Dear Dr. Frischkorn,

Thank you for the thoughtful reviews of our Stage-1 registered report manuscript on individual differences in inattentional blindness. Below we have reproduced your action letter and the reviews in full, and below each comment we explain how we have revised the manuscript in response to the suggestions and feedback. The revision process necessitated an overhaul of parts of the results section due to the addition of a third task. We think the revised version more directly addresses our key questions and also is easier to follow as a result of these changes. Below we provide a summary of the substantive and structural changes, including some improvements that were not specifically suggested by the reviewers but that followed from our discussions after thinking about their comments. Below is a summary of the major changes, followed by our point-by-point responses.

Summary of major changes:

- We simplified and improved the wording of the numbered list of research questions and we added one for the divided attention trials (and added analyses of those to the text).

- Reviewer 1 made a number of helpful design/analysis suggestions. The biggest substantive change we made in response was to add a third inattentional blindness task to the protocol (rather than implementing planned missingness or a multi-session study - see responses below). Although it isn't one that has been used much in the literature, it is different enough from the other two (while still similar in "flavor" to the other cognitive tasks) that we think participants might not be suspicious of it in advance. The addition of a third task fundamentally changed how we evaluated both order effects and the reliability of noticing across measures. That led to a restructuring of the opening of the results section that we felt substantially improved the overall flow and clarity. It also meant cutting several figures and replacing them with different ones. We think those newer figures help as well. Finally, we included analyses to determine whether aggregating across the inattentional blindness tasks was justifiable and we added contingent analyses that will be in the manuscript if it is (and in the supplement if it's not).

- This change, coupled with our ruminations about how best to assess order effects, led us to make a slight change to the questioning following the critical trial and divided attention trial in each task. Rather than asking participants if they had experienced that task before (which can easily be misinterpreted given the similarity of these primary tasks to other experimental procedures), we instead ask participants directly if they were searching for an additional object in addition to performing the primary task. We then use responses to that question to exclude participant data from any inattentional blindness task in which they claimed to have been looking for another object to appear. This measure allows us to separate the effects of task order on task strategy from less important order effects such as fatigue.

- When thinking more about the aggregation process, we decided to eliminate the randomization to attend to the white shapes or black shapes in the sustained inattentional blindness task. That manipulation isn't of particular interest, but it either would have required us to calibrate noticing to be the same in both versions or to further break down the aggregated results (beyond separating

for the easy/hard task). Given that we're not manipulating similarity effects in this study, the differences in overall levels of noticing between attending to white or black aren't inherently interesting. This simplification means that the smallest sample size we would have for any of the primary analyses would be n=500.

- Reviewer 1 made a number of other suggestions, most of which required more minor changes to code or prose, and we made almost all of those.

- Reviewer 2 was concerned with our characterization of load theory. We hadn't intended for load theory to be a focus of this experiment. We were more using the word "load" as a shorthand, and what we actually were manipulating was primary task difficulty. We decided that the most straightforward way to address the reviewer's concerns was to avoid explicitly describing our task-difficulty manipulation as a manipulation of the construct of load. We had already described evidence for task difficulty effects, and sticking to that descriptor rather than making claims about cognitive load simplifies the exposition and doesn't require adding a more extensive review of load theory. This approach allows others to evaluate the implications of our task-difficulty results for load theory, which is not our primary interest (and as the reviewer notes, is "all over the place" in this literature anyway). We give more detail in our explanation below and discuss why that task difficulty manipulation might not map onto the load manipulations used by others (that sometimes involve a secondary task rather than variation of primary task difficulty).

- Reviewer 2 raised concerns about our choice to use a liberal noticing criterion. The criterion matters a lot if the goal is to make claims about processes occurring with or without awareness (which seems to be the reasoning behind the reviewer's comments). But we're not doing that here. Inattentional blindness tasks are not really ideal for making such claims anyway because they rely on a single-trial report of noticing and consequently can't separate sensitivity from response bias. In the context of examining individual differences or when exploring the effect of an experimental manipulation on noticing, we're looking for differences as a function of whether or not people *report* noticing the object. That is, we're looking for variations across conditions using a shared criterion rather than examining absolute levels of noticing. For that goal, the choice of criterion is somewhat arbitrary. But we think there are good reasons to favor adopting a more liberal criterion for noticing. Here are our reasons for that choice:

  The criterion determines which type of error we want to avoid: With a conservative criterion, some people who actually did notice an unexpected object will be treated as having missed it. With a liberal criterion, some people who did not see an unexpected object (but who claimed they did) will be counted as having seen it. That is, a more liberal criterion will mistake some false-positive responses for noticing whereas a conservative criterion will mistake some false negative responses for missing. The question is which is more consequential and likely. Recent work (see this preprint from Nartker et al (2024): https://www.biorxiv.org/content/10.1101/2024.05.18.593967v1) shows that people tend to be conservative in claiming to have noticed something unexpected. That is, they are relatively unlikely to err by falsely reporting that they saw something unexpected when nothing was present.

In other contexts, when participants were explicitly asked about an unexpected change that didn't actually happen, few participants falsely claimed to have seen it (see Simons et al, 2002). In short, participants tend to be conservative by default when asked if they saw something unexpected; they tend to say "yes" only when they are pretty sure that they did see something. Consequently, using a conservative noticing criterion will likely result in misclassifying more people as missers, whereas a liberal criterion likely will not introduce too many misclassifications as noticers.

In sustained inattentional blindness tasks, we have regularly checked for a difference in the pattern of results across conditions when applying liberal and conservative noticing criterion, and we almost never see a meaningful difference in the pattern. The conservative criterion tends to lower the overall "noticing" rates in all conditions without having much effect on the pattern of results across conditions. That's likely because the object is present long enough that if people notice that it's there, they generally have time to identify it fully. So, we don't think the criterion choice will matter much for that task (especially the easy counting version).

In contrast, for the transient task and the search task, the additional object is visible for only a fraction a second. That means people might well notice the presence of something but not have enough time to process it fully enough to recognize it. You can experience that for yourself by viewing the demos:

https://yifanding1.github.io/test/IB_transient_test/IB_transient.html
http://simonslab.com/mot/individDiffs_Cognitive/IB_Search/IB_search.html

Even though you know the object will appear, you may have the experience of seeing *something* appear at fixation but still not be sure what it was. In such cases, you might correctly say "yes" when asked if you saw something while selecting the wrong object on the forced-choice question. The conservative criterion will misclassify such participants as missers even though they were not inattentional blind. Given the brief presentations in these tasks and the default conservative reporting bias people generally have, we expect that pattern to be common (far more so than false reports when people didn't actually see something). We do not want to claim that participants were inattentional blind (failed to notice) when they actually did see an object—that will undermine our ability to examine differences between missers and notices.

For those reasons, we prefer to stick with the liberal criterion as our primary one to make sure we capture all cases in which someone did actually see the unexpected object. And, as we noted, we will discuss any discrepancies in the pattern of results between the liberal and conservative criterion and we will fully report the analyses with the conservative criterion in the supplement. The conservative criterion provides a robustness check to see the pattern of results is the same when shifting the emphasis to minimizing false positive errors rather than minimizing miss errors. We expect the pattern of associations with individual difference predictors to be robust in the sustained task (especially the easy counting task - the criterion hasn't matter much for noticing patterns across experimental conditions in our other work), but for the transient and search tasks, we expect the conservative criterion to result in including many participants who actually did see

something into the “misser” category, resulting in substantially lower noticing rates and perhaps obscuring some individual differences.

We have elaborated our justification for using the liberal criterion in this paragraphs in the manuscript:

> We treated a person as having noticed on the critical trial if they said “yes” when asked about the presence of an additional object. As a robustness check, we also analyzed the data using a more conservative noticing criterion of saying “yes” to having seen something and also correctly picking the shape they saw when given a forced choice but not requiring a correct forced choice. A liberal criterion counts someone who saw something but was unsure what it was (and guessed wrong) as having seen the unexpected object, whereas a conservative criterion treats that person as inattentionally blind when they actually had seen something. A liberal criterion does risk treating someone who falsely reported having seen an additional object as having noticed, but people tend default to claiming they did not see something when they are uncertain (e.g., see Nartker et al, 2024), so the risk of that sort of misclassification is lower. Although we use the liberal noticing criterion as our primary outcome measure for the inattentional blindness tasks, we report the full analyses using the conservative criterion in the supplement, and we note any discrepancies in the pattern of results in the text.

- Reviewer 2’s comments highlighted an issue with the conservative+ criterion that we hadn't fully thought through. We have revised how we discuss that exclusion rule and fully report analyses in the supplement (as we do for the more conservative noticing criterion). We feel it would be inappropriate to use that exclusion rule as part of our primary analysis, though, because primary task performance might be a factor associated with many of our individual difference measures. That is, some people might perform worse on the primary task because they find it hard (a cognitive difference) while still attempting to do the task (the key element of an inattentional blindness task). So, although we did not make that exclusion rule part of our primary noticing criterion, we did clarify the nature of the analysis when applying that exclusion rule and made clear that we would note any discrepancies in the text (it is another robustness check for the pattern of results).

- We did not add a full attention trial and we did not add controls for visual angle for reasons explained in our responses to Reviewer 2’s comments below.

- Where possible, we have tried to address the remaining comments from both reviewers (details below). We have provided a change-tracked version of the manuscript to show the revisions from our original submission, although it is a bit hard to read given the reorganization we undertook.

- The references we cite in our responses are appended to the end of this letter.

- Given that PCI isn’t taking submission of revisions between July 1 and September 1, we weren’t able to complete pilot testing for the search task to calibrate noticing levels to 50%. We do this

sort of calibration for all of our studies and it won't be a problem to get noticing rates into the right range before examining individual differences. In the text we note what might change from our demo version. Specifically, we likely will vary the number or types of ignored items in pilot testing to find parameter values that result in approximately 50% noticing.

We look forward to any further input and guidance.

Sincerely,
Dan, Yifan, Connor, and Brent

****
Action letter
****

Dear Dr. Simmons, dear Authors,

I have recieved thoughtful and comprehensive reviews from two colleagues with different backgrounds regarding the topic of your registered report "Do individual differences in cognitive ability or personality predict noticing in inattentional blindness tasks?". Both reviewers complimented the already strong manuscript, but still suggested several points where the study designed could be strenghtend and optimized or better justified. I generally agree with them and the points they are raising and suggest that you consider their comments and aim to address them or explain should you not see merit in adopting their suggestions.

Apart from that, I think you provided a strong registered report that will provide critical insight into individual differences in inattentional blindness and their relationship to personality and cognitive abilities.I hope that you choose to revise your report.

Kind regards,

Gidon Frischkorn

****
REVIEWER 1
****

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 $\sqrt{(rel\_ADHD * rel\_IB)}$.

> You're right. To the extent that the inattentional blindness measures are correlated, we will include an analysis of an aggregated noticing measure. Near the beginning of the results section, we now evaluate whether aggregation is justified (based on inter-task correlations

and alpha). We have added placeholder analyses and discussion for the prediction of the aggregated measure that will be included in the main text if aggregation is justified (and in the supplement otherwise). Note that we expect the correlations between tasks to be low for a variety of reasons, so we expect aggregation might not be justified. But we agree it is worth considering and reporting (at least in the supplement).

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).

We agree with the goal of including more tasks. After much discussion, we decided that it would be worth adding a third inattentional blindness task to the battery even if participants might become suspicious prior to completing it (and we adjusted the post-task questioning to assess that). We selected a visual-search based task (first used by Cartwright-Finch & Lavie, 2007) that we had developed for a separate study in our lab. To our knowledge, their study is the only one to use that particular task, but the procedures are distinct enough from the sustained and transient inattentional blindness tasks already in the protocol that we felt we could add it without making participants too suspicious in advance. From our examination of the literature, almost all of the other inattentional blindness tasks were either video based (less control), real-world (impossible for this study), or focused on a particular content domain (e.g., driving, radiology). The search task was superficially similar to the other sorts of cognitive tasks included in our battery, so it fit well. We will randomize the order of these three inattentional blindness tasks and will separate them as much as possible within each battery.

After adding a third task, we needed to restructure the results section to analyze order effects and to address the possibility that participants might start expecting an additional object. We also now describe the criteria necessary to justify creating an aggregate measure across the three inattentional blindness tasks. That meant consolidating and reordering some of the analysis of the inattentional blindness tasks and replacing some of the task-specific figures with new ones. Despite the complexities introduced by adding a third task, we think the revisions that were required actually make the results easier to follow (and better justify our approach).

We decided that adding a third inattentional blindness task for everyone and checking whether people reported deliberately searching for an additional object on each one was preferable to the idea of a planned missingness approach Ruben. We did not want to further subdivide our sample for tests of associations with the individual measures. We are expecting that noticing on these three tasks might not be strongly associated, meaning that forming an aggregate measure might not be justified. If not, we will need to examine associations with other measures individually for each inattentional blindness task.

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.

For the *critical trial* where the additional object is unexpected, we are aiming for a 50% noticing rate for exactly this reason. That's something we do in all of our inattentional blindness studies where we are interested in examining differences in noticing across conditions—it's essential to leave room to observe increases or decreases and to avoid ceiling/floor noticing levels. The 90% figure was for the *divided attention* trial where participants know an additional object might appear and presumably are devoting some of their attention to detecting it. We expect most people to notice on that trial.

These comments led us to revisit the literature to look at noticing rates in divided attention trials. Noticing rates can be lower on the divided attention trial, and we might expect performance on the cognitive tasks to be associated with noticing under divided-attention conditions. We still expect noticing to be substantially higher on the divided attention trial than the critical trial. We have rewritten the section discussing noticing rates on the divided attention trial (including dropping the 90% cutoff) and added a plan to analyze those data in the main manuscript (including adding a research question to the list of questions). Here is the new description (the analyses are on pages 38-39).

To make sure readers are not confused about our anticipated noticing levels, we added the following after summarizing the results of the separate tasks (p. 24) and the analyses appear on pages 38-40.

> After the critical trial in each inattentional blindness task, participants completed a divided attention trial with the same primary task except that they now knew that they could be asked about an additional object. Noticing rates often are substantially higher on the divided attention trial than on the critical trial [**If noticing on the divided attention trials were consistently greater than for the critical trials]**: ", and we observed that pattern as well. **If noticing rates on any of the divided attention trials were not higher than the corresponding critical trial:** ", but noticing in [all three tasks | name task(s)] unexpectedly was not higher than on the critical trial." [**If there were meaningful differences in the pattern of results for the divided attention trials across tasks, we will describe them here**.] We might expect performance on the divided attention trials to be associated with measures of cognitive ability because it requires participants to devote attention both to the primary task and to looking for an additional object. We examine that question after looking at individual differences in noticing on the critical trial. The supplement provides the noticing rates on the critical and divided attention trials separately for each of the unexpected objects.

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.

As noted above, after much discussion, we thought it better to include all three inattentional blindness tasks (the original two plus the added search task) in a single battery and to assess whether or not performance differed as a function of order. The first reason is a practical one: our existing IRB protocol does not allow us to track participant identity across separate Prolific studies, so we would need to submit a new protocol to do that (a process that could take months). The second reason is substantive: We are concerned that attrition might be associated with some of our individual difference measures (e.g., conscientiousness, diligence, etc.). We think examining order effects and excluding data from participants who

reported searching for an additional object on the critical trial is simpler and introduces fewer risks to our ability to interpret associations. It also would be useful for future research to know the consequences of embedding multiple, distinct inattentional blindness tasks in a single battery.

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.

There likely are other labs studying IB on Prolific, but we haven't seen many published papers that have (at least not in recent years). Given that we have obtained consistent, reliable results with our current inclusion criteria, we prefer to continue using them. In our revised procedures, we ask participants after the critical trial of each inattentional blindness task whether they were searching for an additional object. That provides a check on whether they recognized the task as an inattentional blindness task and were deliberately searching for the additional object (especially for the first inattentional blindness task they completed).

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.

Good point. In the method section, we cite reliability estimates from the published literature for each task (when they exist) and we also added an "observed alpha / omega" column to the table of descriptive statistics for the personality measures.

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.

We went through the manuscript to make sure that any use of the word prediction was specific to a hypothesized outcome for these measures and not referring to a downstream outcome or dependent variable (at some future point).

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?

> In practice, those of us who have conducted individual difference studies using Prolific in the past (Brent Roberts) have not found substantial differences in performance on these attention checks as a function of factors like ADHD. Still, we have added the following contingent analysis text to the results section:
>
> > **[If any data were excluded due to the attention checks in the survey items, include the following paragraph]** Recall that we excluded data from participants in Study 2 who answered both attention check items incorrectly. Given that performance on those attention check items might be associated with individual differences in ADHD, as a robustness check, we examined whether including those participants would affect the association between the ASRS inattention scale and noticing. The strength of the association with noticing was [about the same | weaker | stronger **where about the same is within ±0.10**] when including those additional participants (transient: r=xx; sustained-easy: r=xx; sustained-hard: r=xx; search: r=xx).

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?

> Thanks for this suggestion. We computed the precision for point-biserial correlations with different sample sizes and population correlations and now report that information instead of the “corridor of stability.” The section now reads as follows:
>
> > Because we will be examining individual difference correlations with performance on those conditions, we also assessed the precision with which we could measure point-biserial correlations of different magnitudes as a function of sample size (see the osf project for the code used in these calculations and provides estimates for other correlation values). For a true correlation of r=0, we expect 95% of correlations with a

> sample size of n=500 to be smaller than r=±0.088 (for n=1000: ± 0.062; for n=2000: 0.044). The precision of measurement increases with larger correlations. With a true population correlation of r = 0.80, the expected sample correlations would fall within ±0.026, ±0.018, and ±0.013 of r = 0.80 for sample sizes of 500, 1000, and 2000, respectively. For most of the individual differences associations in our study, we targeted a sample size of n=1000, but even with our smallest target sample size for an individual difference association (n=500), we can estimate correlations precisely: If there truly is no correlation between noticing and an individual difference measure, with n=500, we would only observe correlations larger than r = 0.09 about 5% of the time and we would observe correlations larger than r = 0.12 only 1% of the time.
>
> In short, we chose a sample size that would provide a more precise estimate of individual differences than any previous study of any individual difference predictor of inattentional blindness, most of which tested small numbers of participants (median n=44 per between-groups sample and only two studies had n > 200; the maximum sample size was n=554 for a study of personality differences; our total sample for personality measures is substantially larger than that maximum sample size; Simons et al., 2024).

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)
```

> See our response to your point #2 above. We are aiming for 50% noticing for the critical trial (the 90% figure was an estimate for reported noticing on the *divided attention* trial, although we've changed the language about that trial too—see comments above). As for p value corrections, we decided that the better approach would be to maintain our focus on estimating effect sizes rather than using NHST. Consequently, we removed the t-tests from the results table and instead report the point-biserial correlations and confidence intervals around them. Also see our response above about the precision of our estimates.

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

> Thank you for these constructive and helpful suggestions—we think they will lead to a more informative project.

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.

> We agree that consistency would be better, and as noted above, we removed the t-tests and now report the point-biserial correlation and its confidence interval (and explain what we're doing) for associations between each measure and noticing. Given that most individual difference studies use r as an effect size, we decided to stick with r throughout.

P. 5 L. 24 and following: Report CIs for all correlations.

> We assume you meant the confidence intervals for the meta-analytic estimates and we added those. If you also wanted confidence intervals for the individual study results (that went into the meta-analysis), we can compute and add those too, but the meta-analytic estimates were our focal ones in the paper we cited (and we now give the n with each r, so it's easy enough to infer how wide those intervals would be for the individual samples too).

"closer to zero" -> report number.

> In that analysis, we had conducted several different bias corrections: trim & fill, limit meta-analysis, and Bayesian meta-analysis. We now report the estimates for all three in parentheses: Trim and Fill: r = -0.001; limit meta-analysis: r = -0.012; Bayesian meta-analysis: d = 0.002 [-0.205; 0.189]. We retained the 'closer to zero' descriptor to make sure readers grok the key point that the bias-corrected estimate is tiny.

P. 10 L 20: report median and max N

> Done.

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.

It appears that requestAnimationFrame largely does handle the frame synchro issue well enough for this sort of purpose. However, we now also implement "delta timing" which accounts for the elapsed time between each frame and adjusts the animation accordingly. Specifically, we modified the `runAnimLoop` function to use delta timing and used the timestamp parameter provided by `requestAnimationFrame` to calculate the elapsed time (in milliseconds since the last frame). That should improve synchronization across browsers, although it likely doesn't matter much in practice given that none of our tasks require millisecond-precise timing. If Ruben had something else in mind for improved timing, we'd be happy to hear about it and to try to implement it.

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.

The link was faulty and we have corrected it. We were linking to the source for the MPQ. The parenthetical now reads: "(for more information about the MPQ, see https://www.upress.umn.edu/test-division/mpq/)"

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

The typo was fixed as part of our overhaul at the beginning of the results section.

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

We added the following to the description of the ASRS inattention task:

> "One small study (Grossman et al., 2015) observed less inattentional blindness with the "monkey business illusion" video (Simons, 2010) among 14 college students with ADHD than among 18 students without ADHD, although the paper did not report controlling for differences in prior familiarity with that or related videos."

Note that the lack of control for prior familiarity is a big concern given their tiny sample sizes and the use of a relatively well known video. It might not be surprising if there were a difference in familiarity with that video or with the earlier "gorilla" video as a function of group, and that might contribute to the observed difference (i.e., perhaps those with ADHD spend more time on YouTube and were more likely to run across the gorilla video).

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.

> We agree that the sections on expected associations among cognitive/personality measures disrupted the flow of the paper. Given that these validation checks provide a sort of positive control to show that we can detect associations when they should be present, we wanted to keep them in the paper rather than in a supplement, so we moved them to an appendix.

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).

> We now explain that we're using Tjur's pseudo R^2 measure when we first introduce the logistic regression analyses: "(the reported pseudo R^2 is Tjur's coefficient of determination; Tjur, 2009)"
>
> We added reporting of a loo-adjusted R^2 estimate for the section where we use cross validation to see if we can identify a set of survey items that predict noticing:
>
> > We also computed a loo-adjusted R^2 estimate (using the brms R package) for each inattentional blindness outcome measure to determine how much variance all of the items can explain.

****
REVIEWER 2
****

The authors propose to undertake a large online study across two separate samples to evaluate individual differences in inattentional blindness (IB). The first will look at cognitive measures (Ospan, rotation, TestMyBrain matrices test) and the second will look at personality scales (BFI-2 for big 5 personality, MPQ Absorption scale, ASRS - ADHD self-report scale, and the FFOCI - five factor obsessive-compulsive inventory specifically fastidiousness, perfectionism, and punctiliousness), with the goal of assessing whether individual differences predict differences in IB. The authors propose to use two IB tasks ("sustained" IB, "transient" IB) in both experiments so that they can assess differences in the most classic IB experimental conditions. This also allows them to examine IB in one task as a predictor of IB in another task. I applaud the authors for the overall methods and design, which appear to be very robust and well

thought out. The analysis plan is similarly well conceived. The authors have used their recent meta-analysis as primary motivation and for generating hypotheses for their proposed RR, which from my perspective is a robust starting position and justification of a study. Taking their meta-analytic findings at face value, and with an aim to collect 1000 subjects per experiment, this study will undoubtedly gain important insights and contribute well to the literature.

Below I have outlined some concerns I have, both major and minor. I consider my expertise to lay within attention and consciousness, so my comments and recommendations are largely drawn toward the IB aspect of the RR (not so much the other individual difference measures). Major concerns are mostly methodological and some conceptual. Minor are mostly to do with the presentation of the manuscript and some possible errors I noted.

Major

Measure of awareness and added robustness checks

The authors propose to use as their primary outcome measure a "liberal" criterion for classifying a subject as a noticer, and then add in two 'robustness' checks ("conservative" and "conservative+"). In principle I am very much in favor of this idea, but I mostly disagree with the way it is done here.

> See our detailed explanation at the beginning of this letter for why we chose to use the liberal criterion as our primary measure in this study. That explanation addresses our reasoning and response to most of the reviewer's comments on this issue, but to further clarify our reasoning, we have added comments below in response to specific aspects of the reviewer's comments. We hope that these additional explanations help explain why we do not think the conservative criterion is a better option here and why we think the discussion of awareness is not directly germain to our research questions.

The proposed liberal criterion is as follows: simply categorize those that respond "yes, I saw something" as a noticer. With such a loose criterion, there is no identification/recognition of the unexpected object as 'proof' that the subject actually saw it. The experimenters are therefore saying they will take the subject at their word, rather than requiring evidence that they indeed had the perceptual experience in question. This is problematic for multiple reasons:

1. We know that differences in response criterion can impact and confound our measures in studies of consciousness (meaning that some subjects are more likely to say yes, even if they did not have the corresponding experience, and vice versa, which may undermine evidence for so-called 'unconscious' processing, see Yaron et al., 2024). One limitation of the standard IB task is that there is not much of a way of checking for this (but we are working on this currently).

So, with this limitation in mind, it is generally better to take a more conservative measure, just to be sure.

> It's true, of course, that response criteria matters a lot when attempting to make claims about the complete absence of awareness (Dan S. has published on such claims of implicit perception in the subliminal perception field). It's also true that inattentional blindness tasks are a lousy way in general to measure "implicit" perception or to validate whether something was entirely outside of awareness because there is no way to distinguish sensitivity from bias. That said, our primary goal in this study is not to make claims about what is happening without any awareness. And, the choice of criterion is less relevant when examining how experimental manipulations affect noticing rates or when looking at whether individual differences are associated with reports of noticing (when the same criterion is applied consistently).
>
> More generally, a more conservative criterion for noticing is the same thing as a more liberal criterion for missing, so there's always a tradeoff involved in determining which is more relevant/important (see explanations at the beginning of the letter).

2. With such a liberal criterion, it is left unclear how the additional data will be treated. For example, what if a subject says "yes, I saw a stimulus" but were incorrect in the forced choice? What would be done then? Will these subjects be treated as equal to those who were correct in the forced choice (all classified as noticers)? To me, that would be inappropriate. As another example, what if a subject says "yes" on both the critical trial and following divided attention trial, but is incorrect in the forced choice task on one and is correct on the other? From the current proposal, it seems these would be treated as equivalent, which is problematic.

> Using a liberal criterion means that we might treat a case that's a false positive detection as if it were noticing. Using a more conservative criterion, we will "miss" cases in which people really did see something but didn't perceive/remember exactly what it was. It's a tradeoff either way. And given that people tend to be conservative about claiming they saw unexpected objects, only saying they saw something when they are fairly sure, we think counting inaccurate identification in the forced choice as evidence of a "miss" will lead to more misclassifications of participants (see the detailed comments to the editor).

There is both theoretical/principled reasons and empirical data to suggest non-trivial differences between different outcome measures of awareness, for example those that rely on mere detection (ie yes/no) compared with those that require some higher order identification/recognition of the stimulus (see Koivisto et al. 2017; Persuh, 2018). So I appreciate that a balance between these competing issues is needed, but I think that an IB study is probably the wrong place to examine

this particular issue, and more to the point am left concerned that this 'liberal' metric is too liberal.

> We agree that inattentional blindness tasks are suboptimal when examining implicit perception (i.e., what happens in the absence of any awareness) because there is no way to separate a person's sensitivity from their response bias in a single trial. Fortunately, that's not what we're doing. Again, we think that misclassifying people as missers is a bigger concern in this context than misclassifying them as noticers, hence our preference for the more liberal metric.

This is particularly so since it is their primary measure (see page 19, line 38-40: "For all correlations of performance on this task with other measures and for attempts to predict noticing of unexpected objects from other measures, we used the liberal noticing criterion on the critical trial"). At its worst, then, this has the potential to undermine a great deal of the results. It could also even undermine the motivation behind the research, since this is motivated from previous studies that have looked at similar issues, and yet such a loose criterion is not common in these prior works (to my knowledge).

> We don't follow the logic here. The liberal or conservative criterion can't be inherently better or worse. It depends on whether you want to be conservative about noticing or conservative about missing. We think the conservative criterion is less ideal in this context (especially for the tasks with only brief presentations of the unexpected object) because it will misclassify more people as missers. As noted in the response at the start of this letter, if you try the demos for yourself, you might well have the experience of seeing something at fixation and not being able to identify it fully. Those are cases we want to count as "noticing," not missing.
>
> More generally, we don't see how this choice of criterion can "undermine a great deal of the results" or "undermine the motivation behind the research." We think these comments might be addressing concerns that apply to claims about implicit processing—what is processed in the absence of awareness. For such claims, ruling out any awareness is essential. And for that purpose, inattentional blindness tasks are a poor choice because the lack of a report of "noticing" on a single trial does not mean the absence of all awareness regardless of the criterion. But that's not what we're doing in this study.

In addition to the above, can the authors elaborate on the choice of the "conservative+" criterion? Since no check of how subjects perform on the distractor task as a possible exclusion is proposed, I take it the rationale for this added robustness check is such an assessment. I thought it was odd to see it as a robustness check in the measure of noticing and am not confident this is adequate.

Perhaps the authors can consider placing it more central to their pipeline, such as by using it in some manner as part of their exclusionary criteria? In a standard IB study this is quite critical because it rules out that, if a subject is inattentionally blind, it is not simply because they are not performing (or cannot perform) the task. I can see some rationale for not needing this in this proposed study since there are other cognitive tasks being undertaken (and so performance on these could in principle be used for a similar purpose), but what about experiment 2 which will only use personality measures? There is also the possibility that authors may wish to retain variance given the research question involves individual differences. Either way, I believe it is worth elaborating on/justifying further.

> After giving more thought to our conservative+ criterion, we agree that it is not really a noticing criterion. As a noticing criterion, inaccurate performance on the primary task would have been treated as a "miss" of the additional object, which is not ideal given that it is not directly tied to the noticing response. We also are reluctant to use this exclusion rule for the primary analyses given that, as the reviewer surmised, performance on the primary task might well be a source of variance tied to other cognitive tasks. We have rewritten that section and now include a full analysis after applying this exclusion criterion in the supplement as an additional robustness check. And, we note that we will report any deviations in the pattern of results in the main text.
>
> In the manuscript, we have expanded our rationale for our choice of a noticing criterion. We have duplicated the revised paragraph here, along with the expanded paragraph on primary-task accuracy exclusions:
>
> > We treated a person as having noticed on the critical trial if they said "yes" when asked about the presence of an additional object. As a robustness check, we also analyzed the data using a more conservative noticing criterion of saying "yes" to having seen something and also correctly picking the shape they saw when given a forced choice but not requiring a correct forced choice. A liberal criterion counts someone who saw something but was unsure what it was (and guessed wrong) as having seen the unexpected object, whereas a conservative criterion treats that person as inattentionally blind when they actually had seen something. A liberal criterion does risk treating someone who falsely reported having seen an additional object as having noticed, but people tend default to claiming they did not see something when they are uncertain (e.g., see Nartker et al, 2024), so the risk of that sort of misclassification is lower. Although we use the liberal noticing criterion as our primary outcome measure for the inattentional blindness tasks, we report the full analyses using the conservative criterion in the supplement, and we note any discrepancies in the pattern of results in the text.

As a further robustness check, we excluded participants who had poor accuracy on the primary task trials prior to the critical trial. For the transient inattentional blindness task, we computed the proportion of correct line-length judgments prior to the critical trial and excluded data from participants who got fewer than 2 of the 3 judgments correct for this analysis. For the sustained inattentional blindness task, we computed the absolute percentage deviation from the correct count on the last pre-critical trial and excluded data from participants who were more than 20% off in their count. For the search inattentional blindness task, we computed the percentage of pre-critical trials for which participants correctly identified whether the search target was odd or even and excluded data from participants who were less than 80% accurate. In principle, this analysis includes only those participants who we know to have performed adequately on the primary task. We did not apply this exclusion criterion in our primary analysis for two reasons. First, performance on the primary task might be a source of variance tied to the other cognitive and personality measures in the study. Second, even if people perform poorly on the primary task, they might still be adequately engaged in trying to do the task, meaning that the exclusion criterion might remove data from participants that should be included when measuring individual differences.

[**NOTE: We will report any meaningful discrepancies between our primary measure and the robustness checks in the analyses, but we have not flagged every possible place where we might do that. Assume that if we observe a difference in pattern other than an overall shift in average percentage noticing (which would be expected and not interesting) we will add a mention of it in text. The supplement provides the full analyses using each of these robustness checks.**]

Absence of full attention trial

I'm curious as to why the authors chose not to include a full attention trial in either IB task? These are not always routinely used, but I do believe they are far more common than not, and there is evidence that its use (for excluding subjects) can influence results (see Hutchinson et al. 2022). To me, there is also sound principled reason for its use as a manipulation check to ensure subjects are able to perceive the unexpected object under normal viewing conditions. It is of course unnecessary if noticing rates end up being at ceiling in the divided attention trial, but since we do not know this in advance, it is a rather essential (and easy to implement) check.

Studies in the literature vary in whether or not they include a full-attention trial. The original studies using the transient IB task (e.g., Mack & Rock, 1998) did so in order to prove that the unexpected object was visible with a 200ms presentation (even though there really wasn't much doubt given that 200ms is a long presentation for detectability/perceptibility). For the sustained IB task, the need for a full-attention trial is diminished because the critical object is

present for 5+ seconds. The lack of a need for a visibility check in our tasks can be verified by viewing the demos—the critical object is obvious when not performing the primary task. In all of our work since about 2010, we have not included full-attention trials, and many in the literature do not as well. The perceptibility question is not really in doubt any more.

It also is problematic to use the full-attention trial as an inclusion criterion (e.g., see White et al., 2018). First, if noticing on the full attention is related to primary task performance, then using it for inclusion/exclusion can undermine random assignment for any tests of between-group condition effects on the primary outcome measure (it would introduce "endogeneity"). In essence, it would be conditioning on an outcome measure. Second, participants might fail to report the object on the full attention trial not because it was imperceptible but because they momentarily mind-wandered or didn't pay attention to the display on *that* trial. In fact, that is often how failures on the full-attention trial are interpreted (because perceptibility of the unexpected object is not in doubt). Given that individual differences in the propensity to mind wander might be associated with the other cognitive and personality individual difference measures, we feel it is suboptimal to use noticing on a full-attention trial as a manipulation check or inclusion criterion.

Characterization of load theory

The authors border on mischaracterizing load theory at points. The reason for this is that the load manipulation in this study is clearly a cognitive load one—in the low load group, subjects will be required to retain one tally and in the high load, two tallies. The authors later themselves state it is a cognitive load manipulation.

We had not intended for load theory to be a core component addressed in this study. We had used "load" as a shorthand in our tables and figures but hadn't intended to make claims about how primary task difficulty maps onto the construct of load or ties into load theory. Rather than expanding our review and discussion of load theory, we instead chose to remove the word "load" from the manuscript and instead describe the manipulation more atheoretically as a task-difficulty manipulation (easy vs. difficult counting task). We think that description better captures our intent and avoids the sorts of concerns raised by the reviewer. Readers then can interpret our task-difficulty findings in their own preferred theoretical framework.

I have a couple key problems:

1) They seem to selectively cite and/or miss key references at points. For example, page 21, line 15-17, which omits citing other research with different findings (De Fockert & Bremner, 2011), as well as two important meta- analyses conducted on load in IB (Hutchinson et al. 2022; Matias et al. 2022), that are clearly relevant here.

Because we changed our terminology to avoid using the term "load," we have not added these citations. We had already discussed research on primary task difficulty effects.

Aside: Part of the issue with the way load has been addressed in the inattentional blindness literature is that what counts as "load" varies widely. For example, the De Fockert & Bremner (2011) study was a dual-task design, focusing on the effects of performing a secondary digit memory task while also performing the cross-judgment task. That is different from varying the difficulty of the primary task itself, and it might explain the seemingly contradictory results: People consistently notice less when the primary task is more difficult, but adding a secondary task likely diverts attention from the primary task which could lead to increased noticing (although see Fougnie & Marois (2007) for a different result). In any case, the effects of primary task difficulty on noticing have been consistent across a range of different tasks.

2) The authors state that cognitive demands can lead to a decline in noticing rates, but cognitive load may be better characterized as an inverted U, where maximal IB should occur when cognitive demands are optimally challenging. So under conditions with some standard degree of cognitive demand, an increase should lead to a reduction in IB—the increased cognitive demands lead to resources inadvertently spilling over to process task irrelevant information. Indeed this is also what leads to its distinctiveness from perceptual load. See De Fockert and Bremner 2011 for a characteristic example of this.

As noted, the De Focker & Bremner (2011) study varied the demands of a secondary, unrelated memory task rather than the cognitive demands of the primary task (our manipulation). We think the best explanation for their result is that attention to the primary task was impaired by the secondary task (also, their effects are based on small samples, with 9/16 noticing under high load and 3/16 noticing under low load). In our case, the load manipulation is part of the primary task, and we don't see any reason to expect a U shaped function given the levels of task difficulty we're using. To our knowledge, no prior study that manipulated primary task difficulty has observed a U-shaped pattern for noticing, although we can readily imagine a primary task that is so hard that participants would give up on performing it and consequently notice more.

Hutchinson et al. 2022 conclude that cognitive load is all over the place, so I appreciate that this issue may not be resolved, but this should not give the authors a free go ahead to simply ignore the issue and paint it however they choose, particularly since they state it is a theoretically relevant manipulation (page 16, line 22-24) and indeed are in a unique position where their RR may add important evidence toward this issue (even if it is not their priority).

We're not sure how we had mischaracterized load, but it's true that the idea of cognitive load in the inattentional blindness literature is "all over the place," conflating difficulty of the primary task with the effects of secondary working memory tasks (among other variants). We have sidestepped this issue by describing our manipulation as focusing on varying primary task difficulty (what it actually does) rather than in terms of the theoretical construct of load (we no longer use the word "load" in the paper). Our results might be informative for those looking to interpret task difficulty manipulations as cognitive load without a need for us to delve into that literature or to make claims about it (which is not our focus or interest).

Characterization of Inattentional Blindness

In my view, claiming that IB's "primary measure" is something that falls outside subjects intentions/attention, is theoretically presumptive and in some respects wrong. I refer specifically to page 6, lines 19-21: "the primary measure in an IB task is noticing of objects that explicitly fall outside the participants' intentions and attention". The reason is because multiple mechanisms underly IB. One of those mechanisms is well characterized by so-called "attention set", since IB is less likely for unexpected objects that sync up with the subjects top-down attentional strategy. In these cases, noticing is boosted precisely because the unexpected objects falls within the subject's 'intentions and attention', and so here this characterization is awkward at best.

My recommendation is to use terminology that is less presumptive and is more in line with descriptors or features of IB, for example task relevance and expectations—the unexpected object is task irrelevant in that there is no explicit response required of it to perform the task, and it is unexpected in that the subject has no prior knowledge it will be presented. IB can be spoken of using these terms without making claims to what it is a primary measure of (which may be more prone to bias based upon an authors preferred theory, see Yaron et al. 2022).

We don't believe our definition is presumptive, but we agree our wording could have been clearer. We see a distinction between effects of similarity to the attended/ignored items on noticing (attention set effects) and an observer's deliberate intention to devote attention to the detection of the critical object that they do not know will appear. The fact that the critical object is unexpected means that participants are not deliberately searching for *it*. That is, they are not devoting attention in advance to trying to find it. The attention-set and similarity effects (something our lab has studied for decades) test whether variations in the nature of the unexpected object or the task affect whether or not people report it. When it falls within the attention set that participants have for the primary task it is more likely to be noticed even if people were not deliberately *searching for it* as part of their strategy. That is, we draw a distinction between the intention to devote attention to something in advance (what we meant by "explicitly fall outside the participants' intentions and attention")—if it is unexpected, they aren't devoting attention to detecting it as part of the task—and the likelihood that it will

draw attention to itself by virtue of similarity. To make this distinction clear, we added the following right after the sentence flagged by the reviewer:

> That is, in all of these cases, participants know that the additional object will appear and either deliberately try to ignore it or try to minimize its influence on their primary task performance. In an inattentional blindness task, though, participants do not know that an additional object might appear. As long as the additional object is entirely unexpected, they have no reason to intentionally devote attention to it in advance (Mack & Rock, 1998).

Stimulus size

The authors propose to undertake their work online which is a great opportunity to allow for a large sample size, something which is especially useful for IB. But I was surprised that they do not appear to propose to use any control for visual angle. This may not have been possible previously in online behavioural experiments, but is now routinely possible. One potential option that exists is to add in a virtual chinrest (Li et al. 2020). I am guessing the authors are apt with javascript since this is how their IB tasks are built. There may still be some noise in the measurements but it is much improved compared with nothing. I recommend the authors add in a virtual chinrest or justify why one (or something similar) has not been used here.

> Viewing distance/angle might be a small, overall source of noise (maybe) for noticing rates, but it is not likely to contribute to the pattern of noticing rates across conditions in a meaningful way. Inattentional blindness studies yield consistent, reliable results without such controls, and the pattern of results hold up in spaces ranging from laptops to lecture halls. It also is not something typically controlled in in-person, laboratory studies of inattentional blindness either (except when using eye tracking, and those studies produce similar patterns to ones that didn't control distance). Although virtual chinrests are a way to handle psychophysics online, in a study like this, there is no way (short of invasive video surveillance) to ensure that participants actually stick to their measured viewing distance. And even if they do, display sizes vary. We could ask for their display size and viewing distance, but in our experience, people also don't report those measures accurately (and they don't seem to matter anyway). Given that most Prolific workers likely are using laptops, the viewing distance and display size are likely fairly constrained in any case. In short, the lack of a control for viewing distance for studies that are not manipulating the "distance" of the unexpected object from fixation likely would add only a small amount of noise to the estimates (if any).

Minor

page 4, lines 24-27: the meta-analysis by Hutchinson et al. (2022) looks at lots of these 'systematic' factors, so would be worth mentioning

We added a citation to Hutchinson et al.'s (2022) review at the end of the first paragraph.

page 8, line 9-13: I am not entirely sure if this is the most common; perhaps consider revising wording to "one of the most common"

We changed it to say "includes a common manipulation..."

page 13, line 24: "There actually was an extra object". I checked both tasks and this is not in the instruction on the transient task. The sustained task also seems to cease after about 1s for me. It could be an issue with my internet/web browser, or the demo version, but I tried on several computers and each time it occurred. It could be worth rechecking.

Thank you for catching the wording inconsistency in the demo and we're sorry for the playback issue. We have updated the demos to make sure the text matches what's now in the paper, verified that the sustained inattentional blindness demo runs for its full duration, and added a new demo for the added search inattentional blindness task.

page 16, line 19-21: and whether the UO went from left-right or right-left? and the UO itself? Probably worth including all randomization factors here.

We made this list complete.

page 19, line 37-38: It is minor, but I recommend including in the main manuscript

We would prefer not to include this level of granularity in the main text. It would require a fairly massive figure/table with 4 objects per task x 4 tasks x 3 criteria (liberal, conservative, liberal with accuracy requirement). We already specify that we will report any discrepancies in the pattern of results, and the pattern for individual objects is unlikely to be of interest to a general reader. All of the information will be fully available in the supplement for anyone who is interested, and all of the data will be available for anyone interested in digging deeper.

page 20, line 8: There is a subheading "sustained IB" and then immediately following: "as for the transient IB task". I get what this is saying, but it is confusing. I suggest simple rephrase.

Thanks for catching this confusing start. That awkward transition was eliminated when we rewrote that section.

page 22, line 15-19: one potential problem is that, unlike many previous studies, only two pre-critical trials are proposed to be used. This might be insufficient for estimating accuracy (in

such a way as to make it comparable to other studies, which seems to be the goal). Other studies using more pre-critical trials will mean greater precision for estimating subjects accuracy.

> Some studies do use more pre-critical trials, but many do not. We're hoping to keep the tasks as short as possible, and primary task accuracy is a secondary measure. We prefer not to add more trials just to increase precision of the accuracy estimates.

page 23, line 15: (xx); is this a missing citation?

> Good catch. The section was overhauled but now includes the appropriate citations in the new spot.

page 34, line 22: bit unclear on why this is described as an exploratory analysis in the RR stage 1. My understanding is that this terminology should be reserved for stage 2 analyses/results that are not pre-registered.

> By "exploratory" we mean that we don't have any focused hypotheses but we know that we and others might be interested in the descriptive results. We don't see an issue with the idea of describing planned exploratory analysis. As some of us have argued elsewhere (Lindsay et al, 2016), there is nothing wrong with preregistering analyses that do not involve specified hypothesis tests; it's a way of specifying what measures we will report. At stage 2, any new analysis inspired by the results we observed in our planned analyses will be explicitly flagged as motivated by our observed results and not preregistered.

Another question - was the unexpected object the same across trials? (so if it is an L for critical trial, it is an L for divided attention) It is a bit unclear at present.

> Thank you for catching this ambiguity. Yes, each person saw the same unexpected object on the critical trial and divided attention trial of an inattentional blindness task (but possibly different ones across tasks). We have made that clearer in the text by noting that the divided attention trial of each task used the same unexpected object as on the critical trial. For example: "Following the critical trial, participants completed a divided attention trial with the same additional object moving in the same direction, followed by the same three questions about the additional object."

References

Cartwright-Finch, U., & Lavie, N. (2007). The role of perceptual load in inattentional blindness. Cognition, 102(3), 321-340.

de Fockert, J. W., & Bremner, A. J. (2011). Release of inattentional blindness by high working memory load: Elucidating the relationship between working memory and selective attention. Cognition, 121(3), 400–408. https://doi.org/10.1016/j.cognition.2011.08.016

Fougnie, D., & Marois, R. (2007). Executive working memory load induces inattentional blindness. Psychonomic bulletin & review, 14(1), 142-147.

Hutchinson, B. T., Pammer, K., Bandara, K., & Jack, B. N. (2022). A tale of two theories: A meta-analysis of the attention set and load theories of inattentional blindness. Psychological bulletin, 148(5-6), 370-396. https://doi.org/10.1037/bul0000371

Koivisto M, Grassini S, Salminen-Vaparanta N, & Revonsuo A. (2017). Different electrophysiological correlates of visual awareness for detection and identification. Journal of Cognitive Neuroscience, 29(9):1621–1631. doi: 10.1162/jocn_a_01149

Li, Q., Joo, S. J., Yeatman, J. D., & Reinecke, K. (2020). Controlling for Participants' Viewing Distance in Large-Scale, Psychophysical Online Experiments Using a Virtual Chinrest. Scientific Reports, 10(1), 1-11. doi: 10.1038/s41598-019-57204-1

Lindsay, D. S., Simons, D. J., & Lilienfeld, S. O. (2016). Research preregistration 101. APS observer, 29.

Matias, J., Belletier, C., Izaute, M., Lutz, M., & Silvert, L. (2022). The role of perceptual and cognitive load on inattentional blindness: A systematic review and three meta-analyses. Quarterly journal of experimental psychology, 75(10), 1844– 1875. https://doi.org/10.1177/17470218211064903

Nartker, M., Firestone, C., Egeth, H., & Phillips, I. (2024). Sensitivity to visual features in inattentional blindness. bioRxiv Preprint. https://doi.org/10.1101/2024.05.18.593967

Persuh, M. (2018). Measuring Perceptual Consciousness. Frontiers in psychology, 8, 2320. https://doi.org/10.3389/fpsyg.2017.02320

Simons, D. J., Chabris, C. F., Schnur, T., & Levin, D. T. (2002). Evidence for preserved representations in change blindness. Consciousness and Cognition, 11(1), 78-97. doi:10.1006/ccog.2001.0533

White, R. C., Davies, M., & Davies, A. M. A. (2018). Inattentional blindness on the full-attention trial: Are we throwing out the baby with the bathwater? Consciousness and cognition, 59, 64-77.

Yaron, I., Melloni, L., Pitts, M., & Mudrik, L. (2022). The ConTraSt database for analysing and comparing empirical studies of consciousness theories. Nature Human Behaviour, 6(4), 593–604. https://doi.org/10.1038/s41562-021-01284-5

Yaron, I., Zeevi, Y., Korisky, U., Marshall, W., & Mudrik, L. (2024). Progressing, not regressing: A possible solution to the problem of regression to the mean in unconscious processing studies. Psychonomic bulletin & review, 31(1), 49–64. https://doi.org/10.3758/s13423-023-02326-x