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.

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).

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.

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.

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.

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

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

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.

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.

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.

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.

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.

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?

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.

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.

A minor note – it'd be useful to specify the types of effect sizes the authors aim to calculate and report 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%. The thresholds for calculating the inverse efficiency index could also be altered. Perhaps it'd be also useful to try different priors in the multiverse analysis.

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.

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).

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 co-authors had access to data, please use the blinded analyst approach.

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