Dear Dr. Chris Chambers,

We are pleased to submit a revision of our 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.

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:

First of all, I'd like to thank the authors for this interesting submission. The main topic of this RR is highly relevant, and the proposed analyses will provide very informative insights into the effectiveness (both short- and long-term) of a simple lump-sum cash transfer on the improvement of the cognitive performance of people living in poverty. I'm very sympathetic to the fact that the authors decided to submit this as a secondary RR (RR with existing data). I'd also like to highlight the rigor of the proposed analyses as well as the authors' transparency. Below, I'll try to provide several suggestions regarding the proposed analyses and will also depict some points which, in my opinion, require further clarification.

Analytic code

Given that the proposal is a secondary RR, it'd be very helpful if the authors provide the analytic code at this stage of the review process (ideally including simulated data). Furthermore, for the sake of transparency, as some of the authors have been involved in data collection/have already had access to data, I'd recommend preparing the script and analysing data using the blinded analyst approach, for instance: (1) The script will contain a random reshuffling of the "experimental condition" variable. (2) The analyst will not know which condition is experimental/control. (3) If needed, the analyst will do debugging. (4) Once everything runs, the code will be uploaded to the OSF project page. (5) The real condition codes will be revealed only afterwards.

Thank you for the comment. Following your suggestion and to ensure independence of the analysis from the results, we have asked researchers not involved with the present project but having access to the dataset to randomize the "experimental condition" variable, and send us this modified dataset. We have uploaded this modified dataset to the project's OSF folder.

Using this modified dataset, we have now created and uploaded the analysis codes to the OSF page of the project. You will find 3 separate codes there: 1) A STATA file which

includes the analyses described in the paper. 2) A Rmd. file which includes the functions we will use to calculate the Bayes Factors. The input for these BF calculations will be taken from the output of the STATA files. 3) An Rmd. file including the code for the mini meta-analysis.

Research questions and theoretical background
The research questions/hypotheses are clearly written and are supported by a solid theoretical background.

In the introduction, the authors cite a "seminal paper" by Mani et al. (2013). I highly recommend reading a commentary by Wicherts and Scholten (2013; doi:10.1126/science.1246680) who show that the evidence provided by Mani et al. might not be very robust (stated diplomatically).

Thank you for the suggestion. Whicherts and Scholet (2013) is a great commentary and its findings are one of the reasons we think our work is a necessary addition to the literature. We also cite the paper on page 5, when stating that "testing the effect of cash transfers in a randomized controlled study will allow us to provide a clearer, and less biased estimation on the treatment effects compared to previously published studies using pre-post designs".
We have now supplemented the intro with an additional reference to the Wicherts and Scholet (2013) paper: "Although Whicherts and Scholten (2013) raised concerns about the robustness of the results, these findings generated interest in the scientific and policy-making community… "

When describing exploratory analyses, the authors state they plan to test mediation models. Even though I like succinct introductions and am a proponent of a data-driven approach, I believe that a bit more space should be dedicated to the mediation models. For example, I can imagine how stress mediates the relationship between poverty and cognitive performance. However, I've got a bit harder time imagining how, say, wearing a weapon (point no. 2 in the conflict measure) affect cognitive performance. If you prefer not to extend the introduction, perhaps you could depict the (psychological) mechanisms behind the mediating models (or an example of a non-obvious one) in supplementary materials, but this is just a suggestion.

Thank you for the suggestion! As we will only conduct the mediation analysis if the primary analysis reveals strong support (BF > 10) for the effects, we have added the following text to the mediation analysis section:
"Worrying is thought to deprive cognitive functioning through intrusive thoughts' effect on mental bandwidths [15,18]. Sleeping rough [21,22], recent hunger [24,25] and depression [20] have been shown to have direct physiological effects on cognition. We hypothesized that conflict may impact cognition more indirectly, by increasing stress or by taxing attention through focusing mental resources on the object of conflict. The items of the conflict index aim to capture behavioral patterns which we assume to be highly correlated with the frequency or severity of conflicts our respondents are engaged in."

Participants and measures
The part is well-written, I've only some minor suggestions. Could the authors briefly explain why the participants in the experimental condition received exactly US$200? Is there any rationale (besides the practicalities – e.g., a trade-off between the sample size and budget) behind this decision? We're at Stage 1, but this point should be worth discussing in more detail once the results are in.

We have now added the following text to the manuscript explaining the reasoning behind the choice of distributing $200's:
"During the preparation of the project, we interviewed a group of local individuals about the start-up cost of a small enterprise estimating the range between $75 and $125. We also assumed that people have other spending pressures and precautionary saving motives. That, combined with our budget constraints is how the $200 was determined."

The process of data collection is nicely described, and the flowchart is informative. However, I'd just like to check - was there any attrition rate or did every participant who completed the baseline survey complete also all the follow-up surveys? Because, frankly, this seems highly Unlikely.

Thank you for pointing out that we have not discussed the attrition rates of the follow-up surveys. Blattman et al. (2017) discuss this issue in great detail in their Appendix and in Table A4.

We have now added the attrition rates for each round of the surveys to the Consort Flow Chart.

Furthermore, we have added the following summary to the 'Baseline and Follow-up surveys' section.

"As described in detail in Blattman et al. (2017, Appendix A3), the authors collected at least five close contacts and all known addresses of the participants and spent on average three to four days for locating respondents per survey to minimize attrition rates. The attrition rate of the overall endline survey was 7.6 percent, which is common in field experiments in developing countries [e.g.,29]. Most importantly, the response rates in all treatment arms were within 0.4 percent of the control group, while the joint significance tests including all baseline covariates yielded p = 0.328, suggesting that the attrition was unsystematic to the treatments."

It's not clear from the text – were the measures administered in a random order?

Thank you for the question.

We have also clarified that in the Baseline and Follow-up surveys section that the questions were always administered in the same order.

Furthermore, we have now added the following text to page 9: "The questions and the incentivized games and tests were always administered in the same order."

Even though the cognitive tasks were incentivized (motivating the participants to better performance), the participants could have been exhausted after the 90 minutes long questionnaire. Might be worth mentioning in the paper.

Thank you for pointing this out. We agree that respondents could have been exhausted after the long questionnaire. Accordingly, we plan to mention this potential limitation in the discussion section of the Stage 2 manuscript.

Just a minor comment but one of the items in the worrying index should be coded reversely.

Thank you again, we have now added this to the description.

Another minor suggestion – the authors state "To estimate the level of symptoms of depression in the participants…". Technically speaking, I'm not sure if it's appropriate to use the term "depression" or "depression symptoms" since some of the items don't correspond with the symptoms of depression as listed in the DSM-5 or ICD-11.

Thank you for the excellent question! These questions do not correspond indeed perfectly to the language used in DSM-5 or ICD-11. Here, instead, we measured the symptoms of depression based on a locally adapted instrument used previously with ex-combatant populations in Liberia (Blattman and Annan 2016). These questions constitute ways of asking about symptoms of depression "e.g., sad or downhearted for low mood" which are appropriate to the Liberian context and which utilize idioms that are common in Liberia when describing mental health and mood. We believe that this method of asking about depression is more culturally relevant and more valid than asking about symptoms using language and idiom which is common in North America or Europe and thus in the DSM-5/ICD-11. We have now supplemented the description of the Depression index with the following text:

"These items were based on previous work adapting depression measurement to local context using idioms which are common in Liberia when describing mental health and mood [43]."

Blattman, Christopher, and Jeannie Annan. 2016. "Can Employment Reduce Lawlessness and Rebellion? A Field Experiment with High-Risk Men in a Fragile State." American Political Science Review 110 (1): 1–17.

Power analysis

The authors write that "Although our design was not optimized to reliably detect the null effect, we calculated the rate of misleading evidence also with the assumption that the null hypothesis is true for each of our hypotheses. The results showed the rates of misleading evidence were < 1% both of the hypotheses as well". When I reproduced the code for power analysis ("sim.H0"), I obtained very different results. Specifically, the null hypothesis was supported only slightly above 20% of the time while the results were usually inconclusive

(above 75%). Perhaps I'm missing something, but the authors might want to check the code just to be sure.

We really appreciate the thorough review of the accompanying analysis code! Before going into the details, note that given your other comment on the potential publication bias, we have conducted a new mini-meta analysis which provided adjusted estimates for the presence of publication bias. (See a longer response on this matter by the corresponding comment). As a result, in the new version of the paper, we will use a lower prior (b=0.34) in the analyses which has also affected the results of the power analyses. Below, in our reply I will use the values stemming from these new analyses.

Based on our own re-review of the code, we believe that the BFDA analysis we applied is correct. As BFDA is not a common power analysis, let us provide a longer explanation of its output.

In row #29 of the analysis code, the code produces the power estimates of the BFDA analysis to detect the effect ("sim.H1"), given that there is an effect. In this case, in the output, the evidence for H1 (BF > 10) provides the correct inferences rate, while the evidence for H0 (BF < 0.1) provides the misleading rate of evidence. This part of the analysis code produces the main results used and reported in the paper. Accordingly, we write that "We found that given that the alternative hypotheses are true, with the parameters detailed above, the model provides correct inference in 82% and inconclusive results in 18% of the simulations for H1 and H2. The long-term rates of misleading were < 0.01% for both hypotheses. These rates suggest the proposed analysis yields correct inferences in more than 82% of the cases for all the hypotheses while incorrect inferences only in less than 0.01% of the cases." Note that this is equivalent to the output of an NHST power analysis. (As the priors both for H1 and H2 are the same, we only conducted one bayesian power analysis, as its results can be applied for both hypotheses.)
However, to make sure that we do not provide misleading evidence even if there is no effect, and as the BFDA also enables us to quantify power given the null effect is true, we calculated the rates of misleading evidence for this case as well. In row #36 of the analysis, the code produces the output of the BFDA analysis for detecting the null effect ("sim.H0") - this is the one which is referred to in the review above. In this case, in the output, the evidence for H1 (BF > 10) provides the misleading rate of evidence, while the evidence for H0 (BF < 0.1) provides the correct inference rate. Between these two values, you can find the inconclusive rate of evidence.This analysis shows that we produce misleading evidence in less than 1% of the cases. However, this data also suggests, as we also write in the paper that…" our design was not optimized to reliably detect the null effect": the analysis showed that given that there is no effect, our study would provide the correct inferences in just above 10% of the cases, while it would provide inconclusive evidence in almost 90% of the cases. Note however, that in our paper, we use extremely conservative inferences rates (BF values above 10 and below 1/10 will be regarded as strong evidence) for the BFs. If we applied BF cutoff values of 3 and ⅓, which is comparable to the standard p = 0.05 in NHST (Jeffreys, 1961), we would have very different results: we would have found correct 71% inference rates in 70.7% of the cases, inconclusive results in 27.5% of the cases and misleading inferences in 1.8% of the cases.

Jeffreys, H. (1961). The theory of probability (3rd ed.). Ox-ford, England: Oxford University Press.

Data cleaning and handling

The authors don't mention any data quality checks or screening for careless responders. Do the data contain such checks? Or were such checks unnecessary given the way the data were collected? Would it be possible for the authors to screen for such participants, say, by examining the longstrings (see Curan, 2016; 10.1016/j.jesp.2015.07.006)? I know that this request might look a bit tricky since the preceding survey and probably also the tasks were delivered verbally but having careless participants in the sample could substantially bias the results. Would be great if the authors consider some possibilities of detecting such Participants.

Thank you for articulating potential concerns regarding data quality. In the new version of the manuscript, following the reviewer's suggestion, we have added an additional exclusion criteria to the main analysis to maximize data quality.
We use the results of the arrow attention task to test the participants' lack of willingness, motivation, or basic ability to provide high-quality responses. (In the arrows attention task, participants were asked to state the direction of arrows presented on a piece of paper). That is, we will exclude all the individuals in the primary analysis, who did not achieve at least 80% success rate in the arrow attention test. Not being able to finish the attention arrow test can signal a general inability or lack of motivation to produce meaningful results.

To control for outliers, the authors aim to winsorize the continuous variables at the 99th percentile. I'm not sure that this step is necessary, especially if the values of the continuous variables are possible to obtain (e.g., a participant will score X on a cognitive task which is 3 SDs above the mean score of your sample). Of course, if there are improbable values (e.g., obvious typos), they should be removed, but at least the multiverse analysis could be performed also without winsorizing the data. The authors use the 60% threshold (either perfect scoring or zero correct answers) as the indicator of ceiling/floor effect. I suggest using also other thresholds (e.g., 50% and 70%) as a part of the multiverse analysis.

Thank you for the suggestion! We have changed a manuscript in a way that we will also conduct a multiverse analysis without winsorizing data, and we will also use different thresholds to indicate ceiling and floor effects. See the Multiverse section below for detailed description.

The authors plan to impute data in several ways (e.g., imputing median values, imputing the minimum value for unfound treated members and the maximum for unfound controls, etc.). Perhaps I'm missing something, but wouldn't it be both easier and technically more sound to perform a multiple imputation using a regression-based technique? For example, the authors could impute data using mice package and then fit the model using the brm_multiple function from brms package.

Thank you again for the question. In our manuscript, we suggest using 4 different ways to impute data: "1) imputing the median value; 2-4) imputing missing dependent variables for the treatment (control) group as the found treatment (control) mean plus (minus) 0.10, 0.25, or 1 SD of the found treatment (control) distribution". We would prefer not to change the current imputation methods, as this way, we also report bounds which we believe is preferable to simply reporting a central imputation (such as a regression-based imputation), and that way we would also keep the paper consistent and comparable to the original Blattman et al. (2017) paper.
Furthermore, after some consideration, in the new version of the paper we have removed the 5ths imputation method, because we believe that in a research design involving individual differences measures, the difference between the minimum value and the maximum value of the individual differences measure would be expected to be larger than any sensible treatment treatment effect. That way "imputing the minimum value for unfound treated members and the maximum for unfound controls" would lead to unrealistic patterns.

Mini meta-analysis and setting the priors
The authors derive their priors based on the mini meta-analysis. Unfortunately, I wasn't able to reproduce the calculations because the data are missing (the path to the dataset is a private file). Could the authors upload the data (or at least the effect sizes) to the OSF?

Again, we really appreciate the thorough review of the paper including the accompanying analysis code! The reason, the data to the Mani et al. (2013) database was missing because - unfortunately - it is not openly available, and it was acquired through email communication from the corresponding author. However, after receiving this first round of reviews, we have contacted the original authors who gave us permission to share the dataset with the reviewers for our study. You can find the data uploaded to the OSF page and the analysis code updated accordingly.

Furthermore, as publication bias is likely to be present also in this field, applying some corrections would be helpful. I'm not sure how well the state-of-art methods for correction of publication bias work on meta-analyses with so few effects, but perhaps the authors could check it out and (unless the published studies were either preregistered or RCTs) correct for publication bias. This could play a major role in determining their priors. Even though the authors plan to calculate Robustness regions for each BF using extreme priors (which is great), correcting for publication bias could lead to even more precise estimates in the main analysis.

We fully agree with the reviewer that we cannot be certain that we did not miss any studies investigating the same phenomena in a similar context. By meta-analyzing the previously published field studies providing causal evidence on the effect of poverty on cognitive functions, we aimed to go one step further from the common practice of using a default prior or taking the effect size from one specific study, but we did not aim to claim the lack of publication bias here. Accordingly, our original approach was to conduct an individual participant data meta-analysis, as it offered numerous advantages such as less information loss than meta-analyses based on the published

summary information on the results (Debray et al. 2015). We reported the results of this meta-regression in the previous version of the paper (b=0.44).
In the new version of the paper, we have applied a different approach and conducted a new meta-analysis using the *trimfill.meta* function of the *meta* package in R, as this method could adjust for the presence of potential publication bias. We have uploaded the code for this new analysis at the OSF page of the project. The new results controlling for publication bias provided lower effectsize estimates (b=0.34). In the new version of the paper, we will use b=0.34 as the prior in the Bayes Factor calculations. Note that we will also produce Robustness Regions for each BF using extremely low and high priors.

Debray, T. P., Moons, K. G., van Valkenhoef, G., Efthimiou, O., Hummel, N., Groenwold, R. H., ... & GetReal Methods Review Group. (2015). Get real in individual participant data (IPD) meta-analysis: a review of the methodology. Research synthesis methods, 6(4), 293-309.

Main analysis
Could the authors explain how exactly they aim to "merge the responses for the 2 and 5 weeks as well as for the 12 and 13"? This is a rather important step in the proposed analyses. I skimmed through the reference paper (Blatman et al., 2017; 10.1257/aer.20150503) and didn't find the details of the procedure (I might have missed it, though). Anyhow, for the sake of reproducibility, it'd be very helpful to describe the "merging" in sufficient detail.

Exploratory analysis suggestion
This is just a suggestion, but given the availability of the follow-up studies, it could be interesting to perform a latent change score modelling and take a look at the dynamics of the effect of lump-sum cash intervention on cognitive performance in short-term (baseline, 2 weeks follow-up, and 5 weeks follow-up).

Again, thank you for asking for the needed clarification. We have now added the following clarification to the corresponding section (pp. 10):

…"Once they agreed with the study terms, participants completed a baseline survey. The remaining four surveys took place 2 and 5 weeks (short-term), and then 12 and 13 months (long-term) after the cash distribution[5]. As the administration of multiple measurements at relatively short intervals have been argued to decrease noise and increase precision for key outcomes [29], the authors collected two data points both for the short-term and the long-term follow-ups. That is, the 2 and 5 weeks, and the 12 and 13 months follow-up surveys intended to measure the same underlying phenomenon, respectively. Accordingly, similarly to Blattman et al.11, we will merge the responses for the 2 and 5 weeks as well as for the 12 and 13 months surveys in our analyses by taking the average of the corresponding results...."

We would prefer not to pursue further exploratory analysis.

A minor note – it'd be useful to specify the types of effect sizes the authors aim to calculate and report for each analysis.

We have now added (p. 13) that we plan to report the standardized beta coefficients for each analysis.

Multiverse analysis
I'd like to appreciate the idea of performing the multiverse analysis – indeed, there are many researchers' degrees of freedom likely to influence the results. Besides the alternatives proposed by the authors, I've noticed some other choices that could be included in the multiverse analysis.

For example, the 60% threshold for the ceiling/floor effect could vary a bit by moving it to, say, 50% and 70%.

Thank you for the suggestion! We have now changed the multiverse analysis procedure in a way that we will report the analysis for each cognitive measure irrespectively to the percentage of individuals with perfect/zero scoring, but signal for each analysis the percentage of individuals scoring perfectly or not having correct answers. That way, instead of only signaling each cognitive function measure as non-sensitive if more than 60% of the participants achieve perfect scoring / zero correct answers, we will report the exact percentage of participants achieving zero and perfect-scoring on the given measure.

The thresholds for calculating the inverse efficiency index could also be altered.

Thank you for the suggestion. Here, we tried but could not identify other alternative thresholds in the literature which would be relevant for our purposes, so we would prefer not to include an additional threshold for calculating the inverse efficiency index. (As any additional variable exponentially increases the number of models we need to build and report in the multiverse analysis, we aim to stick with those which are the most relevant.) However, we are open to change that if the reviewer is aware of other types of calculation which would be important to integrate.

Perhaps it'd be also useful to try different priors in the multiverse analysis.

Thank you for the suggestion! We have now added two new sections to the multiverse analysis. Accordingly, we will repeat all the analyses with three different priors: the output of the meta-analysis (b=0.34), with the smallest (b=0.18) and the largest (b=0.79) effect sizes from the mini-meta analysis described in the manuscript.

Also, the authors aim to include a set of 13 covariates in the regression model. I wonder, is it necessary to control for all these variables, given they are trying to make causal inference and the data comes from an RCT (with participants being randomly allocated to the experimental/control group)? I think that, at least in the multiverse analysis, the model could be estimated without controlling for those variables.

We have now added a new section to the multiverse analysis. We will repeat all the analyses with and without the control variables.

Transparency
I believe that the authors should explicitly state that the present RR/paper is basically a secondary data analysis, as it might not be obvious from the footnote. Also, I mention it earlier but feels like I should emphasize it one more time – given the fact that some of the coauthors had access to data, please use the blinded analyst approach.

Thank you again for raising this point.

We have now added the following paragraph to the beginning of the Methods section: "In the present registered report, we will reanalyze the randomized controlled trial also described in Blattman et al. (2017). The present paper focuses on the effect of the cash intervention on cognitive functioning. At the time of the original publication of Blattman et al. (2017), the authors specifically did not hypothesize any change in cognitive function, and hence excluded it from their preregistration, and focused their paper on how therapy and unconditional cash transfers should affect criminal and violent behavior. "

As described above, we have created the attached analysis code on a blinded dataset.

Once again, I'd like to thank the authors for this submission, and I hope they'll find some of my suggestions useful. Looking forward to reading their responses and the revised version of the RR.
Best wishes,
Matus Adamkovic

Reviewer 2
The current study examines the effects of a poverty alleviating intervention on the cognitive performance of high-risk males in Liberia. In an RTC, participants are assigned to the intervention condition in which they receive a cash lump-sum (n = 251, equivalent to 300% of their income) or a control condition in which they do not receive this (n = 222). Cognitive functioning is then measured in the short and long-term on various performance measures.

There are several strengths to this proposal: (1) it is very well written and clear and tests an important research question with real-world applications; (2) the effect size estimate is based upon a mini meta-analysis and BDFA simulations, (3) there are contingencies in place for floor/ceiling effects and appropriate exclusion criteria, (4) and the sample size is sufficient to provide informative results coupled with a strong rationale for Bayesian evidence to inform the conclusions of the study. With this said, I do have several concerns regarding the theoretical rationale and hypotheses and the potential for methodological flexibility with task measurement indices. After outlining these main points in detail, I also outline some minor points that are easily clarified. In sum, I believe this research has merit and the current problems could be resolved in a revision.

Theory and Hypotheses

The study aims to investigate the effects of a poverty alleviation programme on cognitive functioning and hypothesises that this will improve cognitive performance in the short term (2-5 weeks) and long-term (12-13 months). The hypotheses are written very clearly and focus on aggregate cognitive functioning, as do the primary analyses of the associated tasks. However, I do have concern with aggregating cognitive performance in this way rather than assessing the sub-components of attention, inhibition, shifting, switching, and working memory separately, as well as with the broader theory and use of the term 'cognitive functioning', as I will now detail below:

Cognitive functioning is a general umbrella term, and I believe you are actually testing the sub-components of executive functioning (EF; see Diamond, 2013). The tasks you use within the Method measure attention, inhibition, shifting, switching, and working memory.

What is the theoretical rational for assessing 'cognitive functioning' collapsed across these tasks? Is there any research in the literature which suggests that these sub-components are impacted to varying degrees by poverty? To highlight my point, there is some work within the separate literature on 'stereotype threat' which suggests that the cognitive load imposed by negative stereotypes impacts the EF of updating more than shifting and inhibition (see Rydell et al., 2013). Whilst the phenomenon of stereotype threat is challenged, it does allow us to apply the same theoretical rationale to the current study - it is likely that all of your sub-components of 'cognitive performance' will be improved by the intervention, or is it more likely that this intervention will have a greater effect on some than others? I worry that lumping them all together in the primary analyses may obscure important findings. A greater theoretical rationale is required to explain why you are focusing on 'cognitive functioning/EF' in general rather than its sub-components; this also feeds into hypothesis generation.

Diamond, A. (2013). Executive functions. Annual Review of Psychology, 64, 135-168.

Rydell, R. J., Van Loo, K. J., & Boucher, K. L. (2014). Stereotype threat and executive functions: Which functions mediate different threat-related outcomes? Personality and Social Psychology Bulletin, 40(3), 377-390.

Thank you very much for the comment and questions. Before the submission, we have ourselves had discussions both about the label of the primary index as well as whether the primary analysis should be focused on aggregated or separated measures.

We fully agree with the reviewer, that most of our measures fit under the umbrella of executive functioning. The reason we originally name the index in the primary analysis as cognitive functioning measures, because we also included the Attention arrow tasks (here participants were asked to state the direction of arrows presented on a piece of paper) which task - we concluded - does not fit well into the definition of executive function as stated for example in Diamond, A. (p., 136. 2013): "to a family of top-down mental processes needed when you have to concentrate and pay attention, when going on automatic or relying on instinct or intuition would be ill-advised, insufficient, or impossible (Burgess & Simons 2005, Espy 2004, Miller & Cohen 2001). Using EFs is effortful; it is easier to continue doing what you have been doing than to change."

However, following the reviewers' suggestions, we have made several changes to make our paper more streamlined and precise:

1) **We have now removed the arrow attention task from our primary analysis. This was beneficial as we only have clear executive functions in the primary analysis. Furthermore, this was a necessary step, as -- following reviewer 1's suggestion to include more quality checks in the primary analysis, -- we used the arrow attention task as a quality check. Accordingly, in the new version of the paper, we exclude all the individuals from the primary analysis, who did not achieve at least an 80% success rate in the arrow attention test. Not being able to finish the arrow attention test can signal a general inability or lack of motivation to produce meaningful results in any of the additional cognitive function measures.**
2) **After removing the arrow attention task, we have renamed the index used in the primary analysis as 'executive function index' from 'cognitive performance index' as we focus our primary analysis on executive functions. We have updated the text in the manuscript accordingly. However, as we will include the Attention arrow tasks and the maze test in the multiverse analysis which are not executive function tests, we will still use the term cognitive performance when we refer to the overall results. Given that, and the fact that prior research on the topic was mostly carried out by economists and non-psychologists who called these terms under the umbrella of cognitive functions, we would also prefer to keep the title of the paper we originally suggested "Does alleviating poverty increase cognitive performance?".**

We also agree with the reviewer that we cannot be sure that poverty alleviations measures impact all the subcomponents of cognitive/executive functioning to the same intent. For this reason, we conduct separate analyses with all the subcomponent measures in the multiverse analysis which can reveal the potential differences. At the same time, we would still focus on the aggregated measure in the primary analysis for two main reasons.

1) **First and foremost, if there is any effect, the aggregated measure is the most sensitive way to find it. The short cognitive tests in this study were adapted from commonly used longer neuropsych subtests. These subtests are often administered with a number of other subtests together to get a better estimate of the person's underlying capacity (Groth-Marnat, 2009). One test is likely to be a noisy estimate of the underlying construct by itself.**
2) **Several authors suggest that these are subcomponents of EF and that they should be considered together as a construct e.g.,**
    - **Miyake, A. & Friedman, N. P. The Nature and Organization of Individual Differences in Executive Functions: Four General Conclusions. *Curr. Dir. Psychol. Sci.* 21, 8–14 (2012).**
    - **Miyake, A., Friedman, N. P., Rettinger, D. A., Shah, P. & Hegarty, M. How are visuospatial working memory, executive functioning, and spatial abilities related? A latent-variable analysis. *J. Exp. Psychol. Gen.*130, 621–640 (2001)**

So we would prefer to use the aggregated measure in our primary analysis.

However, again, we would like to emphasize that we agree with the reviewer that it can happen that different subcomponents have different effects. But for this reason, we

conduct different analyses with all the subcomponent measures separately in the multiverse analysis: "To further test the robustness and specificity of the findings in the primary analysis, we will report the results separately for the 6 cognitive function measures which comprised the aggregate measure and include 10 new measures of cognitive function as part of the multiverse analysis. " If there is a differential effect of scarcity on any of the subcomponents, we will find it here and discuss it in the conclusion of the paper.

Groth-Marnat (2009) Handbook of Psychological Assessment 5th Edition. Hoboken, NJ: John Wiley & Sons.

Methodology
Potential procedural flexibility
From the Appendix which outlines the task, it is clarified that reaction time indices for cognitive functioning is measured using a stopwatch (but this is not specified within the manuscript, see below). This measure has the potential to allow for procedural/methodological flexibility because of the researcher's control compared to programmed experimental tasks. The rationale for the equipment used should therefore be specified, even if this is due to technological resources or the environment in which the study was conducted (I understand the constraints posed by naturalistic settings, especially in rural settings such as for street youth in Liberia).

Thank you for the suggestions. We have now added the following description to the 'Baseline and Follow-up surveys' section: "The response time measures were administered using a stopwatch, as in the context of the study it was not feasible to collect data using computerized means."

Details aiding replication
I find myself reading between the lines a little to understand that this is a secondary data analysis of a study conducted originally by Blattmann et al. (2017). This is because a lot of the relevant information about this is appended within footnotes. I recommend transparently stating this within the sub-section "The process of the study"; in this sub-section, statements such as "in the original study" are included without reference to this original study, and this paragraph currently assumes some prior knowledge on behalf of the reader.

You could also clarify Figure 1 by highlighting the focus of the current study; this could be achieved by putting a box around the "cash" and "no treatment" arms, and subsequent follow-up surveys, and stating above the box "current study". This would aid the reader's understanding.

Thank you for the suggestions!

We have now added the following paragraph to the beginning of the Methods section: "In the present registered report, we will reanalyze the randomized controlled trial also described in Blattman et al. (2017). The present paper focuses on the effect of the cash intervention on cognitive functioning. At the time of the original publication of Blattman et al. (2017), the authors specifically did not hypothesize any change in cognitive

function, and hence excluded it from their preregistration, and focused their paper on how therapy and unconditional cash transfers should affect criminal and violent behavior. Cognitive functions were assessed to obtain an exhaustive list of baseline measures. The treatment effects on cognitive functioning were not previously analysed and published beyond the preliminary analyses on a small subset of outcomes (see Blattmann et al., 2017, Appendix D7)."

We have now edited the flowchart to improve its clarity.

Page 8, sub-section "The treatment: unconditional cash transfers". Can you provide greater detail of when the cash transfers were distributed, i.e. after the baseline measures of cognitive functioning. Was this given as a one-off cash lump sum?

We have now refined the wording of the corresponding section to further clarify the process used. The distribution of the cash transfers is detailed in 'The process of the study' section as follows:
"After being recruited and before being assigned to any of the conditions, participants answered a baseline survey. Next, participants were asked to draw chips blindly from a pouch which determined whether they were assigned to participating or not in therapy. Crucially, participants analyzed in the present study receiving no therapy, were not engaged further until the assignment of cash treatments. 10-11 weeks after the baseline survey, all participants were invited to a public draw in groups of 50 where the lump-sum US$200 grants were randomly drawn by a nonprofit organization (Global Communities). Instead of computerized randomization, personal draws were used in order to maximize trust and transparency among the participants. Four follow-up surveys were conducted 2 and 5 weeks, and 12 and 13 months after the cash randomization by a nonprofit research organization".

I have read through the Appendix on OSF which helpfully provides the materials used. This has clarified some aspects of the written Method section, allowing me to make some recommendations that would aid replication by independent teams. First, on Page 9 under "cognitive function assessment" it should be clarified that participants were asked to state VERBALLY the direction of the arrows (otherwise a reader may think they wrote down their answers). The same applies for the other tasks.

Thank you! We have now added these clarifications to the corresponding sections.

Second, and perhaps most importantly, you do not specify procedural details about each task – how many trials were there?

Thank you for the comment. We write the following descriptions about the number of trials by the corresponding tasks:

- **Arrow tasks: "In each version, participants were presented with a series of 32 black or white arrows pointing up or down in an oral setting."**
- **Digit spam task: "Participants were asked to repeat verbally the digits either in the same (forward-digits) or the reverse order (backwards-digits). In case at least one of the two sets of digits were correctly repeated by the participant, the instructor continued reading longer-sets of digits up to a maximum of nine**

digits. That is, the total number of repeated digits was dependent on the performance of the participant (minimum 2, maximum 16). "

- **Maze task: After completing a pilot trial "participants were asked to complete 3 mazes with increasing difficulty in the maze task."**

How was performance measured? (I know this is via a stopwatch from looking at the Appendix, but this is not specified within the manuscript), Are you looking at both RT and accuracy or only one of these, and why? Third, for the arrows attention task, you state that both the number of incorrect answers and the total time of completion were recorded, but you don't specify this for the other tasks. It would be best to specifically state which dependent variables you are analysing here. Amending this is a requirement for PCI RRs which asks reviewers to assess Is the protocol sufficiently detailed to enable replication by an expert in the field, and to close off sources of undisclosed procedural or analytic flexibility?

Thank you for the questions.

We have now added the following description to the 'Baseline and Follow-up surveys' section: "The response time measures were administered using a stopwatch, as in the context of the study it was not feasible to collect data using computerized means."

As for the total completion time, we state the following in "The arrow task" section: "In each version, participants were presented a series of 32 black or white arrows pointing up or down in an oral setting. Both the number of incorrect answers and the total time of completion were recorded."

We are specifically stating for each measure how we calculate the dependent variable for the given analysis.

For the primary analysis we say the following:
"We will use a general executive function index as the dependent variable in the primary analysis. We will calculate the executive function index for each participant by summing the standardized values of the following measures: accuracy scores (number of correctly repeated digits) in the forward and backward digit span tasks; response time (the average logarithmized completion time) in the arrow switching and arrow inhibition tasks; accuracy (number of correct answers) in the arrow switching and arrow inhibition tasks."

For the multiverse analysis we detail the following:

"To further test the robustness and specificity of the findings in the primary analysis, we will report the results separately for the 6 executive function measures which comprised the aggregate measure and include 10 new measures of cognitive function as part of the multiverse analysis. The additional cognitive performance measures will be calculated as follows:
Accuracy scores in the digit span tasks: **To calculate accuracy in the digit span task, we will use the number of correctly repeated digits for each participant both in the backward and in forward digit span tasks separately.**

Response time in the arrow tasks: To calculate accuracy in the arrow tasks, we will use the average logarithmized completion times for each participant in the arrow attention, arrow switching and arrow inhibition tasks separately.

Accuracy scores in the arrow tasks: To calculate accuracy in the arrow tasks, we will use the number of incorrect answers for each participant in the arrow attention, arrow switching and arrow inhibition tasks separately.

Digit span index: To calculate an overall accuracy in the working memory tasks, we will calculate the average number of incorrect responses per participant separately both for the backward and forward digit tasks and standardize the obtained values. A new digit span index will be calculated for each participant by adding the two standardized values.

Arrow attention response time: We will calculate the average logarithmized completion time for each individual in the arrow attention task.

Arrow attention accuracy: We will calculate the sum of correct answers for each individual in the arrow attention task.

Arrow task response time index: To estimate the overall response time across the arrow tasks, we will calculate the average logarithmized response time per participant separately for each arrow task and standardize the obtained values. A new arrow tasks response time index will be calculated for each participant by summing the standardized values.

Arrow task accuracy index: To estimate the overall accuracy across the arrow tasks, we will calculate the average accuracy per participant separately for each arrow task and standardize the obtained values. A new arrow tasks accuracy index will be calculated for each participant by summing the standardized values.

Inverse efficiency index for each arrow task separately: When measuring cognitive performance, the combination of speed and accuracy can increase the efficiency to detect the effects as it can account for a larger number or proportion of variance. However, to obtain unbiased results from these combined measures, the proportion of correct answers need to be over 90% in a given task and have high positive correlation between reaction time and accuracy. Accordingly, to increase the sensitivity of our analyses, we will calculate the inverse efficiency score for the arrow tasks, given that each of the tasks, the proportion of correct answers is over 90% and the Pearson correlation coefficient between reaction time and accuracy is higher than 0.6. For the arrow tasks that do not meet this criteria, we will only report the accuracy and response time.

Maze task (response time): We will calculate the average logarithmized response time for each individual on the maze tasks. Both the response time and the total number of mistakes were recorded.

Maze task accuracy: We will sum the number of mistakes each participant made in the maze tasks (reversed scoring).”

Can the measures used to assess the secondary outcomes of worry, sleep, depression etc. not be included in the Methods section rather than outlined in the Analysis Plan?

As the mediation analysis is only part of the secondary analysis, we would prefer to keep the description of these variables in the corresponding part of the Secondary Analysis. Especially, as we will only conduct mediation analysis for the associations where the primary analysis revealed strong support (BF >10) for the effect, it is possible that no analysis will be conducted using these variables at all. Although, if the Editor thinks that it would be necessary to move this part to the methods section, we would be open to do so.

There is a typo Page 8, line 13 which states “arrow taks” rather than “arrow tasks”.

Thank you! We have corrected it now.

Analysis Plan
I am not an expert in BDFA or multi-verse analyses, so am not able to comment on this aspect of the manuscript. The Bayes factors specified as ‘good enough evidence’ for the alternative and null hypothesis appear to be robust and well justified. There is a clear distinction between primary and secondary analyses.

I have looked through the following analysis code uploaded to OSF: “BDFA.Rmd” and “preliminary_effectsize.Rmd”. For the BDFA.Rmd code I have identified a potential (minor) error, as follows:
From the R code:
```{r}
mani_dat <- read_dta("D:/Dropbox/055_Scarcity meataanalysis/Data/Data for meta-analysis/Mani/Mani_Table1_Data.dta")
```
Should this be read_dat rather than read_dta?

We really appreciate the thorough review of the paper including the accompanying analysis code! The reason, this part of the code was not running properly because we didn’t have the permission to share the data of the Mani et al. (2013) paper. It was previously acquired through email communication from the corresponding author. However, after receiving this first round of reviews, we have contacted the authors of Mani et al. (2013) who gave us permission to share the dataset with the reviewers for our study. Now, you can find the data uploaded to the OSF page and the analysis code updated accordingly.

Potential limitations of the study
I have identified one limitation of this study which may impact the findings. The manuscript states “the cash transfers were unconditional and the final decision on how they would use the money was at the participants’ discretion.” and in another section states the sample were

“vulnerable participants with evident signs of homelessness and substance abuse”. Without knowing what participants spent this money on, but knowing that they had evident signs of substance abuse, how can we be sure that they did not spend the money on substances that would have a deleterious effect on cognitive functioning? With this in mind, it is possible that you may find null effects or effects opposite to your predictions, but you would not be able to pinpoint that this was the explanation. Is there any contingency or analysis you could put in place that would allow you to rule this out? If not, I would acknowledge this as a potential limitation in the Stage 1 manuscript.

Thank you for raising this excellent point which we haven’t considered before! Fortunately, Blattman et al. (2017) in their original paper collected information on the marijuana usage and hard drog usage of the participants in the 2-5 weeks and 12-13 months follow-up surveys (Table 6, p. 1190). Their results suggest that in the short term neither marijuana (b = - 0.017, SE= 0.036, p.adjusted = 0.928) nor hard drug (b =0.02, SE= 0.030, p.adjusted = 0.928) usage was significantly affected by the cash treatment. Similar results were found for the long-term follow-up both for marijuana (b =0.018, SE= 0.035, p.adjusted = 0.935) and for hard drugs (b =0.079, SE= 0.033, p.adjusted = 0.123). We suggest discussing these findings in the discussion of the Stage 2 manuscript.

Minor points
Abstract
It would be good to clarify in the Abstract (and in the main text) whether the lump sum was a one-off payment (given in only one month) [or did they receive this on multiple occasions, if so, how many?]. I also recommend including the sample size allocated to each of the two conditions.

We describe it in the abstract and also in the main text that the payment was a lump-sum unconditional cash transfer.

We have now also added the sample size of the conditions in the abstract as well.

The following sentence is a little confusing because it is not clear whether you are testing the effect of poverty on cognitive functioning, or whether this reinforces existing inequalities. I understand what you mean because I have read the manuscript, but this could be clarified. “In this registered report, we will investigate the impact of a poverty alleviation program on cognitive performance to test this effect”. It may be as simple as removing “to test this effect”.

Thank you! We have now removed it.

It would be useful to specify the mechanisms being investigated (worrying, sleep deprivation, mental-health, hunger, recent conflicts).

We have now added the following description to the mediation section:

Worrying is thought to deprive cognitive functioning through intrusive thoughts' effect on mental bandwidths [15,18]. Sleeping rough [21,22], recent hunger[24,25] and depression[20] have been shown to have direct physiological effects on cognition. We

hypothesized that conflict may impact cognition more indirectly, by increasing stress or by taxing attention through focusing mental resources on the object of conflict. The items of the conflict index aim to capture behavioral patterns which we assume to be highly correlated with the frequency or severity of conflicts our respondents are engaged in.

We also describe the potential mechanisms shortly in the introduction:

"There are several potential pathways through which poverty can impair cognitive performance in the short term. The context of poverty may tax cognitive capacity by introducing scarcity-related concerns or increased anxiety and stress [3,15,18–20]. Furthermore, individuals living in poverty are often sleep-deprived [21,22], and experience more pain [23], conflict [11] and acute hunger [24,25] which can also diminish their cognitive performance. Yet, some effects of poverty only harm cognitive performance over a longer time frame. Deprived access to different resources such as education, physical and mental health care [20,26] or high quality nutrition [27,28], have the potential to create enduring changes in cognitive functioning particularly when experienced during childhood. "

Introduction

Would it be best to have the hypotheses at the end of the Introduction rather than within the Hypotheses and Data Analysis Strategy? In the latter section, you can then simply refer back to these hypotheses.

At the end of the introduction, we outline the main research question, but we would prefer to keep the hypothesis in the Hypothesis and Data Analysis Section as we think the description is clearer and more streamlined this way. However, we can come up with a feasible solution if the Editor thinks that it is necessary to relocate this section.

I am confused by footnote 1 which relates to the sentence "The idea that unconditional cash transfers can enhance cognitive functioning seemed to be radical even a few years ago.1", The footnote states: "1Indeed at the time of the design of the present study, the authors did not expect an effect of the treatments on cognitive performance. Cognitive functions were assessed to obtain an exhaustive list of baseline measures". What changed your mind about this effect?

We would not state that we changed our mind about the existence/non-existence of the effect. But we have been convinced by previous publications that this is an important question to investigate, and we also realized that this dataset could provide a substantial addition to previous literature.

This needs to be clearer to clarify that you haven't yet looked at, or analysed, any part of your data. I would also state the date (if you recall) that the study was designed initially, so this is completely transparent. You could state something like "This is a secondary data analysis of [REFERENCE]. Indeed, at the time of the design of this study (MM-YY), the authors did not expect an effect of the treatments on cognitive performance. Cognitive functions were

assessed to obtain an exhaustive list of baseline measures. This data has therefore not yet been analysed".

We have now added the following paragraph to the beginning of the Methods section:

"In the present registered report, we will re-analyze the randomized controlled trial also described in Blattman et al.[11]. The present paper focuses on the effect of the cash intervention on cognitive functioning. At the time of the design (2009) and the original publication of Blattman et al. (2017), the authors specifically did not hypothesize any change in cognitive function, and hence excluded it from their preregistration, and the focused their paper on how therapy and unconditional cash transfers should affect criminal and violent behavior. Cognitive functions were assessed to obtain an exhaustive list of baseline measures. The treatment effects on cognitive functioning were not previously analysed and published beyond the preliminary analyses on a small subset of outcomes (see Blattmann et al., 2017, Appendix D7)."

On the Fig 1 flowchart, could you add when the cash was disseminated?

We have now added this information to the flowchart.

Methods:
Page 8 states "As a result of the randomization procedure, 22 percent of the participants were assigned to the control arm, and 25 percent into the cash only arm (as well as 28 percent into therapy only, and 25 percent into the joint treatment arm)". For clarity, can you specify how many participants were assigned to each condition in brackets (i.e. 22 percent [n = XXX] were…).

Thank you, we have now added this information to the manuscript.

Phases of implementation – were the 100 participants who were first included in the pilot phase compiled with the participants recruited thereafter? This could be stated explicitly.

Yes, they were. We have now added this information explicitly.

I hope these comments prove useful to you,
Dr Charlotte R. Pennington,
Aston University