Dear Dr. Chris Chambers,

We are pleased to submit the revision of our Stage 2 manuscript "Does alleviating poverty increase cognitive performance? Short- and long-term evidence from a randomized controlled trial" to PCI RR.

We would like to thank you and the reviewers for their constructive comments and helpful suggestions. Below you can find a point-by-point response to all comments in bold. Beyond answering all the comments, we went through the manuscript again and copy-edited several sentences to increase the clarity of the manuscript.

To support the review process, we have submitted two versions of the updated manuscript. One pdf with the final text, and one docx., where the changes are tracked.

We look forward to your comments.

Kind regards,

Barnabas Szaszi on behalf of the author team

Reviewer reports

Reviewer #1:

After reading your Stage 2 submission and the Appendix, I'd like to say that I'm very pleased with the paper. I have a few, usually minor comments, I'd like to share with you. I'll split this review into three parts: (1) adherence to the preregistered protocol, (2) results; (3) discussion.

Adherence to the preregistered protocol:
I'm glad that, considering the complexity of the design/analytic choices, the authors have managed to strictly follow the preregistered protocol. They transparently disclose three deviations from the Stage 1 submission. Each change in the design/analysis (especially the first one - standardizing the executive function index) is justifiable and has strengthened the paper. Naturally, the abstract has been rewritten to reflect the results.

Thank you!

Results:
The authors have found that lump-sum cash transfer usually has a minor positive effect on executive functions both short- and long-term (b = 0.13 and 0.08). However, as they correctly highlight several times across the text, the evidence is non-conclusive (BFs10 = 1.51 and 0.56; Cis usually crossing 0). The findings hold across analytical/data processing choices, however, operationalization of the cognitive performance causes variation in the results.

The results of the primary analysis are comprehensibly reported. The results of the multiverse analysis, too, follow the best reporting practices. However, I'd welcome a more elaborated description of the bottom panel of Figures 1 and 3 – it's a bit less intuitive than the upper panel and maybe one additional sentence would save readers a lot of time trying to understand this visualization properly.

We have added a more elaborated description for the bottom panel now. We hope that it increases the clarity of the figure.

Could the authors double-check the results (and the corresponding code) presented in Figures 2 and 4? I think they are a bit counterintuitive - the largest prior leads to the highest support for the null compared with the other priors (I'd expect it to be the other way around).

Thank you for noticing the inconsistency here! We have indeed mixed labeling of the planned and small prior results. It is corrected now. However, as per your questions, it is correct that the largest prior leads to the highest support for the null (lowest support for the alternative). This phenomenon is not unique to our case. The Bayes factor is the ratio of the likelihood of one particular hypothesis (model) to the likelihood of another. Typically the alternative hypothesis vs the null hypothesis. The BF provides information on which model to choose on the basis of how well the observed data align with the predictions of the compared models. Hence, when our alternative model is represented by larger effects (1.57, high prior) compared to the case when it is represented by smaller effects (0.34 or 0.09), and the observed effects are relatively small (such as 0.08 and 0.13), then the latter model should produce higher likelihood for the alternative hypothesis (i.e., the small prior better aligns the observed data than the large prior). Since the likelihood of the null remains the same for all models of the alternative hypothesis, the evidence for the null is the highest when the likelihood of the alternative hypothesis is the lowest (i.e., when we use the largest priors).

A minor note – Figures 1-4 should actually be Figures 3-6, as Figure 1 depicts the flowchart, and the results of the primary analysis are visualized in Figure 2.

Thank you, we have now changed the numbering of the Figures.

Discussion:
The main findings are comprehensively summarized and logically interpreted. However, the discussion would definitively benefit from adding more references to make more specific comparisons with previous studies as well as to back up the authors' reasoning – the current version of the discussion involves only four sources (three of those papers were co-authored by some of the authors of the present study).

We have now added further references to back up our reasoning.

I'd also suggest adding a paragraph focused specifically on the limitations of the study. It's true that the authors present potential limitations implicitly across the whole discussion, but maybe a standalone paragraph would make things clearer, even though it might require some restructuration of the discussion to preserve the flow. On a similar note, the authors offer only one suggestion for future research on the topic. I'd say that the present study offers more food for thought on which directions could future research in the field of poverty alleviation focus/avoid.

Although we believe that we present several limitations throughout the discussion, we have added now the following paragraph to the manuscript:

“The question of when, why, and to what extent cash transfers affect cognition is far from being answered, which also reflects the limitations of our study. Future work should further examine how different magnitudes of cash transfers and the way they are distributed (lump sum vs. installments) affect cognitive performance; how different demographic characteristics (such as the level of money scarcity, cultural differences, or the strength of one’s social network) and the mode of task administration (online vs. onsite, computer vs. pencil based, oral vs. written) moderate the effect; whether working memory and inhibitory control are affected differently by cash transfers; and whether some specific forms of cognitive control or working memory respond more robustly to poverty alleviation.“

The authors argue that “…it is possible that cash positively impacts working memory more robustly than inhibitory control as working memory is assessed in the digit span index“ and further elaborate on that point in the discussion. However, the results of the multiverse analysis show that the unconditional money transfer has had usually negligible or even a negative effect (especially short-term) on the backward digit span test. Could the authors consider incorporating this finding into the discussion?

Thank you for raising that issue. We agree with that this variance should be mentioned. We have now added the following sentence to the corresponding part of the discussion: “Although note that the effects of cash on the backward digit span test was negligible or even negative for the short term weakening this argument.”

Minor issues:
Please check for typos, spacing, font size, etc.

We have now reviewed the manuscript and corrected several minor problems (see track-changed manuscript).

As always, I hope that the authors will find the comments useful. I’m looking forward to reading the revised version of the paper/discussion!

Thank you again for taking the time to provide such a valuable review! We hope we could address all issues sufficiently in the revised version.

Reviewer #2:

I have undertaken the Stage 2 review of the manuscript “Does alleviating poverty increase cognitive performance? Short- and long-term evidence from a randomized controlled trial” by Szaszi et al. Overall this is an interesting and well-written manuscript which has important implications for the field; the inconclusive results will provide a more balanced literature and a point of reference for future work. However, I do have three main points for revision centering on clarifying the “Hypotheses and Data Analysis Strategy” to avoid reader, ensuring that floor and ceiling effects are fully ruled out by clarifying them, and revising the Discussion section to

ensure that the authors conclusions are justified given the evidence. I detail these as below as well as some minor revisions to improve the written style and structure further.

To note to the PCI Recommender, the OSF link provided on the PCI-RR website did not successfully take me to the project page (https://osf.io/2r8a9/?view_only=1781fb681edc4cdeb61287172cd14ba2 - I got the error "Page not found"). However, I was able to access the data through the OSF link associated with the preprint (https://osf.io/qymaz/ via https://psyarxiv.com/4gyzh). This error might be because the PCI RR link is 'view-only', but I mention it because it might need fixing in the PCI-RR portal.

Thank you for pointing out this issue. We are not sure how to fix that, but we are happy to provide any assistance for the PCI-RR team.

We believe the correct link is the following: https://osf.io/qymaz/.

Major points

My main points for revision center on:

1. Clarifying the "Hypotheses and Data Analysis Strategy"

I am confused by the sub-sections of "Statistical framework" and "Bayes Factor Design Analysis" as they both mention information about the bayes factor cut-offs for evidence. I am not sure which sections relate to which specific analyses (i.e., in the "Statistical framework" section it states that "BF values above 10 and below 1/10 were regarded as strong evidence for the alternative and the null hypothesis, respectively" but then in the "Bayes Factor Design Analysis" section it states that "The long-term rates of correct evidence were calculated as the proportion of iterations where strong evidence (BF > 10) was found for the existence of the effect. The long-term rates of misleading evidence were computed as the proportion of iterations where the evidence strongly supported the null hypothesis (BF < 10)". Can this be clarified?

Thank you for the question. The "statistical framework" part explains the main principles of how we interpret the results, so this part explains the cut-offs also for Bayes Factors. The "Bayes Factor Design Analysis" section does not belong here, as it basically explains in detail how we conducted a bayesian power analysis. "The Bayesian Factor Design Analysis (BFDA) is an alternative of the frequentist power analyses which enables researchers to estimate the informativeness of the study in a Bayesian framework." To calculate "bayesian power" one needs to also add the evidence threshold, so we need to speak also in this section about the cut-off points, but the two sections cannot be merged.

To help the readers to get a clearer understanding of these, we have now edited the BFDA analysis section.

It's also very confusing to then have another sub-section titled "calculation of BFs" with additional information below another sub-section of "specification of the models" – can you group all of the information on Bayes factors together in a more reader-friendly way?

Thank you for the suggestion! We have now moved the calculation of BFs section just after the statistical framework section and before the BFDA analysis section and we have also slightly edited the sections to remove redundancies and to make them more reader-friendly.

On a similar point, you continuously refer the reader to "as described below" and "as described above" which means the reader has to scroll up and down the manuscript to remind themselves: a better structure would be to avoid such language and explain this immediately.

We have now reduced the number of instances with such 'described above/below" references and only included the term where we think they are necessary and helpful for the reader.

2. Stage 2 Review Criteria 2A. Whether the data are able to test the authors' proposed hypotheses (or answer the proposed research question) by passing the approved outcome-neutral criteria, such as absence of floor and ceiling effects or success of positive controls or other quality checks. This criterion addresses whether the data quality is sufficient to be able to test the stated hypotheses, according to the pre-specified conditions in 1E.

On Page 25, you suggest that your data showed no sign of ceiling or floor effects ("Finally, the fact that the cognitive function measures were administered as part of a 90 minutes long questionnaire, could have exhausted the participants leading to floor effects. However, our data showed no sign of ceiling or floor effects". Can you please extend the reason why so that you explicitly meet review criteria 2A. This could be, for example, by proving the range of scores for the executive functioning tasks between the two groups.

As we stated in our original plan, "to ensure that we did not include executive function measures with ceiling and floor effects, in the Stage 1 report, we planned to exclude any of the measures from the calculation of the executive function index and hence from the primary analysis where more than 60% of the individuals achieve perfect scoring or zero correct answers in the given test." Following this plan, we didn't find any executive function measure that meet the exclusion criteria for the ceiling and floor effects. Now, following the reviewers' suggestion, to further ensure the interpretability of our results, we also provide the range of scores for each executive functioning task in the appendix (see Table A1 and A2).

This should also be clarified in Figure 1 by giving the total range that the axis can go to. By including the minimum and maximum range on this figure, you will also demonstrate visually that this was a small, non-significant effect between the two groups.

We are not sure if we understand this suggestion correctly. Figure 1 (using the updated numbering Figure 2 now), shows the standardized scores for which the minimum / maximum values are shown in the picture, and for which the minimum / maximum value

is also a function of the other values due to the standardization process. However, now we have calculated and added the non-standardized minimum / maximum values in the Appendix, so it can be checked as well (see Table A1 and A2).

3. Stage 2 Review Criteria 2E. Whether the authors' conclusions are justified given the evidence. This criterion addresses whether the claims drawn by the authors in their conclusions (including in the Discussion, Abstract, and anywhere else in the paper) are warranted by the data or evidence in hand. Note that PCI RR recommendation decisions will never be based on the perceived importance, novelty, or conclusiveness of the results.

In general, I think the Discussion section needs to be more specific with regards to the interpretation of the results: whilst I understand that the results are inconclusive, the language when interpreting the results is very generic and rather ambiguous at times. Here are some ways in which you might achieve this:

On Page 24 you state: "While we cannot conclude with certainty how these differences add up and interact, we can make a few observations which can put our findings into context. First, while previously published studies used pre-post designs, here the findings are based on a randomized controlled trial which in general provides a clearer and less biased estimate of the true effect size". Here you could expand the final sentence to explain why an RTC provides a clearer and less biased estimate of the true effect size.

We have now expanded the sentence and provided a reference for those readers who need a more detailed explanation of the issue.

On Page 24 you state: "Second, although individuals participating in the study were extremely poor, they were relatively homogeneous and unusual along some of their demographics. This may have influenced the effect in some unknown way: they were all male, from Liberia, between the ages of 18 and 35, and selected to be engaged in high levels of antisocial behavior as well as poor and often homeless." In what precise way could this affect the results? Why would this make it inconclusive? Is there any previous research you could draw upon to suggest that these demographics might matter?

Thank you for the question! Here we do not say, that the listed demographic characteristics made the results inconclusive or necessarily had an effect on the results. We only say that the " contrast between our results and prior studies could be the consequence of some mix of differences", from which one is the characteristics of the population. However, we do not have data on whether these characteristics mattered and we would prefer not to speculate why it mattered if it did. We do highlight these characteristics to enable researchers and meta-analyst to form hypotheses on the factors potentially moderating the main effect in future research. To make this clearer, we have added a short paragraph about potential future research directions.

On Page 24 you state: "Third, we used paper and pencil or verbal versions of three different arrow tests, two different digit span tasks and a maze task to assess changes in cognitive functioning, while previous studies predominantly used computerized forms of cognitive control and intelligence tests. We cannot be sure how the tasks and the way we administered them impacted the results." Okay, but in what way may they have affected the results? Could pen-and-paper instruments be more noisy? Could there be experimenter error? This was something I'd mentioned in my Stage 1 review, to which you'd explained that due to the context

of the study it was not feasible to collect data using computerized means; you might want to reiterate this here.

In general, our answer to the questions above (in what way may they have affected the results? Could pen-and-paper instruments be more noisy? Could there be experimenter error?) would be very similar to the one above. We do not have data on whether and how these effects mattered and we would prefer not to speculate why it mattered if it mattered. But again, we have added a short sentence about the issue in the potential future research paragraph.

Page 24 you state: “Fourth, in the present study, participants were provided with a lump-sum cash of $200. It is an open question, how larger cash treatment or applying monthly installments instead of lump-sum money would have impacted the results. Previous results found that monthly installments vs. lump-sum money had differential effects on people’s behaviour”. In what way do these differences effect people’s behaviour? I recommend being more specific here.

Again, we do not have data on whether and how these effects mattered and we would prefer not to speculate why it mattered if it mattered. But referencing some prior work, we have added some specific hypotheses on how this might matter, added a short description of the issue in the future research paragraph.

Page 24, the Discussion states: “We observed a small, positive effect on executive functions both for the short (b = 0.13) and the long term (b = 0.08) toward the hypothesized direction, but the data provide inconclusive Bayesian evidence to support or reject the effectiveness of the intervention”. I would avoid mentioning the direction of the results given the inconclusiveness of the findings: you word this better in the Abstract by stating “Our main analysis revealed that cash transfers have a nonsignificant effect on cognitive performance both for the short (b = 0.13) and the long term (b = 0.08), but these observed effects are roughly four times smaller than prior non-randomized research suggested, and the evidence is inconclusive”.

We have removed the word positive from the relevant occurrences in the manuscript.

4. Does the manuscript adhere to TOP guidelines?

The data and analysis code are made openly available and adhere to TOP guidelines in this respect. However, I am unable to open the files “STYL_lr_reshuffled.dta” and “STYL_Final_real_data.dta” on my computer. It appears these are STATA datafiles, but I do not have access to this programme. I am also unable to open the file “analyse code.do”. Could these files be exported to an open-source programme and uploaded onto the OSF as additional files that follow the FAIR principles (https://www.go-fair.org/fair-principles/)? I also think it would be useful to have a data dictionary with the open data (e.g., it is difficult to know what ‘exclusion’ refers to without this being described). This is key for reproducibility.

We have now transformed the STATA-based dta. databases to csv and downloaded the corresponding data dictionary from STATA. Now all files are uploaded it the OSF project folder.

Other minor points:

1. The Abstract states: “Our main analysis revealed that cash transfers have a positive, nonsignificant effect on cognitive performance both for the short (b = 0.13) and the long term (b = 0.08), but these observed effects are roughly four times smaller than prior non-randomized research suggested, and the evidence is inconclusive (BFshort-term = 1.21, BFlong-term= 0.56).” I suggest removing the term ‘positive’ so it is clear that these findings were non-significant and small. I also suggest this phrasing throughout; the reason being is that readers sometimes pull out any information that might support their own preconceptions (i.e., that there was still a positive effect, yet it was non-significant).

Following the suggestions, we have now removed the word positive throughout the manuscript where it potentially conveyed equivocal meanings.

2. Page 3, Introduction: comma missing between cited Authors (“Mani Mullainathan”).

Corrected, thank you!

3. The overview on Page 11 could be written more clearly. It could simply state (suggested changes in purple font): “Furthermore, we planned to conduct two exploratory analyses: (1) a multiverse analysis to reveal the robustness and sensitivity of the results to different analytical choices (see “Robustness tests: multiverse approach”) and (2) a mediation analysis to understand the driving mechanism behind the observed effects in the primary analysis. The mediation analysis was planned for those cases where the primary analysis revealed strong support (BF >10) for the effect however we were unable to conduct this because we found no strong support for the effects in the primary analyses”.

Thanks, we have now improved the manuscript accordingly.

4. Page 14 states “Finally, we standardized the executive function index to make it comparable with other results.” Can you please clarify how this was standardized (i.e., was it z-scored)?

Yes, the values were z-scored. We have now added this information to the manuscript by the first appearance of standardization.

5. On Page 15, present tense is still used when it should now be past: “Accordingly, when calculating the BF, we will use”. Please check throughout.

Corrected.

6. Page 16, please refer to Figure 1 in the text.

Added.

7. Page 16 states “Accordingly, we conducted multiple versions of the Intent to-treat analyses specified in the primary analysis section with 6 alternative analytical specifications (with and without control variables x 3 different priors), across 14 alternatively processed datasets (2 exclusion criteria x 7 imputation method) predicting 14

different cognitive function measures." Can you add "as follows" to the end of this sentence to clarify to the reader that the specific information on these specifications comes next.

Added.

8. Page 17, "We repeated all the analysis with and without the control variables as specified in the primary analysis" – it would be helpful to specify these control variables in brackets rather than 'as specified in the primary analysis'. This would aid readability/understanding. Same with "we applied no exclusion criteria on individuals." – specify what this is briefly.

We have now added the list of control variables and improved the wording for the exclusion criteria part.

9. Page 17, remove the term 'new' from the following sentence as it seems like this wasn't planned, when it was in the Stage 1 protocol: "10 new measures of cognitive function". Perhaps 'additional' or 'alternative' would work better here.

We have changed 'new' to 'alternative'.

10. Page 17, comma included where it should not be: "First, we winsorized the continuous variables at the 99th percentile while we also excluded all the individuals, who did not achieve at least a 80% success rate in the arrow attention test". Please remove.

Removed.

11. Page 17, typo? "explorative by nature" should be "in nature"? Again on Page 17, "We repeated all the analysis" should be "analyses"?

Corrected.

12. Page 18, you state "but almost all of the specifications yielded 95% confidence intervals that included zero"; can you specify the number here rather than 'almost all'? 23 In addition on Page 20 you state "Using the planned or small priors, most Bayes Factors are between 10 and 1/10", again can you be more specific? Again, same wording on Page 23 which needs to be clarified.

We have added the % of specification at each of the suggested instances.

13. Page 18 states "The effect of cash transfer becomes larger and always positive when executive functions are assessed with arrow switching accuracy, digits forward accuracy, or digit span index, but become systematically smaller and mostly negative when measured by arrow switching RT, backward digits accuracy or maze task accuracy". Do you want to add this information to the Abstract to extend the description of the results?

Thank you for the suggestion, we have now extended the last sentence of the abstract accordingly. "However cognitive performance varied between the executive function measures, suggesting that cash transfers may affect the subcomponents of executive function differently."

14. Page 21 states “. However, similarly to the short-term results, the way the cognitive performance was assessed seem to matter. The impact of the cash transfer program was larger (mostly positive) when cognitive performance was assessed with arrow switching accuracy, digits forward accuracy, or digit span index, but was smaller (mostly negative) when measured by arrow inhibition accuracy, maze task accuracy, and maze task RT.”. Does this change any non-significant effects to significant effects? Please clarify exactly how this changes the results – it suggests that some effects are significant when cognitive performance was assessed with arrow switching accuracy, digits forward accuracy, or digit span index.

Thank you for the question. First of all, it is important to note that we do not talk about significant, or non-significant results in the bayesian framework we used in the present paper, so we cannot say whether it changed any non-significant effect to significant and vice-versa. But we agree with the reviewer that the language we used was ambiguous and could implicitly suggest that there was a change in the significance levels. To overcome this, we have now modified the corresponding part of the text as follows:

“However, similarly to the short-term results, the way the cognitive performance was assessed seems to matter. The estimated impact of the cash transfer program was mostly positive when cognitive performance was assessed with arrow switching accuracy, digits forward accuracy, or digit span index, but was mostly negative when measured by arrow inhibition accuracy, maze task accuracy, and maze task RT.”.

We would prefer not to make stronger conclusions from these exploratory results beyond reporting the descriptives as the present registered report was not created to answer these questions.

15. The figures have the wrong numbers: Figure 1 is on Page 16, so the figures on Page 23 should be Figures 2 and 3, respectively. Please correct throughout as this has created a ‘knock-on’ effect with all other figure numbers.

Thank you! We have corrected the Figure numbering now.

And finally, thank you again for reading our paper in such detail and making excellent suggestions on how to improve it.