Reply to decision letter reviews: #177

We would like to thank the editor and the reviewers for their useful suggestions and below we provide a detailed response as well as a tally of all the changes that were made in the manuscript. For an easier overview of all the changes made, we also provide a summary of changes.

Please note that the editor's and reviewers' comments are in bold while our answers are underneath in normal script.

A track-changes comparison of the previous submission and the revised submission can be found on: https://draftable.com/compare/ZRslygEGcvBF

A track-changes manuscript is provided with the file: "PCIRR-RNR-Fox et al 2005 replication & extension main manuscript-track-changes.docx"

Summary of changes

Below we provide a table with a summary of the main changes to the manuscript and our response to the editor and reviewers:

Section	Actions taken in the current manuscript
General	Ed: Updated the original methods and added some supplementary analyses for robustness checks. R1: Made some changes to the introduction and methodologies (to ensure it is more up-to-date). R2: Supplemented the original analyses with some additional statistical tests.
Introduction	Ed: Broadened the literature review. R1: Updated some references and included more additional studies.
Methods	Ed: Updated some measures in the methodology (eg. wine years) R1: Clarified that this is a within-subjects study, updated the deviations session, added attention checks and visual depiction in study 2. R2: Removed the extension of patriotism, added an exploratory analysis to address possible failed replication due to misinterpretation of instruction in study 2
Results	R2: Updated the exploratory analysis
Supplementary materials	R1: Added the changes in the deviation tables

Note. Ed = Editor, R1/R2/R3 = Reviewer 1/2/3

Response to Editor: Prof. Chris Chambers

> **Two reviewers, including one of the authors of the target study for the replication, have now kindly evaluated the Stage 1 submission. At a broad level, the submission has many promising characteristics but you will see that the reviews raise a number of substantial issues that will need to be addressed to achieve in-principle acceptance (IPA). I will highlight some of the headline points.**
>
> **In terms of the framing of the replication, a broader consideration of literature would be useful (as noted by Craig Fox), as well as clarification of the 3 rationale for certain predictions and extensions on the original methodology.**
>
> **Other key points include consideration of deviations from protocol and potential negative consequences of doing so (e.g. use of a single sample for all three studies), and alternatively whether keeping certain features of the original methods the same might inadvertently reduce validity due to changes in how the measure will be received now compared to 20 years ago. Keeping a replication study as faithful as possible to the original methods while also ensuring that it provides a theoretically valid replication can be a challenging balancing act -- there are often no perfect solutions, but one option would be to run parallel studies in different samples using original vs updated methods to assess robustness.**
>
> **The reviewers also highlight concerns with possible knock-on effects of completing different measures, and raise queries regarding the patriotism intervention (and the statistical power of that analysis as well as other extensions, noted by Leonardo Cohen).**
>
> **Overall, the reviews are helpful and constructive and I judge that with careful revision this submission can suitable for IPA at PCI RR. On this basis I am happy to invite a major revision and response.**

Thank you for the reviews obtained, your feedback, and the invitation to revise and resubmit. We appreciate a lot for all the suggestions and comments provided and below we provided our point to point response to the reviewers.

Response to Reviewer #1: Prof. Craig R. Fox

> **I appreciate very much the interest and efforts of the authors in attempting a replication of our studies from the 2005 paper as a "registered report." This is an important service to the field.**
>
> **Of the six studies reported in our original paper, one (Study 4) has had failed a literal replication attempt, but was subsequently conceptually replicated. A second study (Study 1) has been previously successfully replicated.**
>
> **The authors chose to attempt to replicate Studies 1, 2, and 5 (Studies 3 and 6 would be a little more difficult to replicate using an online sample of convenience).**
>
> **The authors also propose to collect data on potentially related constructs (desire for choice diversity as a possible predictor of the extent of diversification; patriotism as a possible covariate of partition dependence involving domestic vs. international charities).**
>
> **Due to heavy teaching and service responsibilities this month I've only had a moment to quickly review the Stage 1 manuscript. I provide my major comments and specific comments below.**

Thank you very much for taking the time to review the manuscript. We really appreciate all the comments and your input is very much appreciated in helping us improve our manuscript.

Also thank you for sharing all the information about previous replications.

> **1. Review of literature. I think it may be important to contextualize the Fox, Ratner & Lieb (2005) paper within a larger literature on partition dependence. Although, as noted in the manuscript, there have been few attempts that I am aware of to literally replicate our findings, there are probably dozens of studies by now (many from my lab group and many from other groups) that conceptually replicate findings of partition dependence across a wide range of guises and contexts.**
>
> **The reason I bring this up is that a Bayesian reader will want to know how strong our prior beliefs should be that these phenomena are robust.**
>
> **I'd be happy to provide references, but for now I'll just suggest that the literature review might be broadened, and it might acknowledge that these kinds of findings are not limited to the studies and paradigms outlined by Fox, Ratner & Lieb (2005).**

We appreciate the suggestions on broadening our literature review. When reviewing the literature, we aimed to focus our summary on the literature directly related to the studies and methods used in the target article. A systematic comprehensive review of this literature would be valuable, especially given that it can aggregate samples to get close to accurate effect size estimates and provide indications for possible publication bias and possible moderators. In our case, a review of the literature goes far beyond the scope of a replication of a specific article in a very comprehensive literature, and we chose to keep the scope clear - a well-powered Registered Report replication.

To build on this advice and address this point we revised to include more studies that conceptually replicatde findings on partition dependence in the introduction, hoping that this will offer readers a slightly better understanding on the effect of partition dependence across different contexts.

> **2. Methodological differences. It is often impractical or impossible to conduct a literal replication of studies originally conducted in a specific location long ago. This said, while I expect that the first-order findings from our studies will prove robust to small methodological variations, I do think that some of the variations could possibly prove problematic.**
> **I presume that mTurkers completing three studies back-to-back-to-back in 2022 online will have less personal connection (and devote less concerted attention) to the content of these surveys than did Duke students in-person each responding to a quick single pencil-and-paper study in the early 2000's. For instance, family incomes have increased (affecting Study 1), charities may be less familiar (affecting Study 2), wines are older (affecting Study 5).**

We understand these concerns, and we believe this is exactly one of the reasons why this and other replications are needed, to test the robustness of these effects. We are also hopeful and optimistic about the replications, our experience of now over 80 replications of judgment and decision-making paradigms from decades ago of some small undergraduate sample from a top US university has shown that many judgment and decision-making are robust and with comparable effects (our efforts are summarized on this page: https://mgto.org/pre-registered-replications/), and when there are mixed findings and/or failed attempts we were able to learn something new and important about the target phenomenon.

We have already carefully documented the comparisons between our sample and the target articles as well as the deviations of our efforts, to help readers assess exactly these points.

To address these concerns in advance, we added a limitations section to our discussion aiming to cover these issues. We were planning to complete this section after data collection, in light of the results, yet we see the value of stating our plans to discuss that up front.

> **I outline more such variations that could be an issue in my specific comments below. Mind you, I could imagine everything replicating despite this (honestly, I don't recall there being much that didn't work in our file drawer); however, I also wouldn't be shocked if some differences in methodology end up making a difference here.**
> **For example, there might not be enough variation in familiarity among 2022 mTurkers with wines from 20 years ago to replicate an expertise effect that we observed among Duke graduate students 20 years ago. At very least, I'd like to see the authors acknowledge a little more explicitly the methodological differences. I imagine that if everything does replicate as expected despite these limitations then that only makes the results more convincing, no? And if everything does not replicate perfectly these variations may provide clues to previously unknown moderators worth investigating in future studies.**

We would like to think that the results would be convincing regardless of what the outcome would be, and would help update our knowledge about the phenomenon using these methods and context.

One of the reasons why we believe it is important to run the three studies together is to examine whether there is a specific study or domain that works better than others, and to allow us to address any possible sample concern if one study replicates but another does not.

There is always a tricky dilemma when it comes to replications. If we change the stimuli and fail, and there are countless ways to change the stimuli, then it could be argued that the change is what caused the failure. If we do not change the stimuli and fail, then it could be argued that we did not make necessary adjustments, whatever those may be, to update to the current context. Therefore, in our replication we wanted to stay as close as possible to the original stimuli to allow a possible failure to be more indicative. If that fails, then we could then understand that an adjustment is needed and then proceed to test that in future follow-up studies, if we find that valuable and relevant. If that succeeds, then we provided support for the original and the robustness and generalizability over time and context.

We did not consider the wine years an issue, yet this may indicate our limited knowledge about aging wine, and we now understand that wines this old might indicate an issue. Therefore, we updated the wine years to be similar to the timescale used in the target (~3-5 years old).

> **3. Correcting the record. In a couple of cases I think there are mistakes or important omissions in citing prior literature. See specific comments below.**
> **Specific comments:**
> **p.8, background: FYI: I first used ("coined") the term "partition dependence" in print in Fox & Rottenstreich (2003), in the context of judged probabilities.**
> **Fox, C. R., & Rottenstreich, Y. (2003). Partition priming in judgment under uncertainty. Psychological Science, 14(3), 195-200.**

Thank you for sharing that. We revised the reference accordingly.

> **1. pp.9-10, the film study cited at the bottom of p.9 is something that we successfully ran years ago (and alluded to in the discussion of Fox et al., 2005) but didn't end up including in the Fox et al. (2005) paper writeup (we had more material than we needed). I surmise that the authors must have received the original materials for that study from Rebecca Ratner. But mentioning it here in the paper could be confusing to readers.**

Thank you. Including an experiment that is not well elaborated in the article could be confusing. We changed it to another study in the original article, which concerns the allocation of free lunches to illustrate partition dependence.

> **2. p.10, par. 3: It is not clear what the authors mean about individuals taking partition dependence into account in their own decision-making strategies. What exactly do you have in mind?**

Thank you for raising this issue. We removed this part to minimize ambiguity.

> **3. p.10, last paragraph: My reading of Xing et al. (2020) is that their analyses "did not provide evidence that financial aid status moderated the way income partitions influenced resource allocation." What they did find is that students on financial aid allocated more financial aid to families with lower incomes in both partition conditions.**

Thank you for raising this point. We removed the part concerning the moderation of financial aid status to avoid potential ambiguity to readers.

> **4. p.11, par. 1: actually, although Reichelson et al. (2018) failed to directly replicate close version our original Study 4 results among adults, those researchers did manage to create a successful conceptual replication among children.**
> **See: Reichelson, S., Zax, A., Patalano, A. L., & Barth, H. C. (2019). Partition dependence in development: Are children's decisions shaped by the arbitrary grouping of options?. Quarterly Journal of Experimental Psychology, 72(5), 1029-1036.**
> **They later concluded that the transparency of the task (to adults) may have been part of the issue.**
> **Indeed, in a follow-up they found that their simple paradigm replicated among a separate sample of children but not adults:**
> **Williams, K., Zax, A., Reichelson, S., Patalano, A. L., & Barth, H. (2020). Developmental change in partition dependent resource allocation behavior. Memory & Cognition, 48(6), 1007-1014.**

Thank you. We included the successful replications among children with reference to the studies that you provided. Given that some research has shown mixed results regarding the effect of partition dependence among adults, we feel that this motivates the need for more replications to find out what factors might be at play.

> **5. p. 11, Method: single data collection. The authors should make clear that the study is entirely within-subject rather than between-subjects in parallel.**

Thank you very much for catching that, we added the sentence "This was a within-subject design replication, but in each of the three studies we ran participants between-subjects. " under the session "Hypotheses and findings in target article" to make it clearer.

> **7. p.14: Extensions:**
> **I'm surprised to hear the assertion that there are no off-the-shelf scales for preferences to diversify, but it is an interesting and still not settled question what drives partition dependence in this context. A mindless application of motive to diversify (which is not sensitive to the particular partition presented to participants) could be one mechanism, though there could of course be others. Of course preference to diversify is not really an explanation (several reasons for variety seeking have been proposed in the literature—such as seeking information, overgeneralized concerns about satiation, desire to look adventurous, etc.).**
> **As for patriotism, I'm not persuaded that this will moderate partition dependence. I think I understand the instinct that patriots might be more drawn to U.S. charities than international charities (on the other hand, many patriots like to sponsor international causes in the name of advancing US prestige abroad).**
> **This said, while I could imagine patriotism increasing motivation to donate to domestic causes, it is difficult to see why this measure would otherwise moderate partition dependence (except for the trivial case where some participants give zero to the International United Way regardless of partition).**

Thank you for your suggestion. We removed the extension for patriotism as it seems to be quite irrelevant to include it in the current replication study.

> **8. pp.16-17 Participants: I'm a little concerned that the authors are using a single sample of participants for all 3 studies. Of course this could lead to correlated error across studies, and possible subject fatigue. Presenting participants with multiple partitioned option sets might also inflate attention to differences among partitions. (On the other hand, past research suggests that partition dependence is usually robust to transparency about the partitions used in a study—see e.g. Xing et al. 2020; Fox & Clemen, 2005; Sonnemann, Camerer, Fox & Langer, 2013).**

Thank you for raising this. What you suggested is exactly the reason why we would want to combine the studies and examine them together. Combining the three studies into a single study design and using the same participants may potentially provide us with additional insights that we would not know otherwise. Using your example, if we fail to find support or weaker effects for some of the studies, we could then examine order effects. This may lead to new insights as to how participant fatigue is associated with partition dependence.

Another important benefit is that if one study succeeds and another fails then we can rule out sample concerns regarding seriousness or attentiveness or context (e.g., timing like pandemic), which proved important in many previous replications we ran in addressing various concerns. We believe that by now we have solid evidence that for JDM paradigms the CloudResearch subsample of MTurk is reliable, also in pandemic times, yet this design addresses the specific evidence regarding that specific data collection in that specific time.

> **9. p.21, Table 7: It is interesting that the authors chose to use the same income brackets (and same university) as we did in the study that we ran 20 years ago at Duke. Nominal incomes have risen considerably since that time and of course mTurkers generally have no ties to Duke University or, necessarily, a sense of university tuition and financial aid norms. This said, my guess is that these differences won't stop the authors from replicating our effects.**

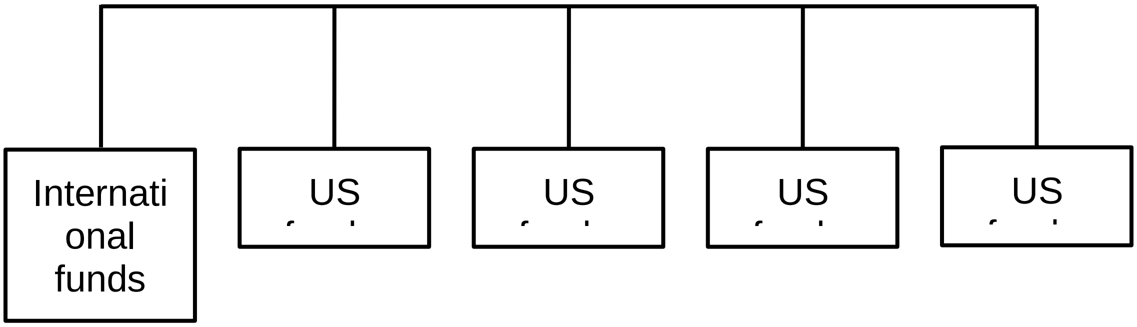
Thank you for raising this point. We updated both the manuscript and supplementary and changed Duke University to simply a university. It is also noted in the deviations session to explain the reason for the change.

Also, the direct replication conducted by Xing et al. (2020) used the same income brackets as well while successfully replicated the effects. Therefore, we think that it may be reasonable to follow the original study.

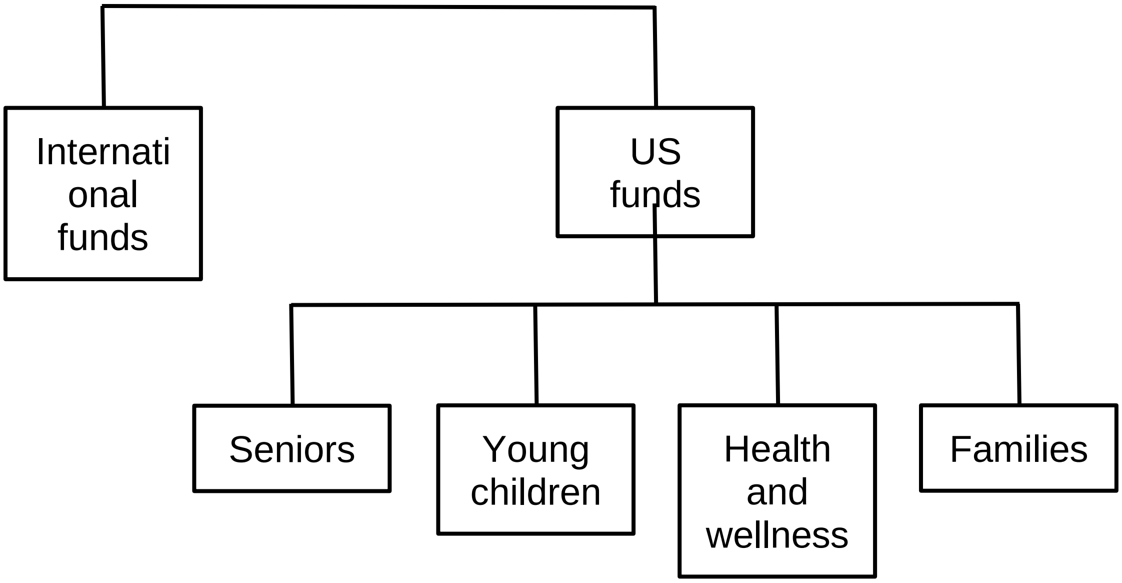
> **10. p.21, Study 2 methods. Procedural differences from the original study that are mentioned on p.23 might be anticipated here. Of course the original study used Duke students (who live in Durham county), an incentive-compatible consequence (a random student's preferences would direct a real donation of $2 per participant); also the funds involved Durham County vs. International United Way Charities—so the 2005 subjects had a closer connection to the decision they were making. This said, I am guessing if the partitioning elicitation is done right (it looks fine based on the supplementary materials) I expect that this shouldn't stop the authors from replicating our original finding. I do think it would be helpful to readers to provide a visual depiction of how the one-stage versus two-stage elicitation is to be implemented.**

Thank you for the suggestion. We added visual depiction for Study 2 under "Manipulations".

Nonhierarchical partition condition:



Hierarchical partition condition:



> **11. p.21, Study 5 methods. Here again, it would be helpful to reproduce the elicitation to make clear how the partitioning was implemented and that the six wines included two wines from each of the grape and region categories. Perhaps it won't make much difference but if I were trying to replicate this study today I probably wouldn't use the same materials: the wines are now 20 years older (much more expensive if at all available, in some cases probably past their prime, etc.).**

Thank you very much for raising the issue of the wine list update, this is something that we did not consider before but seems very relevant to rerunning this replication. We updated the wine list by changing the year of the wines, so that it may make more sense to the participants and readers.

Full details about the partitioning of the wines were provided in the supplementary, and all materials for the study were included in the OSF.

> **12. p.22 Measures: Of course it is an open question how the distribution of wine expertise will differ in this population compared to the 2005 sample of graduate students. For instance, if there is not sufficient variance in this population you would be less likely to find moderation of partition dependence by expertise. Note that in our sample the median number of bottles that distinguished "novices" from "experts" was 4, and of course the current sample is likely to differ. Also note that we only binarized our data for illustrative purposes. In any case, expertise and other indicators of strength of preferences for specific wines should in theory moderate the magnitude of partition dependence effects. The question is whether the measures here will tap into that.**
>
> **13. p.24, Table 9: taking stock of the classification of the replication, I think that it is basically accurate, but elides reasons why differences could be important, as detailed above.**

Thank you for the comment and it is much appreciated. We now modified the original dichotomous measure of expertise to a continuous variable and explained that this is a better strategy to treat the expertise measure as the wine expertise distribution may be different between the replication and the original population, under the "deviation" session.

We will also elaborate on this potential issue in the limitation section in the discussion, explaining that:

> "the effect of expertise on partition dependence may depend on whether there is sufficient variance on the number of wine experts among the targeted population. Since if most participants are "experts", meaning they have greater relevant expertise, they would be less likely to be affected by partition dependence. Hence, the moderation effect of expertise might be less pronounced."

> **14. p.25: the proposed data analysis strategy seems fine to me and basically duplicates our approach from 2005.**

Thank you.

> **15. p.26, outliers and exclusions. I'm glad that the researchers are planning to weed out non-human responses, but I do worry a bit about the attentiveness of an mTurk sample. Many researchers, for instance, use attention filters. I don't feel strongly about this though. It might even be that less attentive participants are more likely to exhibit partition dependence.**

Thank you for the suggestion.

Our experience is that MTurkers are professional survey takers and they have been generally shown to be attentive in our many other replications. Moreover, we did indirectly check participants' attentiveness in the beginning of the Qualtrics, by randomly assigning the answer options of "yes"/"no"/"not sure, probably not", for the first questions in the study outline where participants indicate consent and agreement to participate.

We however do agree that adding more attentiveness checks may help reassure further about the attentiveness. Hence, we added three more attention check questions, and noted in the methods section under the "Attention checks" subsection:

> Given feedback received in the peer-review process we added three items to the desire for choice diversity scale, randomized in order among the other items, which serve as attention checks (1 = Strongly disagree; 4 = Neutral; 7 = Strongly agree): "100 is larger than fifty." (failure: <=4) , "One hundred is smaller than 50." (failure: >=4), "Please select "Agree" (failure: != 6). Failing to answer two out of the three correctly will qualify for exclusion.

Given feedback provided in a number of our other PCIRR submissions (with the same editor) that indicated the disadvantages of that approach (especially using the +-3SD), we decided to take out the outlier analysis.

Response to Reviewer #2: Dr. Leonardo Cohen

> **The authors propose to replicate three studies in Fox et al. 2005 which demonstrate "partition dependence", and extend with new measurements to better explain the drivers behind the phenomenom observed. Partition dependence has been explored in many areas of psychology research and is also known as the 1/N or "naive diversification" effect. I thank the authors for providing such detailed and open experimental materials. While it is important to replicate studies, in particular those done at such specific "WEIRD" original population, unfortunately, I am not sure that an exact replication of this specific set of studies conducted most than 15 years ago makes sense today, without further changes to ensure its generalizability. Perhaps a conceptual replication is more suited.**

Thank you for your review and comments. Much appreciated.

> **For example, in Study 1, students at Duke University were asked about the distribution of financial aid to students at Duke University. There must be an effect of students making choices about financial aid at their own institution. For example, were those students more likely to believe that their answers could affect actual distribution of financial aid (and perhaps even affect themselves or their colleagues), as opposed to asking a broader audience from Mechanical Turk who are not necessarily students and not at Duke? I would go as far as suggesting that many participants from Mechanical Turk will not even be aware of what Duke University is.**
> **I believe that the current proposed study setting makes it much more hypothetical than the original study, and therefore not a true replication. I also propose that the stipend and the income brackets should at least be adjusted for inflation.**

These are great points, and are appreciated. As replicators we often face the delicate dilemma of whether to change something or not, given that we often do not know how any of these change might affect the replication rates, and a small change may have some unintended side-effects, which is why with a direct replication we aim to remain as close as possible to the original, unless we see a clear and obvious potential problem.

We agree that a reference to Duke University might be confusing, and does not seem like a crucial point, and so we removed the reference to Duke.

The replication of Study 1 conducted by Xing et al. (2020) used the same income brackets as the original and concluded a successful replication. Therefore, we consider that an indication that these income brackets still seem relevant so we decided to keep those.

One of the reasons we decided on a combined unified data collection of the three studies was to try and determine whether some effects might be less generalizable than others, and this is something we want to know. Many of the JDM paradigms involve having participants reflect on facing a situation that the participants are not familiar with, involving a decision that is not from their own lives (consider scenarios from negative-positive framing effects having to think like a pandemic decision maker in the badly named "Asian Disease" or scenarios from action-effect having to think like an investment broker, as in the many behavioral economics scenarios).

We used the classification of LeBel et al. 2018 to categorize the type of replication, and categorized the replication of X. We are not sure what categorization you referred to when using the term "true replication", and what that means, and we are happy to make additional adjustments to our categorization given clear editorial guidance.

The bottom line is that when we conduct replications things are different and context changes, the participants change, the time is different, the weather and experiments/method are different, etc. etc. and so we need to be careful and clear in how we think of what a "true replication" is and when one qualifies for that label. The LeBel et al. 2018 criteria is fairly clear about those and to our knowledge is the best reference we have at the moment.

> **Unfortunately Study 2 is also not a true replication because the authors change one of the categories from Durham funds (where Duke is located) to United States funds. While this is important because they can no longer guarantee participants will be local to Durham, surely individuals are more likely to be emotionally attached and loyal to their local community (in-group) than a country-wide community.**
> **Can the authors filter MTurk to a more specific locality, for example, a state - and ensure that location is guaranteed and cannot be faked via IP geolocation identification? (Authors should also be aware that their information sheet does not 4 inform participants that their IP address is being recorded.) Although even at state level, this would not be an exact replication of the in-group affiliation as at county level from the original study.**

Thank you for raising this concern.

Yes, we made adjustments to remove the references to a specific local. You refer to the issue of emotional connection, but there is nothing specific in the partition dependence theory that refers to that point, and we should hope that partition dependence would replicate and would generalize beyond the very specific locale of participants in Durham. In our replication we were aiming to go beyond the specific Duke University students and the Durham students. If this does not replicate, then we would have learned something about the generalizability of that specific study design to a broad sample.

One of the many reasons why we decided to combine all the three studies into a single data collection was exactly because we wanted to check and see if all of the effects can be generalized in different situations or just some. This way, if one study replicates, but not the other, then we gain insights about what may work beyond the context of the original and what might not.

We decided to use the classification of LeBel et al. 2018 to evaluate and determine the type of replication as their criteria seem to be reasonably clear, so we believe that it is the best reference that we can think of at present. This criteria has been used in all our other replication projects, some that already received IPA from PCIRR (e.g., Li & Feldman, 2022; Zhu & Feldman, 2022).

All that said, we are open and would gladly make additional changes about the categorisation given more guidance is provided.

References:

Li, M. & Feldman, G. (2022) Revisiting mental accounting classic paradigms: Replication of the experiments reviewed in Thaler (1999). Received Stage 1 in-principle acceptance from PCI-RR. Retrieved from: https://osf.io/4ps8m/ [IPA]

Zhu, M. & Feldman. G. (2022). Revisiting the links between numeracy and decision making: Replication of Peters et al. (2006) with an extension examining confidence. Received Stage 1 in-principle acceptance from PCI-RR. Retrieved from: https://osf.io/8z6ga/ [IPA]

In Study 5, I am not a wine expert, but I am aware that wines can improve (or indeed, spoil - as some wines are supposed to be consumed closer to harvest than others) with time, depending on the variety, and at different speeds. Using the original list of wines, more than 15 years later, does not make any sense to me. Perhaps these wines no longer exist, perhaps they would taste horrible, or perhaps they are extremely rare and valuable. The authors should consider updating the list of wines. The price of the wines also need to be adjusted for inflation.

Thank you so much for pointing out this issue, this is something that we did not consider before and may indicate our limited knowledge about wine. This is also why we were more careful with that study and added more questions about wine expertise. Given your feedback, we updated the wine list by changing the year of the wines, so that it may make more sense to the participants and readers. Please also see our reply to Prof. Fox on this above.

There were also methodological flaws to the original study that do not warrant replication, and could be improved. For example, the original authors use a t-test to analyse percentage data - data which is bound between 0 and 1. This is not applicable - and might explain the unexplicably large effect size. Perhaps with the limited computing power of 15 years ago

> **this was acceptable, but nowadays much better methods are easily available. For example, a beta regression could be considered (a transformation might be needed to remove zeroes and ones, for example, y' = y*0.998+0.001). If there are many zeroes and ones, a zero-one-inflated beta regression could be used instead, and no transformation is needed. A t-test should still be conducted for the sake of replication and comparison with the original results, but a more appropriate analysis should be run instead to confirm the findings.**

Thank you for the advice and it is much appreciated. We have supplemented the original t-tests with two-proportions z-tests to check the statistical robustness of the results for Study 1 and Study 2.

> **I also believe that the hyerarchical approach in Study 2 can be very misleading. Participants are asked to allocate 100% between national and international funds, then again asked to allocate 100% between national funds. I believe that the majority of participants will not understand that the second allocation of 100% is a sub-set of whatever percentage they first allocated to the national funds. Previous work in numeracy has consistently shown that individuals are very bad at understand percentages. Instead, the authors should say that the second allocation should sum up to the percentage entered in the first allocation for US funds. As a smaller point, will participants be aware of what United Way is - perhaps the original authors knew that students at Duke know what this is, but I can envisage many respondents being unaware of their work?**

Thank you so much for raising this concern. We have rephrased the instruction to “the second allocation should sum up to the percentage entered in the first allocation for US funds.” in study 2, so it may make more sense to participants.

In response to your comment about individuals' consistent confusion in understanding percentages and what was done in Study 2, we added the following, not in the method section in subsection “Study 2: Clarity (exploratory extension)”:

> Given feedback in the peer review process, we were concerned with the clarity of the Study 2 design, the possibility that the hierarchical condition was more complex to understand than the non-hierarchical, and that participants may not have processed the percentages calculations correctly. We therefore presented the participants with a page displaying a summary of their choices and asked them to indicate whether our summary of their decisions was what they intended to choose (0 = “NO, these are not the allocations I intended to make (please explain)”; 1 = “YES, these are the allocations I intended to make”). Those who answered no were given the option to explain further.

> We considered this an exploratory measure to examine if the instructions in the nonhierarchical partition condition would be more clear to participants, in comparison to the hierarchical partition condition in Study 2, with the aim to address any possible failed replications that may be due to a misinterpretation of the instructions.

We then added a two proportion z test to compare the clarity check across the two conditions, as an exploratory direction to address possible failed replications that may be due to misinterpretation of the instructions.

As to whether participants will be aware of what United Way is: We think that the instruction/question for Study 2 may be sufficient in explaining how the two fundings (international and US) differ from each other. For example, we explained in the instruction that an allocation to international funds means that the United Way would allocate donation more specific funds abroad; whereas the US funds would be represented as four different programmes that benefit a variety of people in the country (eg, programmes that benefit seniors, young children, promote health and wellness, and strengthen families).

We also added the following clarification in the scenario, based on the Wikipedia description of United Way:

> [Clarification: United Way is an international network of over 1,800 local non-profit fundraising affiliates.]

> **When analysing Study 5, the original authors discard a considerable amount of data (participants who chose {2,2} or {3,3} types of wines). I do not understand how that can be acceptable. Data cannot be simply discarded for the sake of analytical simplification. I understand that the cells are not independent, which is one of the assumptions for contingency tables and logistic regressions. Perhaps the stimuli need to be updated so that the cells are independent, which I believe could be done with a different set of wines - or a better analytical method must be identified?**

We acknowledge that the original analyses may have some imperfections. We saw our main goal in reproducing what the original did, yet we welcome the opportunity to try and improve.

We therefore added a two-way chi squared test as a supplementary analysis to check the robustness of the results.

> **I also do not understand the rationale for the original measurement of "expertise" for Study 5 based on number of bottles of wine consumed. Is this a proven measure of expertise? What is the cutoff point that makes one an expert? And how can we justify that just because an individual buys more wine than another, they become an expert? Surely there are better measurements of expertise that the current authors could explore. I see the authors have added self-reported questions about wine in their questionnaire, but they do not explain how these will be incorporated into the analysis, and how they have been validated as true measurements of expertise.**

The original authors distinguished “novices” and “experts” according to the median number of bottles of wine they bought in the previous year, which was 4. And they binarised the data (novices: less or equal to 4; experts: higher than 4) for illustrative purposes.

We very much appreciate your advice that this may not be the best measurement for expertise,. Our aim was first and foremost to try and follow the original study as much as possible.

To address this point, we tried to further improve by modifying the original dichotomous measure of expertise into a continuous variable, as a better strategy to treat the expertise measure.

In addition, as you pointed out below, we included several exploratory wine expertise related questions, and following your feedback these are now outlined in the second part of the "Study 5: Expertise (replication + exploratory extensions)" subsection:

> We also included several exploratory measures asking participants to indicate their reasons ("Briefly describe your reasoning for choosing the three wines you selected on the previous page"), eliciting self reported familiarity with wine ("How familiar are you with white wines?"; 1 = Not familiar at all, 7 = Extremely familiar), knowledge regarding wine ("I know a lot about white wines.", 1 = Strongly disagree, 7 = Strongly agree; "How clear of an idea do you have about which characteristics of a white wine are important in providing you maximum satisfaction?"), winery names ("List here the winery names (if any) that you recognized on the previous page"), and reading wine magazines ("How often do you read wine magazines?"; 1 = Never read them, 7 = Read them all the time). We intended to use these questions for exploratory robustness checks, especially in case we fail to find support for the hypotheses.

> **In fact, upon looking at the questionnaire that the authors plan on deploying, I have noticed *many* additional measurements which are not mentioned in the registered report. For example, there are many questions about wine. I guess these might be used for later exploratory analysis. I would imagine that the additional variables being captured might be considered for additional extensions? These analyses *must* all be pre-registered. It is not clear for example if the choice diversity is added to the model as an interaction, only main effect, or both. The way these measurements are calculated is also important. And are the measurements going to be centered or standardized when entered into the model?**

Yes, the additional questions about wine were meant exploratory analyses for robustness checks and we included those in the expertise paragraph under measures, to explain our purpose of including these questions.

The additional measurements were included in the Qualtrics and therefore pre-registered as part of this study, yet we did not plan on conducting any confirmatory tests and therefore did not detail those in our planned data analysis. However, we appreciate the suggestion to include more information to aid readers of the Stage 1 see all the measures we took to address the challenge of expertise.

> **I do appreciate the authors' attempt to create a choice diversity measurement. However, one potential problem come to mind that should be addressed. How does answering the earlier questions in the questionnaire, which shows participants many choices and ask them to diversify between them, potentially influences answers to the diversity metric. Can the latter be measured without influence from the former, given that they are sequential in time (and closely follow each other)? Could a diversity metric be measured separately somehow, perhaps during another session?**

We understand this concern, there are potential issues with whatever decision we make here, displaying this before the replication, after the replication, or in a separate session entirely.

The current setup is quite common in the judgment and decision-making literature. For example, in a different replication project we are conducting on numeracy and judgment and decision-making by Peters et al. (2006), numeracy was measured after the tasks in the same session. In our interaction with the original authors and the PCIRR reviewers we were asked to keep that design, and so we did. We followed that same idea here, especially given that this is an exploratory extension direction and we wanted to ensure minimal impact on the replication itself.

We noted this as a limitation to discuss in the discussion section after data collection.

References:

Peters, E., Västfjäll, D., Slovic, P., Mertz, C. K., Mazzocco, K., & Dickert, S. (2006). Numeracy and decision making. *Psychological science*, 17(5), 407-413.

Zhu, M. & Feldman. G. (2022). Revisiting the links between numeracy and decision making: Replication of Peters et al. (2006) with an extension examining confidence. Received Stage 1 in-principle acceptance from PCI-RR. Retrieved from: https://osf.io/8z6ga/ [IPA]

> **I know that the original studies have been published in a high-quality journal and all these methodological aspects have probably been discussed and covered already by other reviewers. However, I would propose that instead of having an exact replication, perhaps a conceptual replication which improves on the original methods can be employed?**

We have a very specific scope for this project to be a close direct replication. There is much room for conceptual replication, and these are valuable, but they are complementary to direct replications. There is much room to improve and update the original's methods, and we did some of that best we could keeping with the spirit of the original's design, but further improvements and conceptual replications are directions for future research. We noted this in our "limitations and future directions" subsection in the discussion and will reflect on this after data collection when reflecting on the insights gained from our findings.

Minor points:

> **On page 12 the authors state that Study 5 is a "mixed design," I don't understand how that's the case as it appears to be a between-subjects design study?**

Thank you for catching that, that is indeed an oversight. We adjusted this to read “between + 2 predictors” in the table and “between-subject” in the text.

> **On page 16, the authors list which filters they are using, plus "etc", twice. That is very ambiguous. A comprehensive list must be used otherwise the work cannot be replicated in the future.**

We appreciate this concern. We checked the paragraphs thoroughly again and deleted the “etc.” to prevent ambiguity.

> **On page 17 they list the participants as US American - that for me implies citizenship. It should instead perhaps say US-located?**

Our survey will clearly indicate in the HIT on MTurk/CloudResearch as intended for native English speakers born and raised and currently residing in the US. This is based on our extensive experience in running replications, in aiming to minimize cultural variations and potential language and comprehension issues.

> **The patriotism questionnaire for me has many flaws. Just because participants are physically located in the US, it does not mean that they will be US citizens. (See my comment below on citizenship) So questions like "How strong is your love for your country" as very ambiguous, as the respondent might not be thinking of the United States. I also think that the flag burning question might trigger concepts of law and criminality (even though it is not a crime to burn the US flag since 1989 according to a quick search - I am not an expert), participants might think that burning of flags is illegal, and therefore, wrong regardless of my patriotism. This also applies to the question about sharing government secrets - it does not mention which government, so might not be the US, and also it is a crime, so in someone's view could be seen as wrong, regardless of patriotism. I believe these questions might be measuring different concepts.**

Thank you for raising this concern.

After consideration, and given that you and the other reviewer raised this - we agree, and decided to remove the extension of patriotism.

> **The treatment of outliers for me is strange. The authors say that outliers will be excluded if they fail to replicate the original results. However, what if it is the other way around and the replication was due to outliers, and removal of outliers actually removed the replication? I think that the authors need to have a better strategy for outliers - either removal or not removal, regardless of findings. Also it's important to note that the 3SD rule is not the best approach for percentage data and I am not sure how they would apply that on count data (Study 5).**

Thank you for this feedback, this is helpful.

Given your feedback, and similar feedback we received in our other PCIRR submissions, we decided to remove the treatment of outliers and run the analyses on the full sample.

> **The questionnaire mentions to participants that there are attention checks. What are they? And how are participants who failed the attention checks be treated? 6 You ask the following question: "This survey is only intended for native English speakers born and raised in the United States." I'm not sure this was part of the original research, as it's not stated in the original paper. Why are only those born in the US allowed to participate? And how can you guarantee that individuals will answer this truthfully? Furthermore - what are the data privacy implications of collecting potential immigration data from participants? Is there a danger that participants might lie about their citizenship?**

The sections relating to location, sample, and consent are pre-study and serve as qualifications and prerequisites to participating in the study and aiming for our target sample within the large MTurk online labor market pool. We ask in our HIT that only those born, raised, and currently residing in the US take part.

We are not collecting any immigration data, and we will not be checking or verifying citizenship. We are only making it clear to our participants the type of participants we are looking for. We rely on the participants to answer truthfully and to participate only if they qualify, just as we rely on them to share with us their answers to the questions we ask them to answer.

The attention checks are embedded in the pre-study qualifications and consent, given that the options of No/Yes/”Not sure, probably not” are in random order and so require the participants attention to follow to qualify starting the survey. Those who do not answer yes, are asked to return the HIT (task).

In addition, we add several new attention checks, please see our reply to Prof. Fox above, and these are detailed in the “Attention checks” subsection of the method section.

> **How are the authors going to analyse the data in open-ended text questions, there are many of those, such as "Briefly describe your reasoning for choosing the three wines you selected on the previous page"**

These are exploratory, and we have no specific plan on analyzing those. They are included in case we fail to find support for the original's hypotheses and would like to try and better understand how to improve in future studies.

> **The authors do not conduct a power analysis on their extensions - for example, for patriotism. What is the minimum sample size needed to observe an interaction with patriotism?**

These are exploratory directions, and our sample has been powered several times that of the power analysis of the effects in the original. In our sensitivity analyses we included the use of covariates and indicated that we are well powered to detect what are considered weak to medium effects, much weaker than those reported in the original.

In addition, per the specific point about patriotism, in this revision we removed that extension.

> **I also do not understand why the authors asked Qualtrics to randomly generate data, when they could have generated data based on the findings from the original study. Or, alternatively, given that the effect size is probably uncharacteristically too high, the authors could have generated data based on their understanding of what the difference between groups realistically could be (perhaps with a smaller effect size). This would help better determine power, as well as understand what type of sample size they would need to find a significant effect for their extensions.**

The simulated random dataset was meant to demonstrate our data analysis plan. We were able to conduct power analysis on the original based on the provided statistics and did not need to simulate data for that purpose. We provided comprehensive details about all our calculations of the original's effects and how we conducted the power analysis and the sensitivity analyses, so we are unsure what the comment here was aiming for going beyond what we did.