Response to reviewers for "Improving the reliability of cognitive task measures: A narrative review"

Samuel Zorowitz[1], Yael Niv[1,2]

[1]Princeton Neuroscience Institute, Princeton University, USA
[2]Department of Psychology, Princeton University, USA

Formatting note

In the document below, reviewers' comments are indicated in bold typeface. Quoted text from the revised manuscript is presented in quotation blocks; within these excerpts, text that is new is indicated in italicized typeface. All updates to text in the revised manuscript are detailed below.

Reviewer #1

This review discusses strategies for improving the reliability of commonly used cognitive tasks, with examples from the literature. Overall I thought this was a nicely written review that makes a constructive contribution to the literature. My comments are minor suggestions where the authors could expand on the critical discussion of some points.

1. It may be worth considering Rouder et al.'s (2019) preprint [1], which discusses some of these same issues (e.g. how far hierarchical models can get us, whether the required trial numbers for some tasks are prohibitively large). It is perhaps also noting (e.g.) cases where hierarchical models did not improve reliability (Whitehead et al., 2020; [2]).

We thank the reviewer for bringing both papers to our attention. We have cited Rouder et al. at multiple points throughout the revised manuscript, including in the section on calculating reliability to note that hierarchical modeling is not a universal panacea. Specifically, the text now reads (beginning on page 5, line 42):

> Instead, it may be preferable to use trial-level hierarchical models in which observations are organized hierarchically (e.g., individuals within a group, trials within an individual) with variability modeled at both levels. Hierarchical models exert a pooling or regularization effect on person-level variables, in effect correcting for measurement error and improving estimates of reliability [13–15]. The benefits of hierarchical models for estimating reliability has been multiply demonstrated [27, 32–34], *though see [Rouder et al.] for discussion of when these benefits may be limited.*

With respect to Whitehead et al., we would interpret their study as providing further evidence in favor of hierarchical models. Though the reliability of the Simon effect did not differ substantially between hierarchical modeling vs. traditional sum score approaches ($\rho = 0.65$ vs. $\rho = 0.61$), the reliability of both the Stroop effect ($\rho = 0.65$ vs. $\rho = 0.50$) and Flanker effect ($\rho = 0.57$ vs. $\rho = 0.31$) were notably improved.

2. Pg. 5 notes that the type of ICC to use depends on the experimenter's goals - it may be worth expanding this point to encompass discussion of other methods. For example, hierarchical models may lead to higher reliability estimates, but this may not be the most informative estimate if subsequent analysis/applications don't have the same statistical properties (e.g. if the measures are used in traditional analyses, or if it is intended to be used as an individual-level diagnostic).

We thank the reviewer for this point. However, we are unsure of the sort of situation that the reviewer has in mind. For example, if a researcher is using a task measure as an individual-level diagnostic (e.g., for treatment selection for a particular patient), group-level priors from an already-fit hierarchical model could be used to regularize a model fit to a particular patient's data. Because of this, and due to word limit constraints, we have elected to forego adding this point to the main text.

3. Pg. 6: "Practice effects are not inherently an issue for reliability — especially if an experimenter is only interested in the consistency, but not the absolute agreement, of participants' performance over time...". This appears to assume that practice effects would be constant across participants, i.e. that every participant would learn the task at the same rate, therefore not changing the rank ordering. This may require evidence/consideration. With both practice and fatigue effects, there is also a possibility that task length interacts with participants.

We thank the reviewer for this important point. We have now clarified the text to acknowledge that practice effects may impact reliability insofar that they are differentially expressed by participants. The text now reads (page 7, lines 12-13):

> Practice effects are not inherently an issue for reliability — especially if an experimenter is only interested in the consistency, but not the absolute agreement, of

participants' performance over time — but they can become a pernicious issue if *they are exhibited differentially across participants* or if they are severe enough to induce ceiling effects.

Reviewer #2

Summary: In the current study, the authors review the literature on the psychometrics of cognitive task measures with a focus on the construct of measurement reliability. They overview how poor reliability has impacted individual difference research using cognitive task measures and then discuss various different perspectives on how to resolve issues associated with poor reliability.

Strengths: I think this is a very well written piece! The authors cover quite a broad range of perspectives in a small amount of space, and they provide good directions for future researcher on reliability and cognitive tasks.

Major Suggestions: I only have a few "major" suggestions that I believe help correct, clarify, or strengthen the authors' arguments:

1. On page 2, the authors note that "Reliability is therefore a prerequisite for validity: an unreliable task measure reflects measurement error and not the construct of interest." I understand the reliability is often discussed as such a prerequisite, but I think it is a bit of an oversimplification. This often-repeated statement assumes that there is some reliability threshold that, once reached, makes it possible for a measure to be valid. In other words, that one cannot make valid inference when reliability is low. However, I don't believe this is the case. e.g., using equation 1 in the main text, if measures X and Y have a true correlation of yet poor reliabilities of $r_T = 0.8$ yet poor reliabilities of $\rho_X = 0.3$ and $\rho_Y = 0.3$, I can expect to observe a correlation of $r_O = r_T\sqrt{(\rho_X \cdot \rho_Y} = 0.24$. If I assume this observed correlation is the true correlation, then this is an invalid inference. However, if we account for unreliability and disattenuate the observed correlation, we get a valid inference. This is a simple example, but the field of decision theory is full of other counter examples—so long as we account for uncertainty, we can obtain valid inferences (albeit our inferences will be very uncertain if reliability is very low). It seems that the phrase I quoted above takes a black-and-white approach to reliability and validity, assuming a measure is either reliable or not (and subsequently allowed to be valid or not). I think the above points are important because the "black-and-white view" of reliability is apparently widely held—I think it is largely responsible for why researchers have been sounding alarm bells re: individual difference research using cognitive tasks being untrustworthy. I wonder if the authors could make it more clear that we can indeed make valid inference/decisions when reliability is

low, we just need to account for uncertainty/low reliability when drawing inference (which may in some cases stop us from making a decision dependent on our risk preferences, context, etc.).

We thank the reviewer for this thoughtful point. We believe, however, the reviewer is conflating valid *inference* and valid *measurement*. We do not dispute that researchers can make valid inferences using invalid measures. To continue the example above: if we disattenuate a correlation between two measures with poor reliability, as Rouder et al. [1] show, we can expect large uncertainty around our estimate. Insofar that we clearly reflect that uncertainty (e.g., "the true correlation may be as small as -0.6 or as large as 0.6"), we have made a valid inference. That says little, however, about the validity of our measurement. If the reliability of our measure is so poor that our measure predominantly reflects something other than the construct we intend to measure (e.g. noise), then it is invalid.

With that said, the reviewer touches upon an important point: the false dichotomy between "acceptable" and "unacceptable" reliability based on conventions from elsewhere in psychology. We believe the uncritical adoption of standards for reliability (e.g., acceptable reliability ≥ 0.8) is to the detriment of biological psychiatry. (Moreover, in our view, this is at least part of the reason why some researchers view task-based individual-difference research as untrustworthy.) As such, we have addressed this issue in a new paragraph in the discussion (page 14, lines 1-14):

> *We conclude with two important points. First, although we have discussed the importance of task reliability, we have largely avoided the question of when a task measure is "reliable enough". Though it is tempting to fall back on conventional cutoffs (e.g., $\rho \geq 0.7$), what constitutes sufficient reliability in actuality will depend on the goal(s) of the researcher. If the goal is to detect a significant individual-differences correlation, such as between a task measure and self-reported symptom measure, then a task measure with "unacceptable" reliability by conventional standards may suffice (e.g., if a researcher has the resources to collect a sample large enough to be adequately powered to detect a correlation at the attenuated magnitude). On the other hand, if a researcher intends to estimate an individual-differences correlation with high precision, or use a task measure in a high stakes setting (e.g., treatment selection for an individual patient), then high reliability may be required. We cannot overstate the value of simulation studies (e.g., [24]) for researchers trying to determine what level of reliability is required to meet their goals and risk preferences.*

We hope this new addition addresses, at least in part, the reviewer's concerns.

2. On page 5, the authors note that: "... Cronbach's α, requires certain assumptions that are unrealistic for many task data... As such, internal consistency for task measures is instead usually calculated via split-K reliability... Split-half reliability (K = 2) is most common." This claim assumes that split-half reliability

metrics make more realistic assumptions re: task data than Cronbach's α. However, Cronbach's α is analytically equivalent to the average of all possible split-half reliabilities. The way that the paragraph is organized almost implies to me that split-half reliability is qualitatively different from α. I think this is partly true dependent on how data are split, but it would be useful for authors to explicitly explain the relation between α and split half reliability. Then, they can segue into their later discussion on how different splitting choices can be made to relax the restrictive assumptions of α.

We thank the reviewer for this important point. We agree that the framing of the text suggests that Cronbach's α and split-half reliability are qualitatively distinct measures, when in actuality they belong to the same family of measures [3]. Given word limit constraints and the fact that this topic has been discussed at length elsewhere [3], we have chosen to not go into detail here, but rather clarify and qualify the text while pointing readers to more detailed treatments on the topic. The next now reads (beginning on page 5, line 9):

> Calculating the internal consistency of a task measure is more complicated. *The most common measure of internal consistency is Cronbach's α, which is a function of the average correlation across all unique pairs of trials. However, Cronbach's α is an accurate measure of reliability only under assumptions that are unrealistic for many tasks (e.g., equivalence of trials, uncorrelated measurement error; [6, 16]). As such, internal consistency for task measures is instead usually calculated via split-half reliability, where reliability is estimated after trial data have been divided into two halves.* A critical challenge in calculating split-half reliability is in deciding how to partition the data, as estimates of reliability may be also biased if the data partitions violate either of the two above assumptions (for detailed discussion, see [16, 30]). For example, first-second splitting (i.e., partitioning the data into the first and second halves of an experiment) may underestimate reliability due to nonequivalence of the two partitions resulting from practice, fatigue, or other linear time effects. In contrast, odd-even splitting (i.e., partitioning the data into odd and even trials) may cause bias when behavior across trials is non-independent (i.e., measurement error is correlated across trials), artificially inflating the similarity of data across partitions and thereby decreasing estimates of measurement noise and overestimating reliability. Therefore, where possible, a permutation-based approach to calculating split-half reliability is recommended [6, 30]. Here, reliability is averaged across many thousands of random partitions of the data into halves. *(Insofar that Cronbach's α is analytically equivalent to the average of all possible split-half reliability estimates [31], permutation-based split-half reliability provides an approximation to Cronbach's α while avoiding its problematic assumptions.)*

3. On page 5, the authors note that: "traditional sum or mean estimates... may substantially underestimate task reliability... because such summary scores are

contaminated by trial-level noise that... increases the measurement error... By directly modeling trial-by-trial variability, estimates of participants' performance are effectively de-noised. It is also possible to use hierarchical models that partially pool data across participants". I think that there are two lines of reasoning here that are conflated, making the individual points incorrect as stated. Specifically, it is not the modeling of trial-level variability that "denoises" the estimates—it is either: (1) the use of a model of person-level behavior that better characterizes variability, or (2) the combination of trial-level modeling with the hierarchical model. In other words, if I were to only fit a normal distribution using MLE to each person's RT distribution across timepoints and correlate the resulting μ parameters (as an estimate of reliability), I would not expect that the μ parameters would have higher reliability than the sample means (despite the model technically estimating trial-level noise). The hierarchical model induces pooling, which results in a higher estimate for reliability. That said, one of the arguments we make in Haines et al. (2020; ref 13) is that more sophisticated models may lead to better estimates of reliability. e.g., use of a shifted lognormal over a normal model may produce estimates with better reliability because they better characterize the full distribution of behavior (even absent the hierarchical model). I think perhaps the different points about a choice of person-level model (e.g., normal, lognormal, etc. on RT distributions) and choice of group-level model (e.g., normal population distribution across parameters of person-level models) are being (unintentionally) conflated in the quoted statement above, making it appear as if the authors are claiming that estimating variability alone will increase reliability.

We thank the reviewer for this important feedback. We have revised this section of the manuscript to clarify our description of the benefits of hierarchical and generative models (beginning on page 5, line 40):

> As a final point, traditional sum or mean score estimates of performance (e.g., proportion correct responses, mean response time) may substantially underestimate task reliability [13–15]. This is because such summary scores are contaminated by trial-level noise that, in the absence of a sufficiently (possibly prohibitively) large number of trials, increases measurement error (and thus diminishes reliability). *Instead, it may be preferable to use trial-level hierarchical models in which observations are organized hierarchically (e.g., individuals within a group, trials within an individual) with variability 5 modeled at both levels. Hierarchical models exert a pooling or regularization effect on person-level variables, in effect correcting for measurement error and improving estimates of reliability [13–15]. The benefits of hierarchical models for estimating reliability has been multiply demonstrated [27, 32–34], though see [24] for discussion of when these benefits may be limited. Using statistical models that more accurately characterize the latent data-generating process (e.g., using the shifted log-normal distribution to model response times) may*

also improve reliability estimates [14, 35]. For a detailed discussion of hierarchical and generative models in the context of task reliability, see Haines and colleagues (this issue; [14]).

Minor Suggestions:

1. On page 7 the authors note that: "Amplifying the magnitude of an experimental effect… typically increases the range of participants' responses to it." I think it would be worth noting that this runs counter to the "reliability paradox" hypothesis—i.e. that making an effect strong for experimental purposes results in some sort of reduction in between-person variance. If it is "typical" that amplifying an experimental effect actually increases between-person variability, then how is this reconciled with the reliability paradox?

We thank the reviewer for this thoughtful point. As we understand it, the crux of the "reliability paradox" hypothesis is that the most replicable findings in experimental psychology are not necessarily replicable because their effects are large in absolute terms (e.g., the average cognitive interference effect on response times is only 50 ms [1]), but because they are consistently elicited across participants (i.e., low between-participants variability). Amplifying the magnitude of an experimental effect (e.g., increasing both the mean and variance of a cognitive interference effect) may therefore decrease group-averaged effect sizes while increasing reliability. As such, this is not necessarily counter to the reliability paradox hypothesis.

This having been said, we realize readers may read "effect" as "effect size" in our sentence. We therefore changed the text as below (page 8, lines 3-6):

> A primary strategy for increasing between-participants variability is to enhance the experimental *manipulation*. Amplifying the strength of an experimental *manipulation* (e.g., making a task more challenging, increasing the potency of affect induction) typically increases the range of participants' responses to it.

Reviewer #3

This is an excellent and accessible review/guide covering a variety of approaches for improving the reliability of cognitive tasks. I think this will find a broad audience and could serve as a great resource for both established researchers and students in research methods classes.

My comments are relatively minor:

1. The introduction overstates the current utility of cognitive tasks in relation to understanding/predicting/diagnosing psychiatric symptoms. I realize this is not

the focus of the paper, but I think either the authors should briefly make a better case for cognitive tasks as "invaluable tools" in psychiatry or the authors should present a more balanced view of the current state of cognitive tasks in relation to psychiatry.

We thank the reviewer for this important point and agree that our characterization of the value-add of cognitive tasks to biological psychiatry (so far) was overstated. We have tempered our language in the introduction, which now reads (beginning on page 1, line 20):

> *Cognitive tasks hold great promise for biological psychiatry.* When properly designed, such tasks are capable of isolating and measuring specific cognitive processes. Individual differences in performance on cognitive tasks can therefore provide researchers with crucial insights into the cognitive processes underlying psychiatric phenomena. Elsewhere in psychology, cognitive tasks have been useful in predicting important outcomes such as academic achievement [1] and cognitive decline [2]. *Cognitive tasks, then, have the potential to be invaluable tools for refining our understanding of psychiatric symptoms and syndromes.* For a cognitive task to be useful in this regard, however, it must possess sufficient measurement properties.

2. On page 2, the re-worded definition of reliability as "a task measure is reliable if, assuming participants have not changed, it produces the same score for each participant over time" is a bit misleading and does not encapsulate the important distinction between reliability/consistency and agreement that the authors delve into later in the paper. It might be helpful to soften the definition used here so that readers do not understand reliability to mean literally "same score".

We thank the reviewer for this important comment. We have refined our definition of reliability, so that the text now reads (page 2, lines 10-12):

> Finally, the reliability of a task measure characterizes the degree to which it consistently measures some feature of participants. *That is, a task measure is reliable if, assuming participants have not changed, it produces the same scores, or the same ordering of scores, for participants within a single testing session or across multiple testing sessions.* This review focuses on task-measure reliability.

3. As accessibility seems to be a major goal of this review and since the reliability paradox is a central issue in using cognitive tasks for individual differences research, I think it would be helpful to give a concrete example when discussing the reliability paradox on pages 3 and 4.

We thank the reviewer for this helpful suggestion. We have now included a brief example in the paragraph describing the "reliability paradox" (page 4, line 2-4):

> One possible explanation for this finding is the so-called "reliability paradox" of cognitive tasks [17], which states that the often lackluster reliability of tasks is a result of a mismatch in goals between experimental and individual-differences psychological research. In experimental psychology, the goal is often to demonstrate the existence of a behavioral effect. One means of increasing the power to detect an effect is to minimize between-participants variance. This is the exact opposite of what is desirable for individual differences research, where between-participants variance is essential to achieving reliable task measures. *For example, the Stroop effect is one of the most robust effects in experimental psychology; virtually everyone shows a Stroop effect [23]. However, in part due to this fact, between-participants variance on the Stroop effect is often limited [24].* Thus, the tendency in biological psychiatry to adopt the most prominent tasks in experimental psychology—the ones that most reliably demonstrate a behavioral effect—may actually hamstring efforts to study individual differences.

4. Throughout the manuscript, it would be helpful to highlight the magnitude of the reliability improvement that was achieved by each of the highlighted methods (increasing trials, using highly distinguishable stimuli, reducing ceiling effects, enhancing experimental effects, etc.)? I am left wondering whether these methods are capable of changing task reliability scores from unacceptable to good, from good to excellent? Can we expect cognitive task reliability to perform as well as that of self-report measures?

We thank the reviewer for another helpful suggestion. Where we discuss changes to task designs that resulted in an improvement to reliability, we have now provided descriptions of the magnitude of improvement. For example, when describing McLean and colleagues' efforts to improve the beads task, the text now reads (page 7, line 40):

> This new design was effective in preventing participants from becoming aware of the target sequence, which in turn resulted in more consistent responding, which improved the reliability of participants' information seeking scores *(from* $\rho = 0.62$ *to* 0.84*).*

5. While the paper covers a variety of approaches, there is a more in-depth focus on reviewing methods for altering the design of experimental tasks on the front-end rather than reviewing analytic strategies that can be employed to improve reliability, and I agree that these strategies should get the greatest emphasis. However, I do think it would be helpful to expand a bit regarding analytic strategies, as follows:

a. the sentence "It is possible to (re)design tasks to achieve good reliability, even to the high levels dictated by conventional standards [21-24]" is misleading as most

of the referenced papers here did not alter the task design; rather they used more recent analytic strategies to improve reliability. This should be clarified as it is an important take-away for readers to understand that it may be possible to improve reliability in legacy data from traditional tasks by using newer analytic strategies. Same comment in relation to the statement "We hope that this article can serve as a helpful guide for experimenters who wish to *design a new task, or improve an existing task*, to achieve sufficient reliability for use in individual-differences research". Improved reliability may potentially be achieved without redesigning the task itself.

We thank the reviewer for these important points. We agree that citing Waltmann et al. and Sullivan-Toole et al. in this context is misleading, as improvements in task reliability in those studies stem from the use of improved statistical methods and not alterations to task design. We have removed those two citations from the first sentence in question. We also agree with the second point. The text of the second sentence in question now reads (page 4, lines 19-20):

> We hope that this article can serve as a helpful guide for experimenters designing a new task, improving an existing task, *or refining their scoring methods* to achieve sufficient reliability for use in individual-differences research.

b. Finally, either in the final full paragraph of page 5 or in section 3.2.3 'Reducing parameter estimation noise', I think it is worth briefly highlighting that hierarchical models *can* improve reliability but are not sufficient and that lower-level model parameterization is also critical for reliability improvement, as demonstrated in Brown, et al., 2021 and Sullivan-Toole, et al., 2022. Although, I am biased with regards to this last point, so the authors should use their own judgment.

We thank the reviewer for this suggestion. We have revised to text to better highlight that using more appropriate trial-level models can also improve estimates of reliability. The text now reads (page 6, lines 5-7):

> *Using statistical models that more accurately characterize the latent data-generating process (e.g., using the shifted log-normal distribution to model response times) may also improve reliability estimates [14, 35].* For a detailed discussion of hierarchical and generative models in the context of task reliability, see Haines and colleagues (this issue; [14]).

Reviewer #4

This manuscript reviews empirical findings and recommendations on reliability of cognitive tasks. The topic is timely and the review covers a range of areas. This

manuscript will be a good reference and resource for researchers in this field. I particularly appreciated the attention to special considerations for reliability in cognitive tasks that differ from other applications of reliability (e.g., questionnaires). My comments primarily are about clarifying points to ensure they are useful to readers new to this area as well as suggested additions to ensure the review covers all relevant topics. Given the goal of the paper, some of these comments are more picky than most reviews I do to minimize potential confusion for the reader.

Clarifications:

1. In the intro, page 2, starting line 10, the authors define reliability as consistency in rank-ordering across measurement instances. As they discuss later on, however, reliability may not be a measure of rank order and it can measure within-session as well as between-session consistency.

We thank the reviewer for this crucial point (see also Reviewer 3, point 2). We have refined our definition of reliability, so that the text now reads (page 2, lines 10-12):

> Finally, the reliability of a task measure characterizes the degree to which it consistently measures some feature of participants. *That is, a task measure is reliable if, assuming participants have not changed, it produces the same scores, or the same ordering of scores, for participants within a single testing session or across multiple testing sessions.* This review focuses on task-measure reliability.

2. Page 2, line 5 - discriminating power or discriminatory power?

We thank the reviewer for spotting this typo, which has now been corrected.

3. Page 3, starting on line 54 - as the authors note later, other factors besides low within-person variability also affect reliability. In particular, difference scores are not introduced until the section on improving reliability but could instead be introduced with other factors affecting reliability here.

We thank the reviewer for this helpful suggestion. We now mention difference scores as part of the list of factors affecting reliability. The text now reads (page 2, lines 38-39):

> Indeed, task reliability can vary as a function of experiment parameters (specific stimulus set, number of trials, time limits [8, 9]); sample populations (healthy adults, children, psychiatric patients [9, 10]); testing locations (in clinic, online); response modality (desktop, smartphone, virtual reality [11, 12]); *scoring method (component scores, difference scores)*; and estimation method [13–15].

4. The authors discuss both classical test theory and generalizability theory - it may be helpful to quickly define each or otherwise provide context for these terms and the assumptions they make.

Due to word limit constraints, we have elected not to expand on the history and assumptions of classical test theory. Instead, we now cite Allen & Yen's [4] seminal introduction to classical test theory and measurement theory.

5. The authors discuss generative or statistical models as a way to improve reliability at the end of the section on calculating reliability, but I don't see a discussion of this in the section about improving reliability - this topic may be more useful in that section. In particular, models of underlying processes can be more reliable than relying on difference scores, e.g. PMID: 34252724 [5].

We thank the reviewer for this suggestion. We were deliberate in choosing to place our discussion of hierarchical and generative models in the section discussing calculating reliability. Our motivations for doing so were twofold. First, we did not want to devote much space to hierarchical models in the context of task reliability as this topic has received extensive treatment elsewhere [6–8]. Second, we instead wanted to primarily focus on principles of experiment design that impact task reliability, which has received less cohesive treatment. We have attempted to make our intentions for the manuscript more clear in the introduction. The text now reads (page 4, line 14-16):

> The purpose of the current article is to provide a narrative review of approaches to improve task-measure reliability. The article is divided into two main parts. First, we review methods of calculating reliability and discuss some nuances that are specific to cognitive tasks. *This section is not intended to be exhaustive, but rather to point to other publications where these topics have already been discussed at length.* Then, we introduce a taxonomy of approaches to improve the reliability of cognitive-task measures through experiment design and analysis, using concrete examples from the published literature. We hope that this article can serve as a helpful guide for experimenters designing a new task, improving an existing task, or refining their scoring methods to achieve sufficient reliability for use in individual-differences research.

Separately, we thank the reviewer for bringing Weigard et al. to our attention. We now cite their work in the section discussing the use of alternatives to difference scores (page 13, lines 29-31):

> Another approach is to identify alternative measures of task performance. For example, intra-individual response time variability *and cognitive efficiency have been identified as correlates of executive control that can be measured reliably [80, 81] and are altered in psychopathology [82, 83].*

Suggested additions:

1. Page 5, starting line 9 - the overview of pros and cons of different types of internal consistency measurements is helpful. For recommendations, why not randomly assign trials to halves? This does not necessarily require many permutations. Using blocks also has drawbacks: it assumes the task can be split up into several independent blocks of trials, and with few blocks or influential early blocks, has many of the same drawbacks as splitting first vs. second half of trials.

We thank the reviewer for this comment. We do not want to recommend a single random assignment of trials, as this runs the risk of, by chance, underestimating reliability (if trials in each split are not approximately tau equivalent) or overestimating reliability (if trials in each split exhibit correlated errors) [9, 10]. Instead, we echo others' recommendations of a permutation-based approach that uses many random partitions. Indeed, while we agree that using blocks has drawbacks, they are largely unavoidable for certain types of cognitive tasks (e.g. learning tasks with stationary reward distributions), hence we left these in as well.

2. Page 6, starting line 43 - consider adding PMID: 34688897 as citation.

We thank the reviewer for bringing this article to our attention. We now cite Chen et al. (2021) [8] at multiple points throughout the revised manuscript.

3. Page 6, starting line 5 - consider adding PMID: 28864865 as citation.

We thank the reviewer for bringing this article [11] to our attention. However, we are uncertain about the reasons why the reviewer suggested we cite it. The study in question concerns neither individual differences nor task reliability. Though the study does find evidence of practice effects, they are not so severe as to cause range restriction.

4. Section 3.1.4 - the authors suggest that online samples produce more variable data, but empirical results generally show that in-person and online studies do not meaningfully differ (e.g., https://doi.org/10.1525/collabra.17213; PMID: 34267650; PMID: 34267650; PMID: 32883156). [12–14]

We thank the reviewer for this thoughtful point, but we do not agree that the published literature generally shows that in-person and online studies do not meaningfully differ. Gillan & Rutledge [15] recently noted that they required a 30% increase in sample size (from N=461 to N=670) to reliably detect the same association between compulsivity and model-based planning in online vs in-person participants. Moreover, Nussenbaum and colleages [12] actually reported a similar finding; they found that they required a 40% increase in sample size (from N=45 to N=63) to reliably detect the same association between age and model-based planning in online vs. in-person participants. Although Ito and colleagues [13] do not find substantial differences in behavior on average for in-person vs. online settings, they do not compare variability in performance; moreover, only 23 out of their 83 participants completed the study in-person,

such that their study is adequately powered to detect only the largest effects. Finally, Chaytor and colleagues [14] compare two different neuropsychological test batteries between in-person (WAIS) vs. online (TestMyBrain) settings, confounding task and testing context. In short, we believe that the current empirical evidence, if anything, better supports the hypothesis that online data are more variable. As such, we believe we are justified to state, "experimenters should take special care to ensure that an increase in between-participants variance is not offset by a concomitant increase in measurement noise."

5. An additional consideration of online studies, particularly when using common tasks, is that participants may not be naïve: *Stewart, N., Ungemach, C., Harris, A. J. L., Bartels, D. M., Newell, B. R., Paolacci, G., & Chandler, J. (2015). The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. Judgment and Decision Making, 10(5), 479-491* [16].

We thank the reviewer for this important point. We now address participant nonnaïvete in the discussion of recruiting participants online. The text now reads (beginning on page 8, line 39):

> *Separately, online participants may be more familiar with particular experimental paradigms due to previous exposure [49], which may attenuate between-participants performance variability for the reasons previously mentioned (e.g., practice effects). Thus, researchers running experiments online may want to alter task paradigms so that they appear less similar to preexisting 8 versions and/or limit the recruitment of highly-experienced participants [50].*

6. The discussion is too brief and needs a summary of findings and recommendations - what can people do to improve reliability when designing a new task, modifying an existing task, or using a task that is already developed? What are remaining issues in reliability of cognitive tasks and future directions?

We thank the reviewer for this suggestion. Unfortunately, due to word limit constraints, we are unable to provide a substantive discussion section (though we have added an additional paragraph discussing when a task measure is "reliable enough"; see Reviewer 2, point 1).

Reviewer #5

Zorowitz and Niv thoroughly reviewed the definitions of reliability and ways to increase task reliability. Authors suggested two major ways for improving task reliability, which are increasing between-participant variance and decreasing measurement noise. Authors provided theoretical and mathematical formulations to explain why each approach would be beneficial for increasing task reliability. I

think this is an excellent comprehensive guideline for interpreting and improving the reliability of task reliability.

Major suggestions:

1. In the first paragraph in section 3.1.3. (p.7), the authors mention that amplifying the magnitude of an experimental effect could improve reliability. Kucina et al. achieved this by increasing the task demand. I felt that this approach might cause the floor effect discussed in this paper, which would ironically decrease the range of responses. This might be related to the reliability paradox, as an attempt to increase the magnitude of an experimental effect decreases the between-participant variability. Are there studies that address this issue? A careful calibration of task difficulty as discussed in the following paragraph would help resolve this problem.

We thank the reviewer for this point and wish to clarify a misunderstanding. Kucina et al. measure conflict effects via response time, not response accuracy. As such, the issue of floor effects due to increased task demands is obviated. To prevent further reader confusion, we have made clear that those authors measured response times. The text now reads:

> For example, Kucina and colleagues [26] investigated the reliability of cognitive conflict effects *(as measured by response time)* in new versions of several standard cognitive-control tasks (e.g., Stroop, Flanker, Simon) that amplified cognitive interference via two task design features.

2. In the first paragraph in section 3.2.2 (p.9), the authors suggest we use the most discriminating stimuli. It would be worth mentioning the methods for finding such stimuli. Optimal experimental design methods often aim to select stimuli that are most likely to discriminate between participants (or between models). For example, as briefly discussed by authors, adaptive design optimization selects the design that reduces uncertainty about the parameters in a target computational model to the greatest extent. If the model parameters properly reflect a cognitive process of interest, the stimuli that improves the precision of the parameter estimates are likely to be the stimuli that are most discriminating.

We thank the reviewer for this suggestion. In Section 3.2.2, we now cite Embretson & Reise [17] who provide the gold-standard introduction to item response theory (IRT) for psychologists. IRT methods are ideally suited for identifying the most discriminating stimuli. It is worth noting that adaptive design optimization (ADO) methods do not necessarily identify the most discriminating stimuli (defined here, in line with item response theory, as stimuli for which participants with high ability consistently respond in one way [e.g., make the correct response] while participants with low ability consistently respond in another way [e.g., make an incorrect response]). If an ADO method selects a stimulus based only on the expected response, but not

the variability around that expectation, then an ADO method may not necessarily select the most discriminating stimulus.

References

1. Rouder, J., Kumar, A. & Haaf, J. M. Why most studies of individual differences with inhibition tasks are bound to fail (2019).
2. Whitehead, P. S., Brewer, G. A. & Blais, C. Reliability and convergence of conflict effects: An examination of evidence for domain-general attentional control. *Experimental Psychology* **67,** 303 (2020).
3. Green, S. B. *et al.* Use of internal consistency coefficients for estimating reliability of experimental task scores. en. *Psychon. Bull. Rev.* **23,** 750–763 (2016).
4. Allen, M. J. & Yen, W. M. *Introduction to measurement theory* (Waveland Press, 2001).
5. Weigard, A., Clark, D. A. & Sripada, C. Cognitive efficiency beats top-down control as a reliable individual difference dimension relevant to self-control. *Cognition* **215,** 104818 (2021).
6. Haines, N. *et al.* Learning from the reliability paradox: How theoretically informed generative models can advance the social, behavioral, and brain sciences. *PsyArXiv* (2020).
7. Rouder, J. N. & Haaf, J. M. A psychometrics of individual differences in experimental tasks. en. *Psychon. Bull. Rev.* **26,** 452–467 (2019).
8. Chen, G. *et al.* Trial and error: A hierarchical modeling approach to test-retest reliability. *NeuroImage* **245,** 118647 (2021).
9. Parsons, S., Kruijt, A.-W. & Fox, E. Psychological Science Needs a Standard Practice of Reporting the Reliability of Cognitive-Behavioral Measurements. *Advances in Methods and Practices in Psychological Science* **2,** 378–395 (2019).
10. Pronk, T., Molenaar, D., Wiers, R. W. & Murre, J. Methods to split cognitive task data for estimating split-half reliability: A comprehensive review and systematic assessment. en. *Psychon. Bull. Rev.* **29,** 44–54 (Feb. 2022).
11. Howlett, J. R., Huang, H., Hysek, C. M. & Paulus, M. P. The effect of single-dose methylphenidate on the rate of error-driven learning in healthy males: a randomized controlled trial. *Psychopharmacology* **234,** 3353–3360 (2017).
12. Nussenbaum, K., Scheuplein, M., Phaneuf, C. V., Evans, M. D. & Hartley, C. A. Moving developmental research online: comparing in-lab and web-based studies of model-based reinforcement learning. *Collabra: Psychology* **6** (2020).
13. Ito, K. L. *et al.* Validating Habitual and Goal-Directed Decision-Making Performance Online in Healthy Older Adults. *Frontiers in aging neuroscience* **13** (2021).
14. Chaytor, N. S. *et al.* Construct validity, ecological validity and acceptance of self-administered online neuropsychological assessment in adults. *The Clinical Neuropsychologist* **35,** 148–164 (2021).
15. Gillan, C. M. & Rutledge, R. B. Smartphones and the neuroscience of mental health. *Annual Review of Neuroscience* **44,** 129 (2021).

16. Stewart, N. *et al.* The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. *Judgment and Decision making* **10,** 479–491 (2015).

17. Embretson, S. E. & Reise, S. P. *Item response theory* (Psychology Press, 2013).