, one cannot simply assert that an estimator is unbiased by visually inspecting a plot and checking that the fluctuations of the estimate *appear* to be centered

around the population parameter value, as you claimed. That is a shockingly incorrect statement and is *not* how unbiased estimators are defined. As mentioned above, the definition of an unbiased estimator is that the expected value of the sample estimate is equal to the population parameter. To demonstrate this issue, here is an example of the distribution of both biased and unbiased sample estimates of the population standard deviation (set at 3) with increasing sample size N:



If we had not labeled these panels, one would be hard-pressed to identify which was biased and which was unbiased. This shows that visual inspection of the distribution of sample estimate fluctuations around the population parameter value is not sufficient to claim that an estimator is unbiased. To demonstrate that an estimate is unbiased, one has to take its expected value and show that it is equal to the population parameter value. Otherwise, quantities that have relatively small biases can appear to be unbiased.

Furthermore, Reviewer 1 may have incorrectly believed their statistic was fluctuating around the true population values because, as we pointed out in our manuscript, <u>they plotted their estimators incorrectly</u>. This is another good example of why visually and impressionistically assessing whether an estimator is unbiased cannot be a substitute for mathematically demonstrating it.

> **R1. 45)** 9.1-3: Again, this really wasn't our intention. I hope that you understand this now.

The passage in our manuscript that you are referring to is completely correct. Whatever your intentions, what you *stated* in your paper is that you were using the estimator for selection (covariance) for an infinite population (your Eq. 3a) also as an estimator when sampling from a finite population; at the same time you <u>*modified*</u> the infinite complementarity statistic so as to use it with a finite population (left column, 5th page/pg. 2145 in Clark et al. 2019).

There really is no question about what you did. Elsewhere in your paper you confirmed that you always intended to use this infinite population statistic for selection/covariance as an

estimator when sampling from a finite population (see for e.g., Appendix, page 5, bottom two paragraphs).

Whatever your real intentions, the original Clark et al. paper will still be interpreted by other readers in the way that we did; like us, other readers will be just as confused by the incorrect and *ad hoc* definition of 'unbiasedness' used to justify these estimators with *ad hoc* correction factors.

> **R1. 46)** 9.12-13: Yes, agreed – I hope that you now understand the difference between these in terms of what they can be used for. Again, this might be a good place for you to discuss these different purposes.
>
> 10.5-9: Again, this is a bit of a mistake. Even with the ad-hoc correction that we provide, the net biodiversity effects cannot be calculated from the partition, nor was this ever the intention, as discussed at length above. See, for example, the introductory analyses that we provide in our R package – these show these deviations very clearly. The only case where the net biodiversity effect exactly matches that for the raw data is when both corrections are applied simultaneously, which is something that we caution users not to do. That said, I think it is perfectly helpful to point out that our correction does not yield an unbiased estimate of the net biodiversity effect. Please just make sure not to say that the corrected Eq. 3c provides an unbiased estimate.

This is not a mistake on our part. The made-up correction you applied to the net biodiversity effect, was done, as you pointed out, for the same reason, in the same manner and by using the same claims that you made when you created your "unbiased" estimates for selection and complementarity. Specifically, you inserted a correction for the net biodiversity effect so it will "allow an augmentation similar to Eq (3c) [i.e., complementarity effect] in the main text" (Clark et al, Appendix, pg. 6). The purpose for this net effect estimator correction, as with the others, was so that it would allow convergence to the true value as n = N. But as we stated in our manuscript, with the correct statistics for complementarity and selection:

> **"[there will be] no need to arbitrarily insert an ad hoc, mathematically unjustified correction factor into the expression [for net biodiversity effect] in order to ensure that the sum itself will be unbiased – or worse, simply so that it converges to the true value with increasing sample size".**

It's true the authors cautioned against using their new corrected formula for the net biodiversity effect, but they still clearly said it would be valid "in cases where it is vital that the sum of the selection effects and complementarity effects equals the change in relative yield", which is not true, since it is an expression that does not appear to have any purpose. It's a correction factor created as a consequence of not being aware of the earlier mistakes made when creating the estimators for selection/complementarity. It's also irrelevant that the authors were not really interested in estimating the net biodiversity effect, since, as we pointed out above, obtaining a correct net biodiversity effect estimator is a direct

consequence of having found the correct (unbiased) estimator for selection and complementarity in the first place.

In our manuscript we responded to the claims that were stated and published in the original Clark et al. (2019) paper, but if Reviewer 1 still feels that their true but cryptic intentions need to be clarified, they can do so when they write their Response.

> **R1. 47)** 10.11-18: I think this is a case that reads as unnecessarily snide. Is there a purpose to this phrasing? If so, what is it? If not, then why not change it? Especially since your comment is largely based on a misinterpretation of the purpose of the paper, I think this would be a good place to try to kindly explain the difference between the kinds of results one can get from the two types of partitions, rather than to double-down on the misinterpretation. This would be a wonderful service to future readers of both papers.

There is *nothing* "snide" in this passage – it was meant as a matter-of-fact reporting of how the errors for each of the statistics presented by Clark *et al*. fit together. Regardless, we have now removed the text that points out how the combination of the incorrect selection effect and the incorrect complementarity effect will allow one inadvertently, if taken together, to estimate complementarity (two rights make a wrong), as it is redundant at this point.

> **R1. 48)** 10.20-23: Again, I think this is a place where it would behove you to write a bit more charitably. The fact that the metrics do not solve the problem that you would like to solve does not mean that they are unusable for other purposes, nor does it imply that your methods are more usable for all purposes.

That is not what we claimed. We are not claiming ours "are more usable for all purposes", we are claiming that they are the proper or *correct* unbiased estimators for the specific purposes clearly laid out in Clark et al. (2019) – that is, for providing unbiased estimates of selection/complementarity – whereas the biased statistics you provided in this same paper *cannot* be used for these purposes. What is more, we also showed that neither our properly derived unbiased statistics, nor the *ad hoc* and biased estimators that you described can be used for the purposes described in Clark et al. (2019) for a variety of methodological reasons.

> **R1. 49)** 11.4-6: Yes, this is true, and is something that we try to make very clear in the paper – i.e. that we provide unbiased, but noisy, estimates of the statistic.

You did *not* provide "unbiased" statistics for any of the quantities you claimed you did – and the phrase "noisy, estimates of the statistic" is redundant here since estimates have to be noisy to some degree depending on variation in the data set. The problem with your statistics is not the degree of 'noise' or the variance of your estimates around the true population values, but the fact that the expected value of the estimates are not equal to the true

population values. Furthermore, it is precisely this 'noise' that will render any unbiased estimator on its own (without replication) as essentially worthless, as is the case with the Clark et al. (2019) method, where sample replication is not really possible/feasible.

> **R1. 50)** 11.11-14: This is a case where I think your method would benefit from some more applications of real data. Our statement (which applies to our metrics, though not necessarily to your metrics) was merely based on the observation that, based on the observed degree of variability in real world data, we generally found that at sample sizes of n = N/2, the distribution of potential statistics had shrunk sufficiently that the correct sign of the statistic could be recovered at $p < 0.05$. That is, a single draw of n = N species was highly likely to yield an estimate for the effect at the N level that gave the correct "direction" for selection and complementarity effects. And, all else being equal, comparisons across multiple communities were highly likely to retain the correct rank order. These are the main characteristics of interest in most studies that have applied these metrics in the past, so I don't quite understand why you think that this statement is poorly supported. Can you be more specific about why you feel the variability in statistics that we observe in estimates around n = 2 is too large to be useful? Can you support this concern with real data?

We were specifically referring to the large variability in the numerical estimates of complementarity and selection that emerged when taking repeated samples. By now suggesting that you are only interested in whether the sign of the sample estimate of selection/complementarity effect is consistent with the sign of the true population parameter, you have shifted the goal posts. Nowhere in your paper is it explained that you consider your sample estimate adequate as long as its sign matches that of the population parameter! This is particularly problematic since these sample estimates are numerical and thus suggest that quantitative comparisons can be made.

> **R1. 51)**
> 11.17-20: I'm afraid you misunderstood the figures here. As noted above, the statements we make about accuracy are really based on variability expected for single observations. The homogeneous replication test was really only to compare effects of observation error vs. sampling uncertainty. As discussed above, observation error almost always trumps effects of sampling uncertainty for all n and N, which I think is a major unsolved problem.
>
> 11.24-12.3: Ah! Now I understand why you were concerned about the effects of heterogeneity. No, in this sentence, we are not advocating that, e.g. monoculture data be taken from one site, and used to calculate statistics at other sites. Rather, we are advocating comparisons such as those that we show in Fig. 5. That is, the meta-analysis should compare trends in selection and complementarity across different sites as a function of some covariate (e.g. diversity), but the site-level effects should always be calculated using data from nearby monocultures. Otherwise, of course, the statistics are not very useful.

Hopefully this should make it a bit clearer to you why we care about finding a statistic where there is no a priori trend with community size N? That is, if we thing, e.g. that diversity should correlate negatively with selection effects, and we test this by comparing selection and complementarity across many different communities in different sites (within which the statistics are calculated using available monoculture data at or near that site), then it is of course very important that covariance doesn't increase with (N-1)/N. Otherwise, we would fail to meet the basic assumptions of ordinary least squares, which requires that all observations be drawn from the same underlying distribution. And, worse, we would observe trends even if the only difference were sample size, not underlying ecology.

In any case, I hope that you can rephrase this section. The idea that we would combine information on monocultures from multiple different sites is wrong.

11.6-12: Again, I think you are misunderstanding the standard deviations on our simulations. Note, as said in the main text, these are standard DEVIATIONS, not standard errors. So, they describe variability at the level of individual simulations, not variability in the mean across all 20,000 simulations. E.g. even for 10 replicates, we have very low uncertainty in statistic values in Fig. 4. and Fig. S1, especially for n >= N/2. Are you simply mistaken? Or can you explain why you think that these standard deviations are incorrect? And, can you support that claim with real data? Note, all the data we used for our analyses are available through DOI's in our paper, so feel free to use the Jena data for your analyses if you'd like.

We understand the difference between standard deviation and standard errors, and the fact that you used the former. The issue is not with the use of the standard deviation *per se*, but rather the fact that any particular random simulation will generate potentially large departures from the true population parameter value of interest. Taking the expected value as you did in your figures will yield reasonable predictions because the bias in your estimators is relatively small. However, the point of this section was to show that averaging results from many replicates (sometimes as few as 100 but sometimes as many as 20,000) is necessary to obtain a numerical approximation of the population parameter value of interest (even if your approximation remains inaccurate since your estimators are biased).

R1. 52)

13.22-14.5: Again, I'm a bit confused by your comments here. The lines in these figures show the cross-simulation mean +/- standard deviation. From your comments, I think maybe you incorrectly read these results as though they were for standard error of the mean? But in any case, the intervals show standard deviation at the level of individual simulations – so, e.g., about 68% of simulated values from the 20,000 iterations fell somewhere within the reported interval.

If you made a mistake, could you please correct the text to fix it? If not, can you please explain in more detail why you disagree with how the intervals were calculated? Again, doing so with real data would be idea.

We are referring to the results in our Figure 1, not those in your paper. Here, our results show the wide range of values that the estimates can take across the different simulations. We are not assuming that the error bands represent standard errors. The point we are trying to make is that individual simulations will yield very different results.

R1. 53)
13.14-20: Again, I think that this suggest that you didn't quite understand what the simulations were showing. E.g. for the high diversity plots, even for a single sample in Fig. 4, at n=N/2 about 85% of simulations returned the correct sign for the statistics, and by n = N *3/4, it was closer to about 99%. Moreover, even for cases such as the low-diversity plots, where simulations based on single measurements had a high probability of overlapping zero, relatively low replication (e.g. analyses that span about 10 plots or sites) again provide estimates that differ from zero at p = 0.2-0.01. This, for example, is how we are able to detect significant trends in Fig. 5, even though the estimates for individual plots are highly variable.

Can you explain what aspects of these statistics are undesirable? In general, this performance aligns with, or is less problematic than, effects of observation error on analyses. Again, I'm all for tackling the observation error problem, but I don't see how the analyses we show are inconsistent with these statistics being useful for hypothesis testing.

You are once again redefining success, shifting the goal posts and lowering the bar in a way that was never described in your paper! Nowhere in the main text or appendices did you ever explain that you considered a simulation successful if the sample estimate its produced matched the sign, not the magnitude, of the population parameter of interest. We were simply making the accurate point that the numerical value generated by the estimator in any particular simulation can be different from the population parameter value by orders of magnitude. But even a reasonable estimate of the sign will still require a level of replication that will simply be unavailable to experimenters, particularly in datasets with sizable variation or small magnitudes of values!

R1. 54) 13.25-26: Can you explain what kind of test you apply to determine that individual replicates depart from the mean trend "significantly"? I'm afraid I don't quite understand the idea of individual replicates significantly departing from the distribution that they are drawn from.

We were simply explaining that the estimates from individual simulations can significantly depart from their expected values/population parameter values. We have replaced the term

significantly (which can suggest some kind of statistical test) with "by orders of magnitude" [p16: line 14].

> **R1. 55)** Similarly, what do the intervals in Fig. 1 show? If they show mean +/- standard deviation, then I would actually say that your results are quite consistent with ours – that is, that even though they are noisy, even individual instances are likely to be useful for hypothesis testing. From the text, though, it seems like these might be minima/maxima? If that's the case, then I think this provides even better evidence that these would be useful statistics for testing hypotheses.

As explained in the text, the bands in the figure correspond to the min/max values observed across the simulations. These results can in no way be interpreted as strengthening your claims! The figure shows that individual simulations yield orders-of-magnitude errors in the estimate of the population parameter values for selection and complementarity.

> **R1. 56)** As a general note, it is usually discouraged to show min/max ranges in figures, rather than, e.g. a 95% confidence interval, since the range is so strongly determined by the number of iterations tested. Apologies if I am misunderstanding something.

We are fully aware that the min/max/range are biased estimates of the spread of a distribution (with the min decreasing and the max [and thus the range] increasing with sample size). The point of using the min/max was not to provide an unbiased estimate of the true population min/max parameter but to show the *extreme variability* of the sample estimates of complementarity and selection across replicate simulations. This extreme variability, as we explained in the text, is what one risks measuring when one is unable to either conduct proper replication or sample from the full finite population for each sample draw. Both features – replication and sampling from the full population – will be unavailable to an experimenter when applying the experimental approach Clark et al. (2019) suggest.

Since the number of replicate simulations does not change across the x-axis or the panels of Figure 1, the comparisons are perfectly legitimate. Additionally, because the min/max of selection and complementarity will remain the same or become more extreme with increasing sample size (i.e., they can never become less extreme with additional replicate simulations), the fact that their range is so large with even 10 replicate simulations shows how unreliable the estimates selection and complementarity will be even when using our correct unbiased estimators. Finally, we are interested in showing the variability across individual simulations so the 95% confidence bands for the mean that you suggested would be inappropriate here.

> **R1. 57)** 14.10-16.4: In general, I think that this section makes a very good point. But, I think it's a bit unfair to attack our article with this point, as we make it too. In our analyses, we use only data from sites that are within a few hundred meters of the monoculture plots, and for which conditions are similar to those in the Jena plots.

As we explained in our Comment, geographical proximity is not sufficient to make these comparisons possible. Even if the environmental conditions were identical, differences in community composition would be sufficient to make comparisons between nearby fields harboring slightly different communities impossible. Hence, the point we were making is fair.

More troubling, however, is that the issue of incommensurability that we brought up regarding the ability to make comparisons across different communities as suggested by Clark et al. still appears to elude Reviewer 1: different communities *cannot* be treated as though they were different samples taken from the same larger regional species pool/community. That is, they cannot be treated as different statistical samples from some still larger statistical superpopulation. This is because the relative yields of a species are a *structural property* of each specific experimental community as a whole, not just of a given species itself (unlike monoculture yields).

> **R1. 58)** We furthermore repeatedly try to remind practitioners that their results will only ever be as good as their samples are random. And so, I completely agree that non-random sampling of communities and monocultures (e.g. omitting monocultures that all share a particular property, or sampling monocultures under one condition and mixtures under another) will produce nonsensical results. But, we state this many times in the text. I would politely ask that you try to point this out somewhere in your article – my sense from the current text is that it sound like you are alluding that we were advocating applying our method to highly non-random samples.
>
> One other point – You say repeatedly that cases that stray from random sampling will produce results that are "meaningless". Again, I think that this is a statement that really needs to be tested against empirical data. While it's true that non-random sampling violates assumptions behind the partition, it is not true that all non-random samples will provide poor estimators. In fact, it seems likely that one can get away with samples that are quite non-random before they start influencing the estimates. This is a general problem for all empirical attempts to apply statistics to data – we can never actually fulfil the theoretical assumptions of the tests that we use, so instead we just have to hope that the test are robust to these assumption violations.
>
> Our analysis of empirical data shows that even with somewhat non-random sampling, and sites that are somewhat different from the Jena plots, we are able to detect trends that are consistent with theoretical assumptions, and that match patterns from within the Jena experiment. Jointly, these provide evidence that, at least under some circumstances, it should be possible to apply these methods empirically.
>
> Can you offer some counter-examples with real world data? Or with simulations that are meant to mimic real world data? Especially valuable would be a sensitivity analysis that helped show how big a violation of assumptions we can get away with before results become "meaningless".

The issue is not with the statistical robustness of the results but with comparing results from experiments that are fundamentally incomparable. The simple scenario outlined in our Comment shows that different community compositions can systematically shift the dRY values. The dRY values observed in an experiment are not the properties of individual species alone, that is, they are not properties that can be assigned to individual species in the way that certain types of properties or attributes inhere in individual entities (for example, the way gold can be characterized by a particular density, or carbon by a given atomic weight). Rather, what dRYs represent are in large part *structural properties* of the experimental community that the species are a part of. This is why different BEF experiments cannot be considered as though they were samples drawn from the same larger statistical population unless the population is literally made-up of all possible combinations of community compositions, which is clearly nonsensical.

> **R1. 59)** 16.5: As noted above, I'm not sure that this is the right place for this second set of points. It doesn't seem very related to the primary point of the paper anymore – i.e. discussion of sample-vs-population partitions of the Loreau and Hector framework. My advice would be to either put this in an appendix, or to publish it as a separate manuscript.

As we explained, this section is almost more important than the others because it explains that even if one used the correct unbiased estimators that we derived, one should avoid using them unless the hidden linearity assumption in the Loreau-Hector partitioning scheme can be verified. And in particular, it addresses the *very specific arguments* that were made in Clark et al. (2019) in order to justify statistically extending the Loreau-Hector partitioning.

> **R1. 60)**
> 17.2-4: This seems a bit strange to me. Can you explain? Obviously, one can replace the null model in the Loreau and Hector partition, and still separate dRY and M into components that can be explained via their covariance, and components that cannot.

We explained this in our Comment starting in the following sentence:

> **"The problem of partitioning is deeper than just the issue with the baseline assumption. It is true that the artificial inflation of the net biodiversity effect that occurs when the relationship between density and ecosystem functioning is nonlinear in monocultures (Fig. 2) can be avoided as long as one uses the true ecosystem contributions of each species at the midpoint of their carrying capacities as a null baseline in the Loreau-Hector partitioning scheme instead of the (default) midpoint of the monocultures (Fig. 2). However, this means that the net biodiversity effect cannot be estimated at the surface level of observed ecosystem changes, as has been done for two decades. Accurately estimating the net biodiversity effect under such nonlinearity would instead require a herculean effort: the measurement of not just monoculture and relative yields but also the density-ecosystem functioning relationship of each species in monoculture to properly determine the true null**

baseline (i.e., the ecosystem contribution of each species at the midpoint of the carrying capacities)."

R1. 61)

Again, I agree that attempts to build mechanistic interpretations of the Loreau and Hector partition that go beyond the simple null model that I describe above are, to say the least, dubious. But, I am not sure that anything that you say here can be take as evidence that the null model is wrong. Rather, it's just a description of differences between monocultures and mixtures. I'm perfectly prepared for you to argue that the null model is uninteresting, but that's a different argument.

The problem is not that anyone is trying to build a mechanistic interpretation of the Loreau-Hector partition. We simply explained the issues with the default baseline in the Loreau-Hector partitioning scheme that arise due to effects of nonlinearity in monocultures (Fig. 3). This has *nothing whatsoever* to do with the mechanisms that structure ecological communities.

R1. 62)

20.7: This seem a lot like a restating of your Ecology paper. Is this correct, or am I missing something? If it is largely a restatement, may I suggest just citing the paper and giving a brief description of what you are trying to convey? Sorry if I'm missing something that is conceptually new here.

This is not a restating of our original paper at all. It is an attempt to address the very *specific methodological claims and justifications* made in Clark et al. (2019), specifically that the Loreau-Hector partitioning can be statistically extended because the method can made to work for any baseline. As we explained, under nonlinearity, this is technically true if one *only* wanted to measure/estimate the *net biodiversity effect* alone – and even then, this would be possible only with tremendous effort and changes to the standard BEF experimental design. However, we also needed to point out to the reader that even if it were possible to estimate the net biodiversity effect by statistically extending the Loreau-Hector approach, the partitioning of this net effect in order to estimate selection and complementarity still would be impossible because of the fundamental non-measurability of these two component effects under nonlinearity – and as Reviewer 1 has reminded us several times, it was these effects and not the net biodiversity effects that are the real focus of interest in the Clark et al. paper. Our completely *new and previously unpublished arguments* and discussion explain how Clark et al., while technically correct regarding the baseline for the Loreau-Hector null model, still overlooked the non-measurability and thus non-estimability of the selection and complementarity component of the net biodiversity effect. In fact, the arguments we presented demonstrate clearly for the first time how all attempts to partition the net biodiversity effect into *any* smaller components (not just selection or complementarity) will always fail under nonlinearity.

It is not enough to simply cite the original paper, we also need to explain the subtle difference between the ability to measure the net biodiversity effect and the inability to do the same for selection/complementarity, since this distinction has been widely misunderstood by the BEF community. This elucidation of the issues was necessary in order to address Clark et al.'s *specific attempts to justify* their extension of Loreau-Hector; and since their problematic method and justifications were *published* in this journal (a methods journal), it is incumbent on us to explain why the justification does not hold. More so, it is important that no one attempt to statistically extend the Loreau-Hector method in a similar manner, or by using the same justifications, which will be likely if Clark et al.'s mistaken rationalizations go unchallenged in the literature.

Reviewer 2

R2.1)
I read with great interest Pillai and Gouhier's Forum response to Clark et al.'s "How to estimate complementarity and selection effects from an incomplete sample of species". After careful reading of both the response and the original manuscript, it is clear the authors have shown that Clark et al.'s approach is flawed both in terms of conceptualization and implementation.

We are glad Reviewer 2 agreed that the conceptual and mathematical mistakes that we uncovered in Clark et al. (2019) were real.

R2.2) In terms of the response itself, there are number of improvements to be made to enhance the readability, ease of comprehension and grounding of assertions in the literature. First, the key concept of bias needs to be defined early on and the authors should show the bias, not the estimator, at the end of their derivations. With such an approach would make it obvious for the reader why Clark et al.'s estimators are flawed and need to be revised.

We agree and have added a mathematically explicit and formal definition of statistical bias early in the text, as it now appears the concept may not have been as self-evident or as well understood by ecologists as we had originally assumed [Eqs. 1-2 on p5: 5-11]. We also now explicitly show/quantify the biases in the Clark et al. (2019) estimators [see Eqs. 21-22 on p11: 6-11].

R2.3) Furthermore, it might be useful to also talk about and define the statistical concept of consistency, which Clark et al. seem to be confusing with bias.

We agree. We have mentioned statistical consistency and explained that it should not be confused with bias [p8: 4-10]. .

R2.4) Another element that requires improvement is the introduction and definitions of parameters throughout the text. It is very easy to become confused due to the avalanche of symbols that, in many cases, are not defined before being used. I think it

would make a great deal of sense to quickly set out the notational rules for symbols early on for population parameters, estimators, etc. I would suggest that this could be done by creating a symbol/parameter table.

We agree and have revised the entire text by making sure definitions of each parameter were clearly stated before it is used [for e.g., see p4: 10-24, p5 & p9-p10]..

R2.5) I would also say that the derivations themselves are not always presented in the most straightforward manner as the authors skip steps that they then call out afterward (page 5, lines 19-21, for example). I understand the desire to keep derivations brief and to the point, but I find it better to lay out certain steps when it involves changes of indices of summations and such, as they can be missed easily on first reading (e.g. equations 10 to 13).

We tried to keep the derivations as brief as possible in order to avoid significantly lengthening the text. We have agreed to add intermediate steps to the example referenced above and elsewhere when they are necessary to improve readability.

R2.6) With regards to link to the broader literature, the authors commonly do not cite where certain statistical results come from and provide no derivation. For example, on page 13, they say "This is simply a consequence of the Law of Large Numbers and of how unbiased estimators are supposed to work". Perhaps the authors are assuming a great deal of statistical background from their readers, but I would posit many ecologists may need references for derivations on estimators and the like. Wackerly et al. "Mathematical Statistics with Applications" chapters eight and nine have a number of results and theorems that could be cited within the text. In particular, "the sum of any two unbiased estimators…" on page 11 is theorem 9.2 in the sixth edition of Wackerly et al.

We agree and have added a reference [p15: 10]..

R2.7) I do think that the authors should be clearer that the 'net biodiversity effect' does not need to be overestimated (not all density-ecosystem property relationships are expected to be concave). The fact of that convex functions also fall prey to the non-identifiability under non-linearity is clearly demonstrated in their univariate example and we would expect similar situations when the functions are neither concave nor convex.

We agree. Figure 2 and the accompanying text referred to an overestimation of the net biodiversity effect because the law of constant final yield leads to a concave relationship between density and ecosystem functioning in the kinds of plant communities considered by Clark et al. (2019). However, this relationship could certainly be convex in other systems and thus cause the net biodiversity effect to be an underestimate. We have clarified this in the Figure 2 caption and the accompanying text [p19: 9-12].

R2.8) My last major comment is that the manuscript is quite repetitive. Based on the numerous references to misunderstandings of their previous manuscript, it seems like the authors want to emphasize as much as possible that their results are highly general and not limited to a specific case. Some repetition is necessary, but it can bog down the text. I would recommend streamlining the univariate and multivariate cases by removing about a paragraph of text from each.

We agree and have streamlined this section of the text, including cutting repetitive text and a paragraph in the subsection mentioned. We have also tried to motivate the section better in the beginning by framing it more explicitly in terms of the distinction between the measurability and non-measurability of the different biodiversity effects. Hopefully this will help reduce the overall repetitiveness of this section.

R2.9) Page 4, lines 8-9: It should be noted somewhere that Clark et al. use different symbols for the same quantities (i.e. N is used instead of Q, and n is used instead of N).

Agreed. We have added text explaining this early in the introductory paragraphs of the manuscript [p3: 23-24 to p4: 1-3].

R2.10) Page 4, line 20: should be Cov[x_i,y_i]

Agreed and fixed.

R2.11) Page 6, equations 4 and 6: you introduce the * notation for the covariance, but do not define it here. I know that this is supposed to be a finite sample covariance, but you mention that Clark et al. simply use the infinite population covariance, so the * is unnecessary.

Agreed, this was a typo we caught after submission and fixed.

R2.12) Page 7, line 19: |Ω| refers to the cardinality of the set. I think that should be spelled out in words in this case.

Agreed. We have added text to clarify [p9: 6-7].

R2.13) Page 7, line 21: “If absent from the sample…” What is absent? A particular species? Not clear from context.

We have clarified the text [p9: 12-13]).

R2.14) Page 8, equation 12: x¬_bar and y_bar are introduced, but only defined on line 13-14. Try to make it clearer what they represent, i.e. the true population means.

Agreed and fixed. We have significantly cleaned up the exposition of this section in general.

> **R2.15)** Page 10, line 2: First time that Φ is introduced and is never defined throughout the text.

Agreed and fixed. We now properly define both \phi and \Delta \phi [p12: 3-6].

> **R2.16)** Page 10, lines 14-18: There is no insight gained in emphasizing that the errors nearly cancel themselves out for one estimator. Please cut.

Agreed and fixed. We have removed it.

> **R2.17)** Page 14, lines 14-22: Is this block quote necessary? Clark et al.'s statement is already summarized and paraphrased before being quoted.

Unfortunately, this is such an extraordinary statement that it is almost impossible to believe it was made without the formal proof/verbatim quote provided. And as we see with Reviewer 1's referee comments, even with the verbatim quote the authors still attempted to spin it as something it was not (for example, see responses to Reviewer 1: [R1. 57] and [R1. 58]). That is why we believe the blockquote should remain.

> **R2.18)** Page 17, lines 18-23: I am not sure what the added value is in referencing how Loreau and Hector did not address a criticism in a different paper. I would recommend cutting it.

We respectfully disagree. It is important to show that the nonlinearity issue was not properly addressed in a previous paper, which is why Clark et al. were able to not only inappropriately extend the method, but also to justify their extensions based on the Loreau and Hector (2019) Comment *alone*, without even a reference to the arguments the Comment was purportedly addressing. We would have been very happy to remove the line that Reviewer 2 has suggested here, but this glaring omission by the authors, and their brazen sidestepping of the issue, left us no choice but to at least mention in passing that the fatal problems with the LH method were never addressed and thus could not be bypassed in the way the authors of Clark et al. (2019) assumed they could be. Something clear should be put on the record so that future researchers cannot pretend that the Loreau and Hector Comment (2019) will allow some sort of carte blanche when developing new statistics or methods based on the LH partitioning (a particularly important message in a methods journal).

> **R2.19)** Figure 2: The 'Reality' label is unhelpful for the reader. It would be better to say 'monoculture density effect on ecosystem function' or something like that.

Agreed. We have replaced 'Reality' with 'Observed ecosystem functioning based on law of constant final yield' in the figure legend.

Reviewer 3

R3.1) This paper is a comment on the recent paper Clark et al. (2019). The authors raise serious concerns about this paper, especially about the estimator used therein and they demonstrate how to derive unbiased estimators properly.

First of all, I would like to say that I also have some concerns about Clark et al. (2019), the least I can say is that the way maths are done in that study (especially in the appendices) is quite hard to follow (and I would like to believe that I have decent knowledge in probability and statistics). Therefore, I unsurprisingly concur various statements make by Pillai and Gouhier, for instance,

This turn of phrase side-stepped the fact that their accurate estimations came about from averaging the results of 20,000 random draws from the total species pool.

This is indeed quite problematic. Also, even though equation (1) is quite simple, it looks like the bias was missed by Clark et al. (2019), which is worrisome.

In my opinion, the points discussed up to page 16 are worth being published (but there is one important issue, see below).

We are glad that Reviewer 3 agrees with the issues we outlined in the Clark et al. (2019) paper.

R3.2) However, from page 16 to page 25 the authors are repeating what they have already said in no less that two entire papers (Pillai and Gouhier 2019a, 2019b). I think the authors should simply write that Clark et al. (2019) sadly ignored their recent paper (Pillai and Gouhier 2019a), but I do not see why they should elaborate on the non-linearity issue over and over. Note that I agree that Loreau and Hector (2019) does not address the fondamental issues described in Pillai and Gouhier (2019a) very well but Pillai and Gouhier (2019b) makes that very clear already.

We respectfully disagree. Although the original paper in Ecology (referred to as Pillai and Gouhier [2019a] by Reviewer 3) introduces the problem that arises in the Loreau-Hector partitioning scheme because of nonlinearity in *monocultures*, this issue appears to have been misunderstood by the BEF community.

For instance, numerous papers have been (or are about to be) published that blatantly misread our mathematical arguments, such as Bloom et al. (2019) which misunderstood our point that it was the nonlinear relationship between density and ecosystem functioning in *monocultures* (not mixtures) that would yield nonsensical estimates of complementarity and selection, or where the issue of nonlinearity was simply ignored/unacknowledged (such as Delory et al. 2019, Kothari et al. 2020). Loreau and Hector (2019)'s Comment on our paper in

Ecology made similar errors by not understanding that it is nonlinearity in *monocultures* that can masquerade as a biodiversity effect in *mixtures* under the Loreau-Hector partitioning scheme. Hence, even though we made this point in our Ecology paper, it was never really understood. Despite this, we have no intention of addressing one-by-one the numerous misuses of the Loreau-Hector method that have come out since our paper (e.g., Bloom et al. 2019, Delory et al. 2019, Kothari et al. 2020). But the argument made in Clark et al. (2019) is different precisely because it is a *methodological extension* based on specific justifications that appeared on the record in this journal.

More importantly, we are *not* repeating ourselves in this section, nor are we repeating what was "already said in no less [than] two entire papers". It is also *completely incorrect* of Reviewer 3 to claim based solely on his/her subjective impressions that we are "elaborat[ing] on the non-linearity issue over and over". We have published literally *only one* paper on this issue, and even then, were prevented through backroom malfeasance during the peer-review process at Ecology from responding to the muddled misreading of our paper represented by Loreau and Hector's Comment. The second reference to Pillai and Gouhier (2019b) made by Reviewer 3 is just an *arXiv preprint* that will sadly never be published in a peer-reviewed journal because Ecology's former Editor-in-Chief decided to remove the editor handling our paper and then proceeded to reject our replies for political reasons. This is why Clark et al. (as well as others such as Delory et al. 2019) felt emboldened enough to brazenly cite the Loreau and Hector Comment, without even acknowledging the paper or arguments the Comment was purportedly addressing. It is simply too easy for researchers to ignore preprints and continue misapplying the Loreau-Hector partitioning scheme, which is why it is so important that a section of this peer-reviewed manuscript be dedicated to this issue.

What we are addressing in this section of our manuscript are the very *specific methodological claims and justifications* made in Clark et al. (2019) – justifications that are very likely to be repeated again by others when developing statistics in biodiversity research without a formal correction placed on the record.

What Clark et al. did in their paper was conflate the subtle distinctions between the problems entailed by nonlinearity when measuring a total *net change* in a system, with the problems that arise when trying to measure any smaller components or *partitions of that net change*. The measurability issue that arises from this subtle distinction was something that was implicit in Pillai and Gouhier (2019) but not spelled in sufficient detail to avoid ongoing confusion, and the justification made in Clark et al. (2019). This is the *first time* that we are explicitly teasing apart this issue – a task made necessary by this particular methods paper.

In this section of our manuscript we argued for the first time, using an easy to grasp univariate example, how any *net change* defined by a *path-independent* function (i.e., change in a conservative vector field) can be measured as simply the difference between the function at the final and initial states of the underlying variable. Any net change such as the net biodiversity effect can be treated as a *state function* that is only dependent on these initial/final states. What this means for the LH method is that we *can* technically, by

controlling for the nonlinearity issue, measure the net biodiversity effect. As such, Clark et al.'s argument that the nonlinearity issue will not matter because the LH method will work for any baseline is, in part, correct.

Because the net biodiversity function is a state function (path-independent) it is technically *measurable* and thus, in theory, unbiased statistics can be developed for its estimation. *However*, Clark et al.'s justification based on the robustness of the LH null baseline will not extend to or allow a further partitioning of this net biodiversity effect into smaller components; specifically, the selection and complementarity effects will remain non-measurable, and hence non-estimable. Thus, although the effects of nonlinearity can be partially controlled for in the LH method, the statistical *non-identifiability* of selection and complementarity will remain and will always prevent (*a priori*) the development of any unbiased statistics for estimation. This holds (except in the rarest of cases) even for the correct estimators we derived.

We now clarify and better motivate the beginning of this section so other readers also understand the subtle distinction we are trying to address here [p19: 25-26- p20:1-17].

It is extremely important to note that even after we explained this in our Comment, Reviewer 1 reiterated, yet again in their referee report (e.g., see response to Reviewer 1, R1.35), the claim that their approach is not affected by nonlinearity because of the robustness of the null baseline:

> **"[…] if we stick to the strict interpretation of the Loreau and Hector model that I discuss above (i.e. monocultures in 1/N of the area, vs. a mixture in the full area), *then I am afraid that your critique doesn't hold anymore* (since we aren't trying to pretend that we can interpolate between mixture and monoculture arrangements)." [emphasis added].**

Reviewer 1 has demonstrated better than we could have, precisely why the confusion over how nonlinearity affects measurability needs to be cleared up and placed on the record! What Reviewer 1 is trying to suggest is that the effects of nonlinearity can be controlled for by adjusting the baseline of the Loreau-Hector method so as to not inflate the *net biodiversity effect*. Unfortunately, Reviewer 1 does not realize that the argument that nonlinearity can be controlled for when measuring the net biodiversity effect will not carry over for *selection* or *complementarity*. The subtle distinction between how nonlinearity may allow measurability in the net biodiversity effect while at the same time ensuring *non*-measurability in selection/complementarity is now the most important service this paper can render to the field.

This part of our Comment, like every other aspect of our Comment, is a response to *published mistakes* in the Clark et al. 2019 paper – not a rehash of points made elsewhere. And at this point it is clearly not enough to simply cite the original Pillai and Gouhier (2019) paper; Reviewer 1 has already now tried to wave off the seriousness of the nonlinear issue by

repeating their paper's original arguments in their referee report. Simply citing our original paper without also formally debunking the very specific justifications used in Clark et al. (2019) will almost certainly allow its authors to once again frame the issue to their advantage so as to dismiss the non-measurability problem with a wave of the hand when they write their own formal Response to our manuscript. But even more importantly, we need to ensure that no one else attempts to statistically extend the Loreau-Hector method using the same justifications, which will almost be certainly be the case if Clark et al.'s mistaken rationalizations remain unchallenged in the literature.

> **R3.3)** Importantly enough, I do have a major comments regarding equation (13) which I believe is wrong. This is quite problematic cause if the authors want to prove that Clark et al. (2019) they should do it right. I demand the authors to prove me wrong on this. So here is the problem, if I am right this is simply following all what the authors have done before. But if I am correct then the authors forgot [...]. So what am I missing? If I am right, this should change the rest of the section. If I am wrong then the authors should detail how they jump from (12) to (13).

Former Equation 13 (now Eq. 14) is correct. We fear that Reviewer 3 may have made the transition between lines Eq. 13 (former 12) and Eq. 14 (former 13) far more complicated than it need be. First consider that the following holds

$$\sum_{i}^{N}\sum_{j}^{N} x_i y_j = \sum_{i} x_i \sum_{j} y_j = N^2\,\overline{x}\,\overline{y}.$$

Thus we have

$$\frac{1}{N}\left(\frac{n-1}{N-1}\right)\sum_{i}^{N}\sum_{j}^{N} x_i y_j = \left(\frac{Nn-N}{N-1}\right)\overline{x}\,\overline{y}$$

Substituting the above expression into the second term of Eq. 13 (former Eq. 12) will yield Eq. 14 (former Eq. 13) as shown in our manuscript.

> **R3.4)** Lastly, I would like to mention that I deplore the authors' taste for controversy, but given the depiction of the authors contribution, I can tell they know what they are doing and readers will certainly be happy to know to whom their complaints must be addressed. I must say though that I remain quite unsure that this is an efficient way to communicate science.

We are sorry that Reviewer 3 gets this impression, but we are not going out of our way to seek controversy. Above all, this comment has no bearing on the validity of our mathematical results. It is also inappropriate, in general, for Reviewer 3 to proffer his/her personal or hearsay opinions regarding our purported "taste for controversy" – especially when it is not even true! We are one of only a few teams of researchers who have pointed out the many and significant issues in BEF. Doing so has cost us a lot of time, energy and frustration due to the political nature of the review process. None of this has been easy for us and, as one can

see with the decision rendered with this paper even *clear-cut mathematical results* can be rejected when objective statistical definitions are treated merely as 'differences of opinion'. However, we believe that we are providing an important service to the broader community by identifying critical issues in the foundational methods of the field that have gone unrecognized for decades.

R3.5) In general for the proofs, I would suggest to the authors to mention the linearity of the expectation operator and the Koenig-Huygens theorem.

This appears to be a distinction in the way statistics is taught to anglophones and francophones. In North America, these properties are derived from scratch, even in introductory statistical courses with no mention of the anachronistic label 'Koenig-Hygens theorem'. It's almost impossible to add this label to the intermediate steps of the mathematical demonstration without breaking the flow of the argument. Additionally, nothing is gained by adding an anachronistic label to what is an obvious and self-evident algebraic operation.

R3.6) l.6-7 may be worth citing the original paper (Loreau and Hector 2001) (and Price's equation)?

Agreed. We have now expanded this section to make it more clear [p4: line 6].

R3.7) l.8 (x_1,y_1) instead of (x_1,y_i)

Agreed and fixed. Thanks for catching this error!

R3.8) l.17 I would mention that it is a simple of Koenig-Huygens theorem.

We respectfully disagree. This would only confuse the reader – almost no one in the anglophone world is taught this anachronistic label for a self-evident algebraic operation in introductory probability or statistics courses.

R3.9) l.20 right-hand side should be $\sum_{i}^{n} = \mathrm{Cov}(x_i, y_i)$ instead of $\sum_{i}^{n} = \mathrm{Cov}(x_i, y_j)$

Agreed and fixed. Should be Cov(x_i , y_i)

R3.10)
l.21 " given that" => "under the assumption"?

No the phrase "given that" is correct, since x and y are not, by their nature, independently distributed. The fact that each species *i* sampled comes as an x_i and y_i pair is 'a given' and is not merely assumed (i.e., our derivation is not an argument contingent on an assumption).

R3.11) Page 5: might be worth saying explicitly what are the bias.

Agreed. We have now explicitly stated what the bias is here [p6: 1-5]

R3.12) Page 9: again I would mention the use of Koenig-Huygens theorem

We respectfully disagree. This would only confuse the reader – almost no one in the anglophone world is taught this anachronistic label for a self-evident algebraic operation that is proven in introductory probability or statistics courses.

References

Bloom, E. H., T. D. Northfield, and D. W. Crowder. 2019. A novel application of the Price equation reveals that landscape diversity promotes the response of bees to regionally rare plant species. Ecology Letters 22:2103–2110.

Clark, A. T., K. E. Barry, C. Roscher, T. Buchmann, M. Loreau, and W. S. Harpole. 2019. How to estimate complementarity and selection effects from an incomplete sample of species. Methods in Ecology and Evolution 10:2141–2152.

Delory, BM, Weidlich, EWA, von Gillhaussen, P, Temperton, VM. When history matters: The overlooked role of priority effects in grassland overyielding. Funct Ecol. 2019; 33: 2369–2380. https://doi.org/10.1111/1365-2435.13455.

Kothari, S., R. Montgomery, and J. Cavender-Bares. 2020. Physiological responses to light explain competition and facilitation in a tree diversity experiment. bioRxiv:845701.

Pillai, P., and T. C. Gouhier. 2019. Not even wrong: the spurious measurement of biodiversity's effects on ecosystem functioning. Ecology 100:e02645.

Pillai, P., and T. C. Gouhier. 2019a. Not even wrong: Reply to Loreau and Hector. arXiv:1910.13563.