Task reliability considerations in computational psychiatry

Craig Hedge, Aline Bompas and Petroc Sumner

Cardiff University

Word count – 1458

Author Note

Craig Hedge, CUBRIC - School of Psychology, Cardiff University; Aline Bompas, CUBRIC - School of Psychology, Cardiff University; Petroc Sumner, School of Psychology, Cardiff University.

This work was supported by the Wellcome Trust (104943/Z/14/Z).

Correspondence concerning this article should be addressed to Craig Hedge[1] and Petroc Sumner[2], School of Psychology, Cardiff University, Tower building, Park Place, Cardiff, CF10 3AT, UK.

Emails: hedgec@cardiff.ac.uk[1], sumnerp@cardiff.ac.uk[2]

Telephone: +44 (0)29 2251 0276[1], +44(0)29 2087 0091[2]

Measuring psychological abilities or traits is trickier than it seems from the published literatures (1). We try to study abstract psychological constructs like ‘inhibition’ or ‘impulsivity’, but we can only measure these indirectly through behaviours such as reaction times or self-report ratings. When we apply these measures in clinical and individual differences research, our goal is typically to understand why a patient group appear to be (e.g.) more ‘impulsive’ than healthy controls, or to use measures of ‘inhibition’ to predict a clinical outcome. There are many potential pitfalls and wrong turns in the path towards achieving this goal. In our recent work, we have shown that our intuitions about what makes a good cognitive task can sometimes lead us astray (2,3). Here, we discuss how these issues intersect with the goals of computational psychiatry.

We recently described what we referred to as the ‘reliability paradox’, where tasks that produce robust and replicable effects can simultaneously have poor reliability for individual differences (2). Take the widely used Stroop task (4) as an example. Participants are typically instructed to quickly and accurately identify the font colour of a written colour word and ignore the word meaning. We then subtract reaction times or errors in congruent trials (e.g. ‘red’ in red font) from incongruent trials (‘blue’ in red font) to create a ‘Stroop cost’, which is assumed to reflect the individual’s ability to overcome their automatic tendency to read the word. The Stroop effect is easily reproducible in a variety of samples (5), and as such we might consider it to be a ‘reliable’ task.

However, when we correlated individuals’ Stroop costs in reaction times between sessions performed three weeks apart, we observed a test retest reliability of ICC=0.64 across two studies. This is much lower than we’d hope for clinical research (6), and other widely used and replicable experimental effects had even poorer retest correlations. Figure 1 illustrates the consequences of this. If we simulate performance on a test assuming a reliability of r=0.64, 28% of individuals who place in the top half of performers on the first occasion will place in the bottom half at time two (and vice-versa; Figure 1A). Further, this noise means that we need sample sizes 2.5 times larger than we would expect if we had not considered it (Figure 1B).

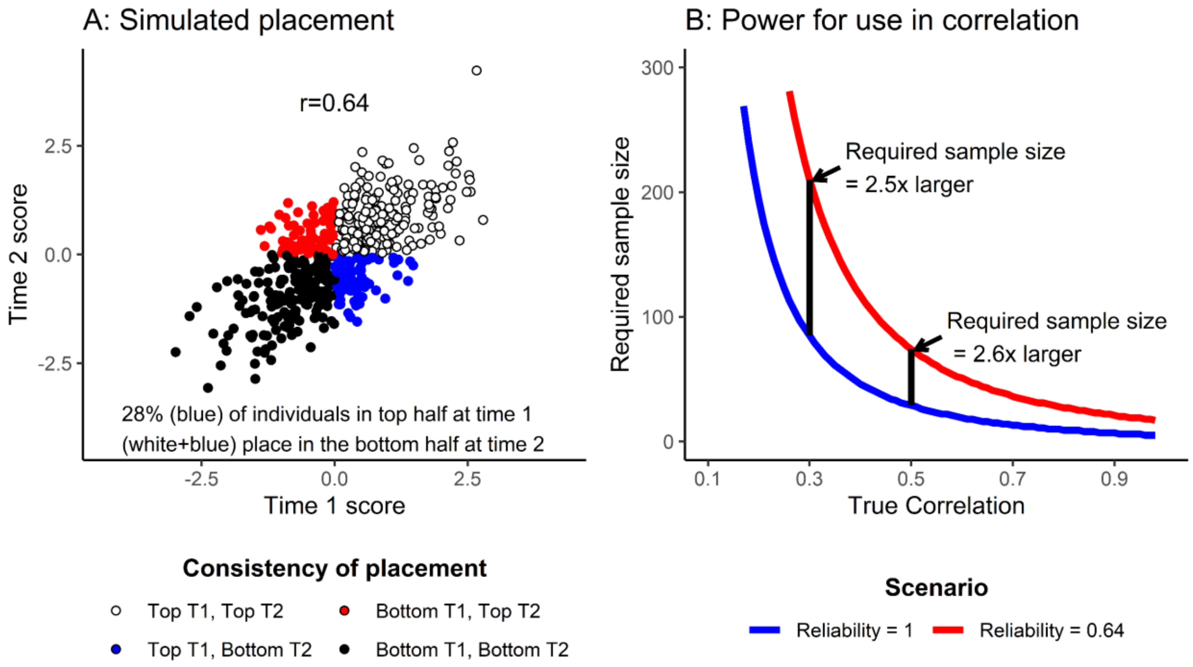

Figure 1. *Panel A shows simulated scores on a cognitive test performed on two occasions assuming a test-retest reliability of r=0.64. Only a subset of the total simulated population (N=1,000,000) are plotted. Even a measure that is considered to have 'good' reliability can frequently lead to misclassifications at an individual level, which is important for clinical applications. Panel B shows the results of a power analysis where we assume that we are correlating our test scores with an external variable (e.g. psychiatric symptoms). The blue line shows the required sample sizes (y-axis) for different underlying correlations (x-axis) both our cognitive test and our measure of psychiatric symptoms could be measured without noise (i.e. if the reliability were r=1). The red line shows the required sample sizes if we assume the same underlying correlation, but now both variables have a reliability of r=0.64. For example, if we assume a medium underlying correlation (r=0.3) between test performance and symptoms, our required sample size increases from N=85 to N=210 in the*

presence of this level of noise. For more details, see (2). *Our power analyses use a two-tailed significance test with a significance threshold (α) of 0.05 and a power (1-β) of 0.8. The code to reproduce these figures is available on the Open Science Framework (https://osf.io/wnhjm).*

This reliability paradox occurs because opposing characteristics make a task successful in experimental and individual differences research. The success of the Stroop task in experimental contexts reflects its ability to nearly always produce a significant average effect. This works because the effect is present and of similar size in most people. In contrast, individual differences research requires variability in the measure in order to reliably identify high and low performers. So the very consistency that makes a task popular in experiments counts against us for individual differences.

A second problem is that alternative ways of measuring the same psychological construct don't always correlate with each other (even after taking reliability issues into account). This is most stark when alternative ways of calculating a measure from the *same task* don't correlate; for example the Stroop effect is sometimes measured using reaction times, and sometimes by using errors. But these alternative Stroop effects measured in the same participants often do not correlate well with each other (3). The same is true for many tasks where one can choose to use reaction times or errors (3,7). The choice of which to use is rarely explicitly justified in papers.

So how do we go about reliably measuring the psychological abilities and traits that we are interested in, and link them to psychiatric or neuropsychological symptoms? Can computational models help us? In the first instance, models help us to be explicit about what we are trying to measure (8). Constructing a cognitive task is like constructing a scene in which to view the cognitive mechanisms of interest. In doing so we implicitly and explicitly make choices about what other functions are also involved. Take the Stroop task again as an example. It is used as a window into processes of cognitive control or inhibition. The task inevitably also draws upon visual processes and word reading. If you are a slow reader, you will have a smaller Stroop effect (try it as you learn a second language; the Stroop effect can measure how automatic you are becoming). An appropriate model provides a way to disentangle these processes.

Models can also help us understand problems and how to overcome them. For example, most task performance is influenced by whether individuals favour accuracy or speed. It is often assumed that subtracting conditions (e.g. incongruent Stroop trials minus congruent Stroop trials) controls for this, but we have shown using model simulations that it does not (3). This actually explains why Stroop effects measured with speed or accuracy do not always correlate. This understanding, in turn, points us to ways to lessen this problem, by, for example, changing the task set up, instructions or constraints on participants to reduce the variability in speed-accuracy behaviour (intermixing trials within blocks, emphasising speed) (3).

Using a model does not inherently overcome the reliability paradox. For a model-based analysis to uncover reliable individual differences in a cognitive ability, we still rely on a task that can capture different levels of that ability. A solution to this is to assess reliability early in task development, and in multiple populations. For example, healthy participants may all perform at ceiling in tasks that are intended as screening tools for dementia. The correlational reliability in a healthy sample would therefore be low, but the task could still be sensitive to differences between healthy controls and individuals with dementia. In that case, correlational reliability may be less important than diagnostic sensitivity. Recently, Paulus et al. (9) proposed a roadmap for computational psychiatry which includes evaluating properties such as reliability and sensitivity to the target population at early stages. There are also several potential indirect benefits in their proposals, including optimising factors such as trial numbers. Increasing the number of trials can give us more precise measures, which improves reliability by decreasing measurement noise. But task duration is important in clinical research, so it is useful to evaluate the balance of reliability and duration early on in a project. This can be done with simulation and subsampling (see supplementary material D of 2).

Finally, we consider two examples where the application of a model has led to improved reliability relative to traditional measures. As we noted above, we should not expect models to create reliable variance where there is none. However, the impurity of behavioural measures may lead to a reliable cognitive ability being masked by confounding processes. A recent example of this can be seen in the application of a choice reaction time model to the dot-probe task (7). The dot-probe task is similar to the Stroop task in its construction. It traditionally measures attentional bias by using the difference in reaction times to a probe preceded by either threat-related stimuli or neutral stimuli. And like with the Stroop task, factors such as speed-accuracy trade-offs will also contribute to the size of attentional bias costs (3). Instead, the authors reasoned that the attentional orienting effect should manifest in the portion of the reaction time that precedes the individual making a decision about the probe. When attentional bias was calculated from a model parameter representing this non-decision time, they observed improved reliability compared to the traditional behavioural subtraction.

Another recent example compared the reliability of a range of statistical and modelling approaches in a reinforcement learning task (10). They observed improved reliability when it was calculated within a hierarchical modelling framework, relative to both traditional behaviour and non-hierarchical versions of the same models. Here, the benefits appeared to reflect the statistical properties of the approach rather than the theoretical assumptions of any given model. It is noteworthy that when we apply a modelling approach we are often doing more than simply applying our theoretical knowledge of the underlying mechanisms. We may combine different behavioural measures (e.g. reaction time and accuracy) or utilise other sources of information about the sample being tested or our prior beliefs about the range of underlying parameters. All of this can help emphasise signal over noise in the data.

At first glance, the questions “Am I measuring the cognitive ability I want to measure?” and “How reliable is my measure of cognitive ability?” may seem like one in the same. However, improving our ability to reliably distinguish between individuals or groups is not an automatic result of improving the mapping between our theories and our measures. Both are essential goals for computational psychiatry, and psychology more broadly, and achieving these goals relies on evaluating, reporting, and optimising our measures (1,9).

Disclosures statement

The authors reported no biomedical financial interests or potential conflicts of interest.

References

1. Flake JK, Fried EI (2019): Measurement Schmeasurement: Questionable measurement practices and how to avoid them. *PsyArXiv Prepr*. Retrieved from https://doi.org/10.31234/osf.io/hs7wm

2. Hedge C, Powell G, Sumner P (2018): The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behav Res Methods* 50: 1166–1186.

3. Hedge C, Powell G, Bompas A, Vivian-Griffiths S, Sumner P (2018): Low and variable correlation between reaction time costs and accuracy costs explained by accumulation models: Meta-analysis and simulations. *Psychol Bull* 144: 1200–1227.

4. Stroop JR (1935): Studies of interference in serial verbal reactions. *J Exp Psychol* 18: 643–662.

5. Ebersole CR, Atherton OE, Belanger AL, Skulborstad HM, Allen JM, Banks JB, *et al.* (2016): Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *J Exp Soc Psychol* 67: 68–82.

6. Nunnally JC (1978): *Psychometric Theory*, 2nd ed. New York: McGraw-Hill.

7. Price RB, Brown V, Siegle GJ (2019): Computational Modeling Applied to the Dot-Probe Task Yields Improved Reliability and Mechanistic Insights. *Biol Psychiatry* 85: 606–612.

8. Huys QJM, Maia T V., Frank MJ (2016): Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature Neuroscience*, vol. 19. pp 404–413.

9. Paulus MP, Huys QJM, Maia T V. (2016): A Roadmap for the Development of Applied Computational Psychiatry. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, vol. 1. pp 386–392.

10. Brown VM, Chen J, Gillan CM, Price RB (2020): Improving the Reliability of Computational Analyses: Model-Based Planning and Its Relationship With Compulsivity. *Biol Psychiatry Cogn Neurosci Neuroimaging*. https://doi.org/10.1016/j.bpsc.2019.12.019