Review of PCI-RR: Registered Report: Do individual differences in cognitive ability or personality predict noticing in inattentional blindness tasks?
Signed review by Ruben Arslan

In this Stage 1 Registered Report, the authors plan to collect data on individual differences in inattentional blindness, relating them to better understood individual differences in cognitive ability, personality and ADHD. I am very interested in the answers to their research questions. For me, it was the first time reading a Stage 1 RR writing in this choose-your-own-adventure style and I really liked the clarity of knowing how the results would be described. Many of the design features of the study are also very well-developed and seem to reflect that the authors thought deeply about this. There is a developing literature that relates individual differences in classic cognitive tasks to each other and to other traits, but the authors face an especially challenging situation because inattentional blindness measures require participants to be unaware that something unusual might occur. Therefore, they basically get only one trial per task, which makes it harder to reliably infer individual differences.

I think the manuscript is already strong. I have a few minor comments and some larger comments that would require more work if the authors are up for it. To be clear, I'd be very curious to read the results of the study as planned. I don't think these comments identify any crucial flaws, just a better use of resources. I feel that the study could be better designed to optimize its ability to accurately quantify associations between inattentional blindness propensity and other traits and to diagnose why manifest correlations are low (which is the expected result).

1. I've read a few papers in recent years (e.g. Frey et al. 2017, Eisenberg et al., 2019) that take a number of tasks taken to reflect e.g. risk preference or self regulation and relate them to self reports, behavior, etc. Their projects were easier, because the nature of inattentional blindness tasks limits the amount of retries you get/different tasks you can study. Still, their findings may be informative. Many such studies find not only that correlations between task behavior and self reports/real world behavior are weak, but also that correlations between tasks that nominally tap the same construct are weak, and that the retest reliability of these tasks is weak. (In many studies I've read, there is no attempt to quantify something like an internal consistency analogue. If there was, these may also be weak, but since there's no hope of doing that for inattentional blindness, let's ignore this).

If you now find weak .15 correlations between ADHD and IB, how will you interpret this? This wasn't clear to me even with the very detailed registered report. You do not seem to be only interested in the manifest correlation between that single event of noticing or not, rather your two tasks reflect your hope to get at a latent trait/propensity for inattentional blindness.
But if the reliability of your IB measure is really low (based on prior work you seem to expect ca. $\sqrt{.13}=.36$), you shouldn't expect a large correlation, as the expectable correlation is bounded by $\sqrt{(rel\_ADHD * rel\_IB)}$.
Currently, you mention reliability, but only plan to compute a correlation between two IB tasks. Of course, interpreting the square root of that correlation as reliability is not a very robust measure of the reliability with which you've tapped into the latent propensity. As you write, maybe one of the tasks is not a good measure of IB. So, could you deploy additional tasks? I'm convinced by your logic that you can at most risk doing two tasks per person with

adequate spacing between them. However, given the planned sample size/available budget I think you can do better than using the same two tasks for all people. It is already part of your design that you vary some of the surface features of the tasks (presumably ignorable) and the cognitive load. But you could also randomly have groups do different inattentional blindness tasks. In such a planned missingness design, the goal would be to have sufficient power to estimate the bivariate correlations between all pairs of tasks. You would be able to get a better assessment of which tasks cohere, you'd be in a better position to estimate reliability, and you'd have a stronger claim that your results generalize to the propensity for inattentional blindness rather than just behavior in a single task. It would also be easier to assess latent correlations between your predictors and IB, which would probably speak more directly to your research questions. This is under the assumption that there are several more tasks (well, even I know some) that you consider valid (I don't know about this).

2. In the planned results section, you mention that you expect that >90% of participants will notice. Given that there are design features under your control (such as cognitive load) which will reduce that percentage, I'd think you should strive to get a rate of approximately 50% to optimize your power to detect associations. Maybe that's not possible, I'm not knowledgeable about these tasks. But psychometrically, if you only have two items, you don't want them to be easy/have low discrimination.

3. Then: You are concerned that participants will be wary on the second inattentional blindness task. I think you're right to be concerned, although I have no idea what this will cost you in terms of reduced sample size and generalizability. Have you considered separating the two tasks into two ostensibly different studies on Prolific? You could, through the account of a different researcher, recruit only those who participated in the first part of the study (with the first task) and then run the second task in a new study. By separating the tasks across studies, you probably reduce the expectation for the second task. Naturally, if you choose to do this, you might have more dropout between studies. I find it hard to judge whether that'll be more dropout than dropout resulting from participants who tell you they expected another odd stimulus in the second task.
There would also be some time lag between the studies, but you *could* actually capitalize on this and estimate retest reliability.

4. The following point about your design also made me think about your recruitment.
> We used settings to automatically exclude for eligibility people who had completed any of our prior Prolific studies assessing inattentional blindness.
Is there reason to believe your lab is the only one studying attentional blindness on Prolific using the tasks? If you have reason to believe familiarity might be high (after all, some Prolific users have done thousands of studies) maybe you want to restrict your sample to users with a number of Prolific studies less than x under their belt.

5. In the manuscript, I currently don't see plans to report any estimates of reliability/measurement error for the cognitive ability measures or personality questionnaires. I am guessing this will be added. However, I'd find it more interesting to see latent correlations rather than only manifest correlations, especially given that some of the measures are fairly brief. You frame all your research questions as "predictions". Maybe I misunderstood this and you actually mean prediction about future responses in IB tasks

(then using only manifest variables might make sense), but I thought you're probably only using it in the statistical sense.

6. You plan to include two "Attention check" items among the survey items. The following exclusion rule is planned:
> For study 2, we excluded all survey data from participants who answered both attention-check items incorrectly, but we retained data from the inattentional blindness tasks for those participants.
I know these types of items are standard, but since you're interested in participants with ADHD, maybe a robustness check is in order to see whether associations with ADHD differ if you exclude people who failed the attention check?

7. Regarding your sample size justification based on Schönbrodt & Perugini, 2013: It's a lovely paper, but as has recently been pointed out to me, the use of "stable" to describe precision is quite unusual for readers who have not read that paper (and know that they define the "corridor of stability" as ±.1). Why not just report the precision with which you estimate correlations at N=1000? Or actually simulate the power you'll have to detect a realistic effect? I only did a quick simulation, but it seems to me that your power is not actually that high, at least if you do the planned Bonferroni correction with 13 predictors.

```
pvalues <- c()
N <- 1000
for (i in 1:10000) {
adhd <- rnorm(N)
latent <- 0.3 * adhd + 0.95 * rnorm(N)
task1 <- ifelse(latent + 1.5 * rnorm(N) > -2, 1, 0)
task2 <- ifelse(latent + 1.5 * rnorm(N) > -2, 1, 0)
pvalues <- c(pvalues, cor.test(adhd, task1)$p.value)
}
round(cor(cbind(adhd, latent, task1, task2)), 2)
>        adhd latent task1 task2
> adhd   1.00   0.26  0.07  0.08
> latent 0.26   1.00  0.36  0.37
> task1  0.07   0.36  1.00  0.11
> task2  0.08   0.37  0.11  1.00
mean(pvalues < .05/13)
> .67
mean(pvalues < .01)
> .78
```

Also, Bonferroni correction is too conservative. Many of your 13 predictors will be highly correlated with each other. For the Big Five, you do not even have a specific hypothesis, so I'm not sure whether you'd divide by 13 or 8. Either way: It is good that you use multiple indicators of cognitive ability, of absorption/distractibility etc. But it is not good if your more thorough assessment of individual differences effectively reduces power. You could use latent variable modeling and only report correlations with individual measures as a robustness check. Or you could use a different correction like Benjamini-Hochberg. Or you

could simply preregister an alpha of .01, which is my rough guess for how much false positive inflation you should expect given the measures you have.
Even with an alpha of .01, you only have 78% power for a latent correlation of .30. However, if you manage to modify your tasks to have a noticing rate of 50% (see point 2), you get 95% power with the same sample size. Or 92% power with a noticing rate of 70%.

```
pvalues <- c()
for (i in 1:10000) {
  adhd <- rnorm(N)
  latent <- 0.3 * adhd + 0.95 * rnorm(N)
  task1 <- ifelse(latent + 1.5 * rnorm(N) > 0, 1, 0)
  task2 <- ifelse(latent + 1.5 * rnorm(N) > 0, 1, 0)
  pvalues <- c(pvalues, cor.test(adhd, task1)$p.value)
}
round(cor(cbind(adhd, latent, task1, task2)), 2)
mean(pvalues < .01)
```

Okay, so these are my larger points on where I see room for improvement in the design. I hope this is helpful.

Some more minor points:

In the introduction you discuss effect sizes in terms of *r*. Is this point biserial? It's not explained. Does anyone find that intuitive to interpret for a group difference? You switch back to Cohen's d in your own Results section. I would find it easier to read if you were consistent or reported both at least occasionally.

P. 5 L. 24 and following: Report CIs for all correlations.
"closer to zero" -> report number.

P. 10 L 20: report median and max N

The Javascript implementations of the tasks do not implement frame synchronization as far as I can tell, though they do use requestAnimationFrame (which is good). Lack of frame synchro will presumably lead to somewhat variable presentation times as browsers determine how many frames to show for the requested duration depending on various factors. As far as I know, the impact will be slight and is negligible for non-psychophysics research, but given that it's only a single result per task per person and device differences would be confounded with individual differences, maybe it's worth considering updating the code to follow the state of the art. I haven't evaluated this in depth and just wanted to bring it to your attention.

There is a link in the MPQ Absorption scale paragraph that leads to a page only accessible by password. Also, the full items for the MPQ are part of your online supplement, so the link may at best mislead readers to think they may not see the items.

P. 19 L 18 "would *not* measure"

This prior small study investigating ADHD and inattentional blindness isn't cited, probably it should be:
Grossman, E. S., Hoffman, Y. S. G., Berger, I., & Zivotofsky, A. Z. (2015). Beating their chests: University students with ADHD demonstrate greater attentional abilities on an inattentional blindness paradigm. Neuropsychology, 29(6), 882–887. https://doi.org/10.1037/neu0000189

Many of the results that do not relate to your primary research question but rather to validating your procedures (e.g., intercorrelations between cognitive tests) could be reported in a supplement.

The k-fold cross-validation to find items associated with noticing is a nice touch. You should mention which Pseudo-R2 you'll report for the logistic regression. You could also do a Lasso regression or similar with all items to see how much variance all items can explain in cross-validation (e.g., loo_R2 in the brms package).