Exploring reliability heterogeneity with multiverse analyses: Data processing decisions unpredictably influence measurement reliability

Sam Parsons[1]

[1] University of Oxford

Author Note

Submitted to Meta-Psychology. Click here to follow the fully transparent editorial process of this submission. Participate in open peer review by commenting through hypothes.is directly on this preprint.

This work was supported by an ESRC grant [ES/R004285/1]

I would like to thank Ana Todorovic for her insightful feedback on an earlier version of this manuscript.

Correspondence concerning this article should be addressed to Sam Parsons, Department of Experimental Psychology, University of Oxford, New Radcliffe House, Radcliffe Observatory Quarter, Oxford, OX2 6AE. E-mail: sam.parsons@psy.ox.ac.uk

Abstract

Analytic flexibility is known to influence the results of statistical tests, e.g. effect sizes and *p*-values. Yet, the degree to which flexibility in data-processing decisions influences the reliability of our measures is unknown. In this paper I attempt to address this question using a series of reliability multiverse analyses. The methods section incorporates a brief tutorial for readers interested in implementing multiverse analyses reported in this manuscript; all functions are contained in the R package *splithalf*. I report six multiverse analyses of data-processing specifications, including accuracy and response time cutoffs. I used data from a Stroop task and Flanker task at two time points. This allowed for an internal consistency reliability multiverse at time 1 and 2, and a test-retest reliability multiverse between time 1 and 2. Largely arbitrary decisions in data-processing led to differences between the highest and lowest reliability estimate of at least 0.2. Importantly, there was no consistent pattern in the data-processing specifications that led to greater reliability, across time as well as tasks. Together, data-processing decisions are highly influential, and largely unpredictable, on measure reliability. I discuss actions researchers could take to mitigate some of the influence of reliability heterogeneity, including adopting hierarchical modelling approaches. Yet, there are no approaches that can completely save us from measurement error. Measurement matters and I call on readers to help us move from what could be a measurement crisis towards a measurement revolution.

Keywords: reliability, multiverse, analytic flexibility, data processing

Exploring reliability heterogeneity with multiverse analyses: Data processing decisions unpredictably influence measurement reliability

The vermilisitude of our conclusions rests on the quality, and the strength, of our evidence. Our evidence rests on the bedrock of our measurements. The quality of our measures defines the quality of our results. Without adequate focus on the validity of our measures, how can we be assured that we are capturing the concept or process that we are interested in? Without any attention to the reliability of our measures, how can we be sure that we are capturing a phenomenon with any precision? Psychological science – I concede some areas are better than others - has a guilty habit of neglecting these foundations.

In a recent paper, my colleagues and I argued for a widespread appreciation for the reliability of our cognitive measures (Parsons et al., 2019). Briefly; low reliability places doubt on the veracity of statistical analyses using that measure; measurement reliability restricts the observable range of effect sizes; and failing to correct for measurement error makes comparing effect sizes between, and within, studies difficult. These issues are compounded by the sad observation that the reporting of reliability (and validity) evidence is woefully poor. Scale validity and reliability is not routinely examined, and many scales are adapted on an ad hoc basis with little or no validation (Flake, Pek, & Hehman, 2017). In other cases scales fail to pass deeper psychometric evaluation, including tests of measurement invariance (Hussey & Hughes, 2018). This likely reflects issues with more superficial approaches to establishing validity evidence - i.e. reporting Cronbach's alpha, stating it is adequate, and moving on. I concede that pockets of psychological science take a more enlightened approach. However, I feel it is reasonable to argue that the field at large is doing well in our measurement practices. Most relevant to this paper; it is the exception rather than the norm to evaluate the psychometric properties of cognitive measurements (Gawronski, Deutsch, & Banse, 2011).

An important reminder: estimates of reliability refer to the measurement obtained - in a specific sample and under particular circumstances, including the task parameters. Reliability is therefore not fixed; it may differ between populations, samples, and testing conditions. Variations of a task may lead to the generation of more or less reliable measurements. For example, the stimulus presentation duration will likely influence the cognitive processes involved in completing the task, perhaps leading participants to perform more consistently in one version, relative to another. Reliability is a property of the measurement, not of the task used to obtain it. Strictly speaking, we cannot state that a task is unreliable; although we might observe a consistent pattern of unreliability in measurements obtained that causes us to question further use of the task.

Thankfully, there is a growing awareness that measurement matters (Fried & Flake, 2018). A valuable term, Questionable Measurement Practices (QMPs), was recently added to our vernacular by Flake and Fried (2019). QMPs describe “undisclosed decisions researchers make that leave questions about the measurements in a study unanswered” (page 2). I hope that QMPs and the importance of measurement become as widely discussed as the parallel idiom, “Questionable Research Practices” (QRPs). Most importantly, wider discussion of these practices should make it clear to all researchers that we make many potentially impactful decisions in the design of our measures, our data processing or cleaning, and our data analysis.

In this paper I was concerned with the influence analytic flexibility on measurement reliability, specifically in data processing or data cleaning. I was inspired in part from numerous papers reporting the unsettlingly low reliability of dot-probe attention bias indices (e.g. Jones, Christiansen, & Field, 2018; Schmukle, 2005; Staugaard, 2009). I was also inspired by other work investigating alternative analyses and data processing strategies, with the intention of yielding a more reliable measurement (e.g. Jones et al., 2018; Price et al., 2015). I was interested in visualising the influence of data processing steps on reliability. My rationale was that as all too often the focus is on decisions made in the beginning (task design) or at the end (data analysis) of the research process. I felt the intermediary stage in which the data is processed is often unexplored in relation to QRPs and QMPs. To fully explain my rationale, we first take a walk through the garden of forking paths.

Analytic flexibility and the garden of forking paths

Every result presented in every research article is the culmination of many decisions made by one or more researchers; the sheer number of combinations of valid decisions is likely uncalculatable. The “garden of forking paths” (Gelman & Loken, 2013) is a useful analogy I use throughout this paper to illustrate this. With each decision that must be made, however arbitrary, the researcher comes to a fork in their research path, and selects one. To add a little suspense, there will be many cases when the researcher does not notice a fork in the road. Perhaps the researcher unconsciously makes the same turn as always, their feet working of their own accord. These forks in the path, the decisions researchers make (whether they are aware or not), may be reasonably combined to make a near uncountable number of paths. Each path also leads to a location; some paths end close to one another, and other times the paths diverge wildly. We can think of the end of the path as the statistical result our researcher arrives at.

The researcher has to decide their path, based on the soundest justifications they can make at each fork (e.g. Lakens et al., 2018). Of course, psychological science has become fully aware of the detrimental effects of selecting one’s path retrospectively, based on where the path ends or the results most exciting to the researcher (read as: $p < .05$; e.g. Simmons, Nelson, & Simonsohn, 2011). Extending the metaphor (hopefully not too painfully): Backtracking the route and selecting multiple paths until reaching a preferred location is akin to walking between the well-tended paths and stomping across the flowerbeds, damaging our colourful science garden. Analytic flexibility is not inherently bad. However, we must acknowledge the ramifications. The effects we observe, or do not, are potentially influenced by all of the decisions made to arrive at them. Thus, a range of possible effects may have been observed that are all equally valid based on the analytical decisions made.

In discussions of analytical flexibility, focus is usually given primarily to decisions made during statistical analysis. For example; should I control for age and gender? Do I reason that this is model more appropriate over that one? Or, where should I set my alpha and how should I justify the decision? Discussions of analytical flexibility often concern issues around *p*-hacking and other QRPs (intended or unintended). However, as Leek and Peng (2015) note, *p*-values are the tip of the iceberg; not enough scrutiny is given to the impact of the many steps in the research pipeline that precede inference testing. I agree. In my estimation, flexibility in measurement and data handling do not receive the scrutiny they deserve. If the garden of forking paths concerns analytic flexibility, then measurement flexibility decides which door one enters the garden through in the first place.

Mapping the garden of forking paths with multiverse analyses

A multiverse analysis (Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016) offers us a "GPS in the garden of forking paths" (Quintana & Heathers, 2019). The process is remarkably simple. First, we define a set of reasonable data processing and analysis decisions. Second, we run the entire set of analyses. We can then examine results across the entire range of results. Specification curve analysis (Simonsohn, Simmons, & Nelson, 2015) adds third step allowing for inference tests across the distribution of results generated in the multiverse (for insightful applications of specification curve analyses, see; Orben & Przybylski, 2019; Rohrer, Egloff, & Schmukle, 2017). In this paper I use "specification" to refer to each combination of data processing decisions in the multiverse analysis.

Multiverse analyses enable us to explore how a researcher's – sometimes arbitrary – choices in data processing (e.g. outlier removal) and analysis decisions (e.g. including covariates, splitting samples) influence statistical results, and the conclusions drawn from the analysis. From this we can examine which choices are more or less influential than others, as well as how robust the result is across the full set of specifications.

A reliability multiverse from many data processing decisions

In this paper I report multiverse analyses exploring the influence of data processing specifications on the reliability of a calculated measurement. I used openly accessible Stroop task and Flanker task data generously shared by Hedge and colleagues (Hedge, Powell, & Sumner, 2018). Following previous work in this area (Parsons et al., 2019), I was interested in the stability and range of reliability estimates on cognitive-behavioural measures. Broadly, I was interested in the positive and negative impact of data processing decisions on reliability. It is possible that certain analytic decisions tend to yield higher reliability estimates; it may be that particular combinations of decisions are also better, or worse, than others. Beyond that, I was interested in the range of estimates. A small range would suggest that measure reliability is relatively stable as we make potentially arbitrary data processing decisions while walking the garden of forking paths. A large range suggests hidden measurement reliability heterogeneity. This is potentially an important, and underappreciated, contributor to the replicability crisis (Loken & Gelman, 2017). Alternatively, this could be a herald for a crisis of measurement.

Methods

Data

Data were obtained from the online repository for Hedge, Sumner, and Powell ((2018); https://osf.io/cwzds/). Full details of the data collection, study design, and procedure can be found in Hedge et al. (2018). These data are ideal for our purposes as they a) contain many trials, helping us obtain more precise estimates of reliability, and b) include two assessment time-points approximately 3-4 weeks apart, allowing us to explore both; internal consistency and test-retest reliability. The data were collected from different studies; for simplicity in this paper, the data across studies were pooled (n = 107 before any data processing – note that this may be different from the sample size presented by Hedge et al. due to differences in data processing). I explored data from the Stroop and the Flanker tasks. Interested readers can find the data and code used to perform the multiverse analyses and generate this manuscript in the Open Science Framework repository for this project (https://osf.io/haz6u/).[1]

Stroop task. Participants made keyed responses to the colour of a word presented in the centre of the screen. In congruent conditions the word was the same as the font colour, whereas, in incongruent trials, the word was a different colour from the font colour. In a neutral condition, the word was not a colour word. Participants completed 240 of each trial type. The outcome index we explore here is the RT cost, calculated as the average RT for incongruent trials minus the average RT for congruent trials.

Flanker task. Participants made keyed responses to the direction of an arrow presented in the centre of the screen. The central arrow was flanked by two other symbols. Congruent trials presented flanking arrows in the same direction as the central arrow, whereas incongruent trials presented flanking arrows in the opposite direction to the central arrow. There was also a neutral condition in which flanking symbols were straight lines. Participants completed 240 of each trial type. The outcome index we explore here is the RT cost, calculated as the average RT for incongruent trials minus the average RT for congruent trials.

[1] I used the following R packages for all analyses and figures, and to generate this document: R (Version 3.6.3; R Core Team, 2018) and the R-packages *Cairo* (Version 1.5.12; Urbanek & Horner, 2019), *dplyr* (Version 0.8.5; Wickham et al., 2019), *forcats* (Version 0.5.0; Wickham, 2019a), *ggplot2* (Version 3.3.0; Wickham, 2016), *gridExtra* (Version 2.3; Auguie, 2017), *papaja* (Version 0.1.0.9942; Aust & Barth, 2020), *patchwork* (Version 1.0.0; Pedersen, 2019), *psych* (Version 1.9.12.31; Revelle, 2019), *purrr* (Version 0.3.4; Henry & Wickham, 2019), *readr* (Version 1.3.1; Wickham, Hester, & Francois, 2018), *splithalf* (Version 0.7.1; Parsons, 2020), *stringr* (Version 1.4.0; Wickham, 2019b), *tibble* (Version 3.0.1; Müller & Wickham, 2019), *tidyr* (Version 1.1.0; Wickham & Henry, 2019), and *tidyverse* (Version 1.3.0; Wickham, Averick, et al., 2019)

Multiverse analysis

In a personal effort to make my research reproducible, and also help others perform similar processes I have developed simple functions to perform the multiverse analyses reported in this paper. The key functions are: *splithalf.multiverse*, *testretest.multiverse*, and *multiverse.plot*. These functions appear in the the *splithalf* package (Parsons (2019); starting at version 0.7.1) and are also provided separately in the supplemental materials.

The functions use R packages *splithalf* (Parsons, 2019) and *psych* (Revelle, 2017) to estimate internal consistency and test-retest reliability, respectively. Moreover, this section acts as a brief tutorial with the aim of helping interested readers conduct their own reliability multiverse analysis with different data or sets of specifications (also see full code: https://osf.io/haz6u/).

Step 1. Creating a list of all specifications. No data were removed before the multiverse analysis. To my knowledge, there are no fixed standards in the literature for processing Stroop or Flanker data. I identified six decisions common to processing RT data, though there are many more. For simplicity I stuck to RT difference scores as the outcome measure of interest. However, there are very different analytical techniques that might be applied to RT tasks such as this (for example, multilevel modelling and drift-diffusion modelling approaches). The decisions were as follows:

- *Total accuracy.* Researchers may opt to remove participants with accuracy lower than a pre-specified cut-off; for example 80 of 90 per cent. I used three options; 80%, 90%, and no cut-off.

- *Absolute response time removals.* Researchers will often remove trials faster than a minimum RT threshold and trials that exceed a maximum RT threshold. I use minimum RT cut-offs at 100ms, 200ms, as well as no cut-off. And, I use two maximum RT cutoffs; 3000ms, and 2000ms.

- *Relative RT cut offs.* After absolute RT cut-offs, researchers can decide to remove trials with RTs greater than a number of standard deviations from the mean (sometimes called relative cut-offs or trimmed means). Three SDs from the mean would remove very extreme outliers; two SDs from the mean is common. I have not seen researchers use one SD from the mean as a cut off, as it is likely a too conservative threshold. As I was interested in a wide range of possible specifications, I included one standard deviation. I use no relative cut off, and one, two, and three SDs from the mean cutoffs in the multiverse.

- *Where to apply the relative cutoff.* The decision to remove trials based on a SD cutoff comes with its own decision. Namely, at what granularity? We could remove trials with RTs greater than 2SDs from the participant's average RT, for example. We could also remove trials with RTs greater than 2SDs from the mean RT within each trial type (congruent and incongruent, for example). I included both options; participant level, and trial type level.

- *Averaging.* Most often the mean RT within each trial type is calculated, and may then be analysed directly, or a difference score calculated to analyse. Researchers may opt to use the median RT instead. I included both options.

The number of possible combinations (data processing specifications) quickly increases with every additional option. Here we have 3 × 2 × 3 × 4 × 2 × 2 = 288 possible specifications. We can specify our list of decisions as follows:

```
specifications <- list(
 ACC_cutoff = c(0, 0.8, 0.9),
 RT_min           = c(0, 100, 200),
 RT_max            = c(2000, 3000),
 RT_sd_cutoff      = c(0, 1, 2, 3),
 split_by           = c("subject", "trial"),
 averaging_method  = c("mean", "median")
)
```

Step 2. Run all specifications and extract reliability estimates. From this decision list, we have a complete list of 288 data processing specifications. In the multiverse analysis the data is processed following each specification parameters, before estimating the reliability of the resulting outcome measure. In the following example code we perform a multiverse analysis on the Stroop data from Hedge et al.'s first testing session. First, we run *splithalf* on the full dataset and save the output into a splithalf object.[2] We then pass this object into *splithalf.multiverse*. The only required inputs are the specification list, and our saved splithalf object.

```
splithalf_1 <- splithalf(data = subset(Hedge_raw_Stroop, time == 1),
                         permutations = 500,
                         var.ACC = "Correct")

multiverse_1 <- splithalf.multiverse(input = splithalf_1,
                 specifications = specifications)
```

The output contains useful information, including; the data, the expanded specification list, each processed dataset, and information about the call. Most important is the "estimates" data frame, which contains the reliability estimate for each specification. The "removals" list can be used to inspect the number of participants and trials remaining following data reduction. "CI" can be used to inspect the median and 95% CI of all reliability estimates.

[2] Note that if users want to run a multiverse on multiple task conditions (e.g. happy and sad stimuli, or time 1 and time 2) they must specify separate multiverse analyses

Step 3. Visualising the multiverse. I find that one of the joys of multiverse analyses are the visualisations; because sometimes science is more art than science. The results section is centred on these visualisations. We use the output from *splithalf.multiverse* in the function multiverse.plot to visualise the specification curve of reliability estimates. It can be called with the following;

```
multiverse.plot(multiverse = multiverse_1,
         title = "My first multiverse")
```

I explain the visualisations in the results section. I have also added functionality to visualise multiple multiverses in the same plot. To do so, the user can specify a list of multiverse objects. For example:

```
multiverse.plot(multiverse = list(multiverse_1,
                                 multiverse_2))
```

Inferences from the multiverse. It is not my aim in this paper to make inferences from these reliability multiverse analyses as one would in a specification curve analysis (Simonsohn et al., 2015). One could use this method to perform inference testing against the curve of reliability estimates. However, it is not clear what this would add: testing whether the reliability estimates significantly differ from zero is a low bar for assessing the reliability of a measure.

Descriptively, it will be useful to explore the full range of estimates. The multiverse objects contain a “CI” object to help extract the median and 95% percentile estimates. The user can obtain these by running `internal.1_Stroop$CI`. I have also provided a small function *threshold* to inspect the proportion of reliability estimates above or below a set threshold. Users can specify whether they are interested in the point estimates or the Confidence Intervals. For example, the following will return the proportion of estimates above a 0.7 threshold (often the lowest bar to describe internal consistency estimates as “acceptable”).

```
threshold(multiverse = internal.1_Stroop,
          threshold = 0.7,
          use = "estimate",
          dir = "above")
```

Analysis plan

In total, I performed six multiverse analyses following the steps described above. Separately for each of the Stroop and Flanker task data, I examined internal consistency reliability at time 1 and at time 2, as well as test-retest reliability from time 1 to time 2. Internal consistency was estimated using 500 permutations of the splithalf procedure for each specification (5000 is standard, but 500 was selected to reduce processing time). For each multiverse I report the median estimate and it’s 95% Confidence Interval, the proportion of estimates exceeding 0.7, and the range of estimates in that multiverse. In addition to visualising each multiverse, I also include visualisations overlapping the internal consistency multiverses from time 1 and time 2. These overlapped plots allow us to visually inspect whether the pattern of reliability estimates following the full range of data processing specifications are comparable across each time point.

Results

I include a visualisation for each multiverse analysis. The reliability estimates are presented on the y axis at the top of the figure; each estimate is represented by a black dot and the 95% confidence interval is represented by the shaded band. The x axis indicates each individual multiverse specification of processing decisions (288 total), displayed in the “dashboard” at the bottom of the figure. The vertical dashed line running through the top panel and the bottom dashboard represents the median reliability estimate. This line is extended through the dashboard to demonstrate that the estimate is derived from the unique combination of data processing decisions, including (from top to bottom, in order of processing step); 1) participant removal below total accuracy threshold, 2) maximum RT cutoff, 3) minimum RT cutoff, 4) removal of RTs > this number of SDs from the mean, 5) whether this removal is at the trial or subject level, and 6) use of mean or median to derive averages.

Stroop Time 1: Internal Consistency. The median reliability estimate was 0.77, 95% CI [0.69,0.91]. Estimates ranged from 0.68 to 0.91. 96% of the reliability estimates were > 0.7.

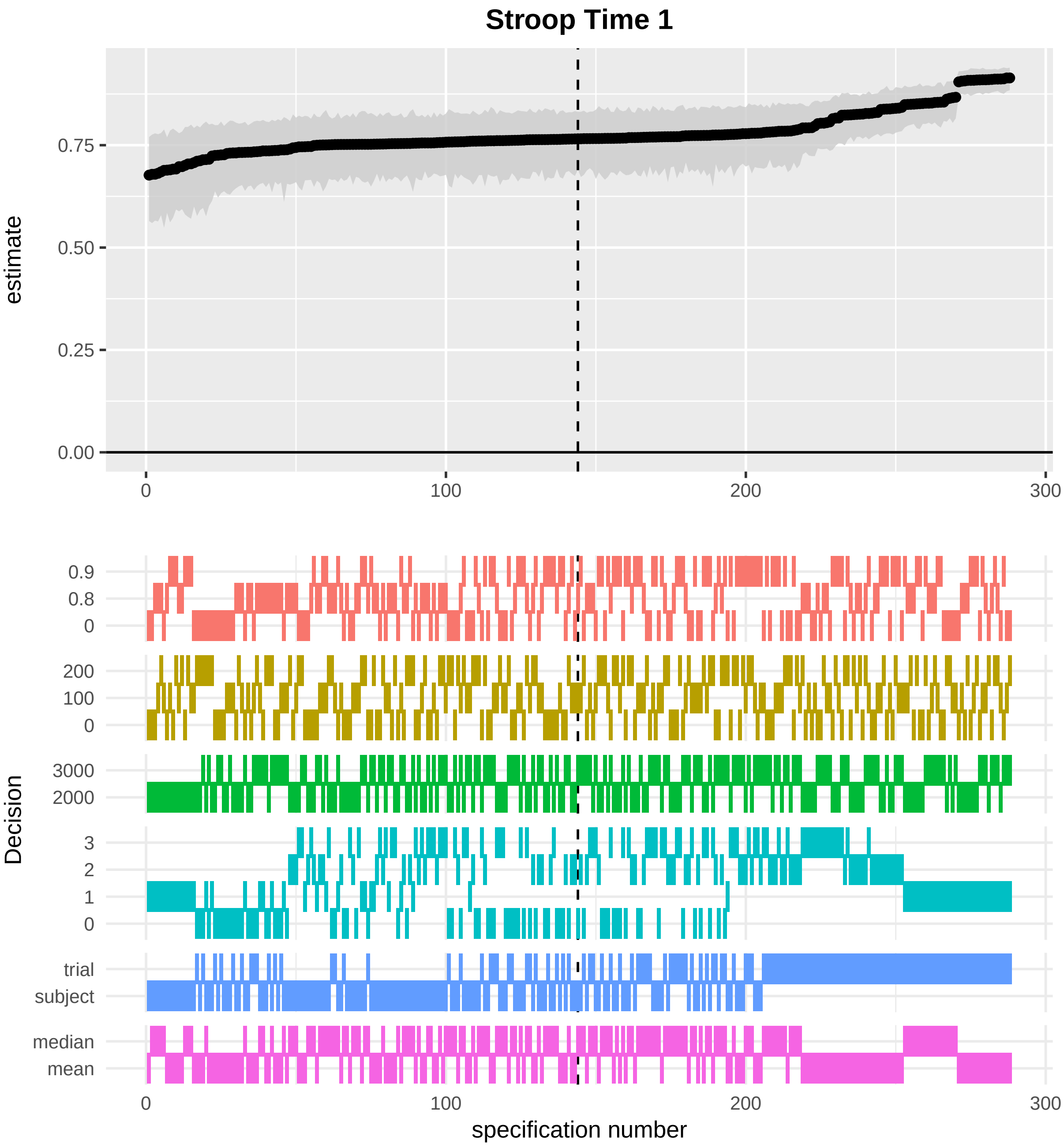

Figure 1. Internal consistency reliability multiverse for Stroop RT cost at time 1

Stroop Time 2: Internal Consistency. The median reliability estimate was 0.67, 95% CI [0.61,0.88]. Estimates ranged from 0.59 to 0.89. 29.00% of the reliability estimates were > 0.7.

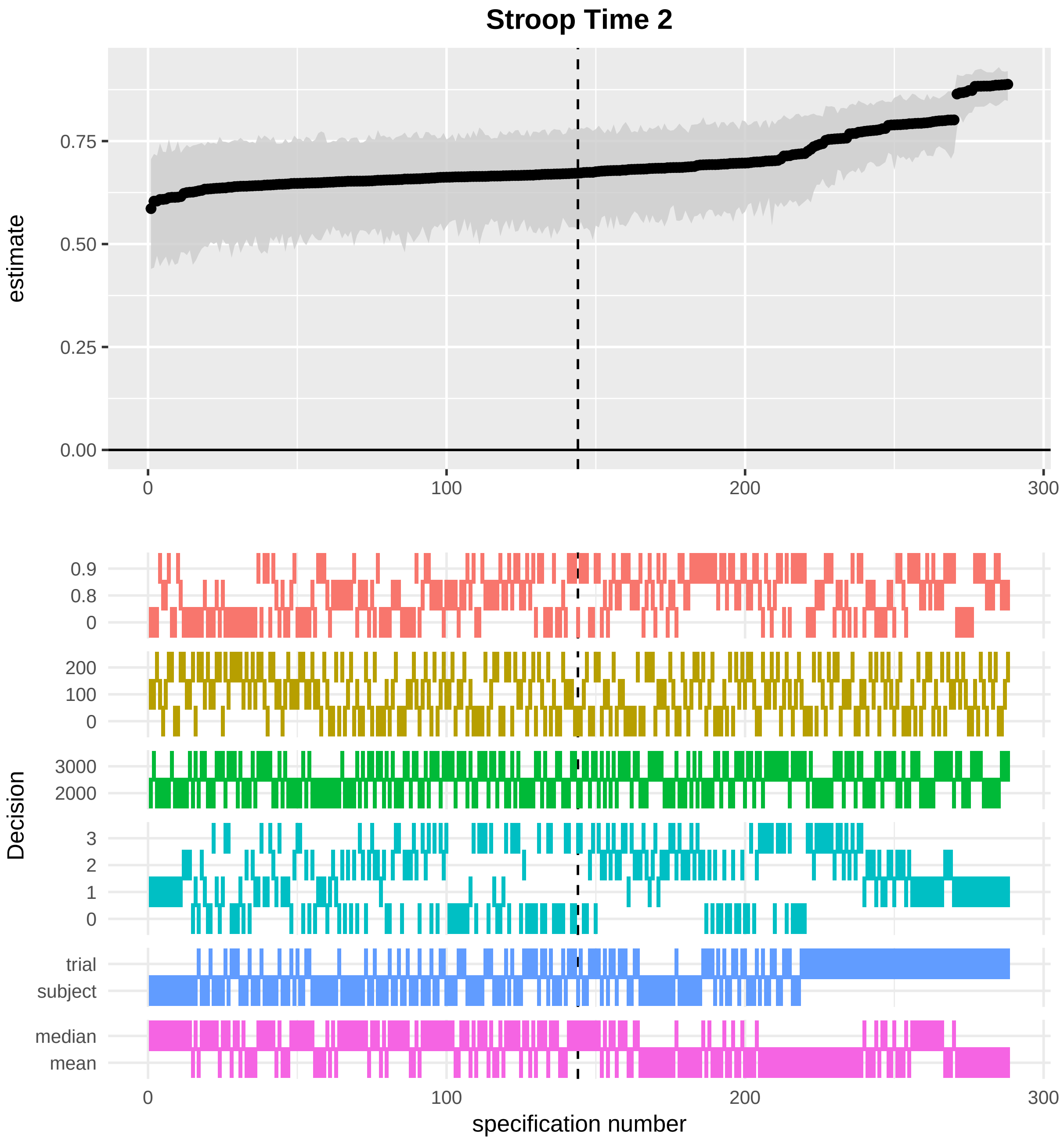

Figure 2. Internal consistency reliability multiverse for Stroop RT cost at time 2

Stroop: test-retest. The median reliability estimate was 0.56, 95% CI [0.50,0.63]. Estimates ranged from 0.47 to 0.63. 0% of the reliability estimates were > 0.7.

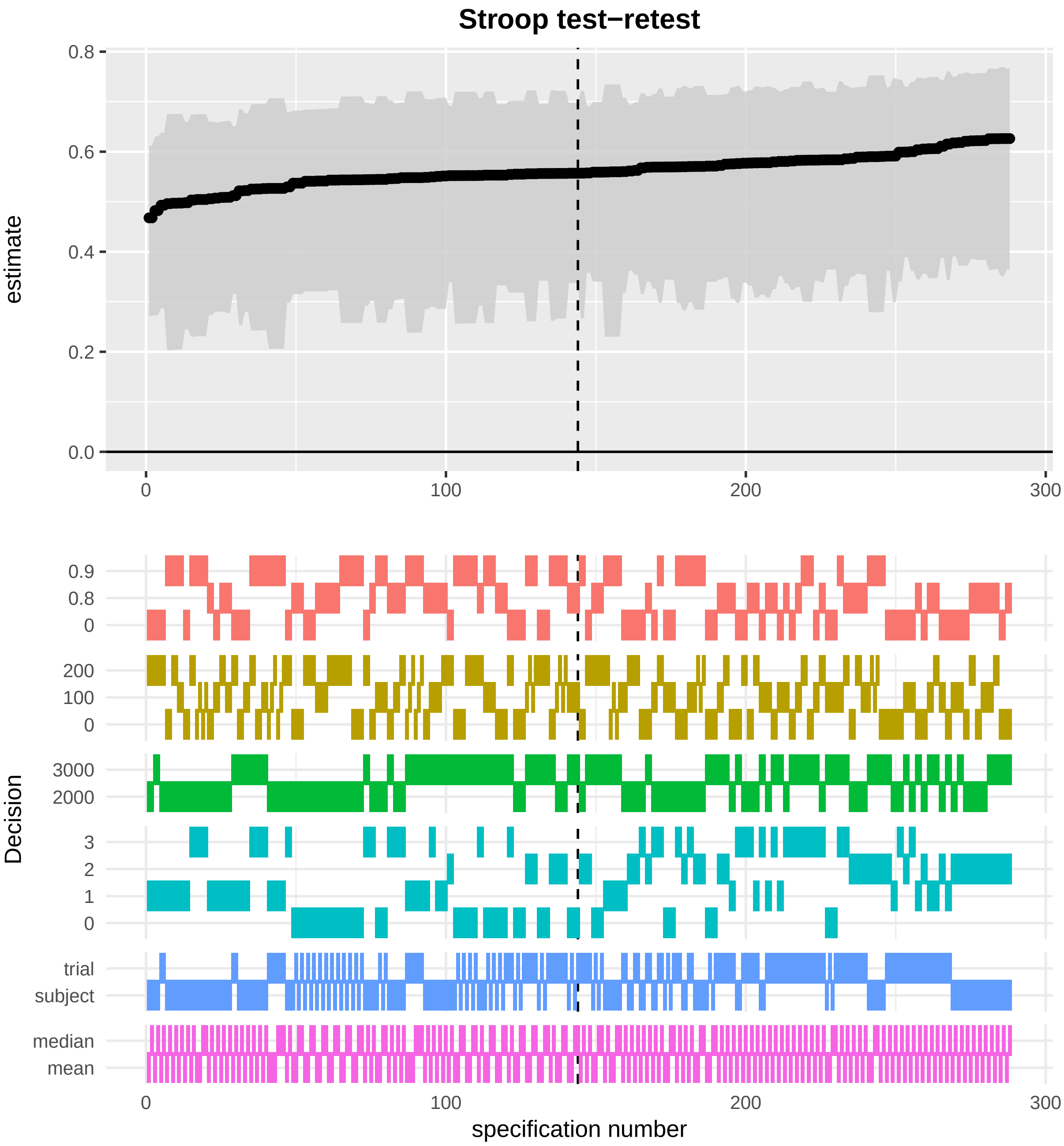

Figure 3. Test-retest reliability multiverse for Stroop RT cost

Flanker Time 1: Internal Consistency. The median reliability estimate was 0.84, 95% CI [0.70,0.92]. Estimates ranged from 0.63 to 0.93. 98% of the reliability estimates were > 0.7.

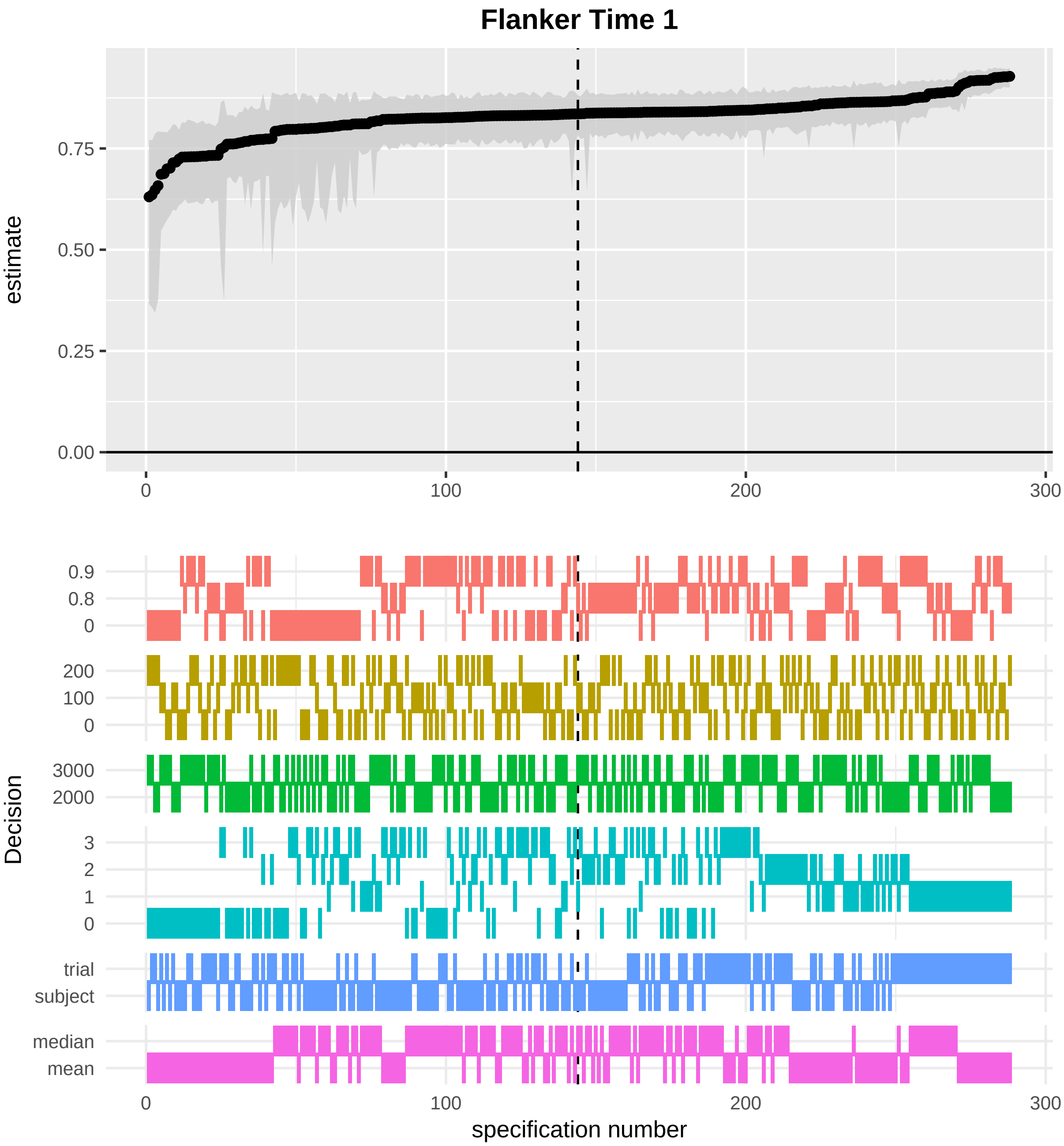

Figure 4. Internal consistency reliability multiverse for Flanker RT cost at time 1

Flanker Time 2: Internal Consistency. The median reliability estimate was 0.74, 95% CI [0.63,0.88]. Estimates ranged from 0.59 to 0.90. 66% of the reliability estimates were > 0.7.

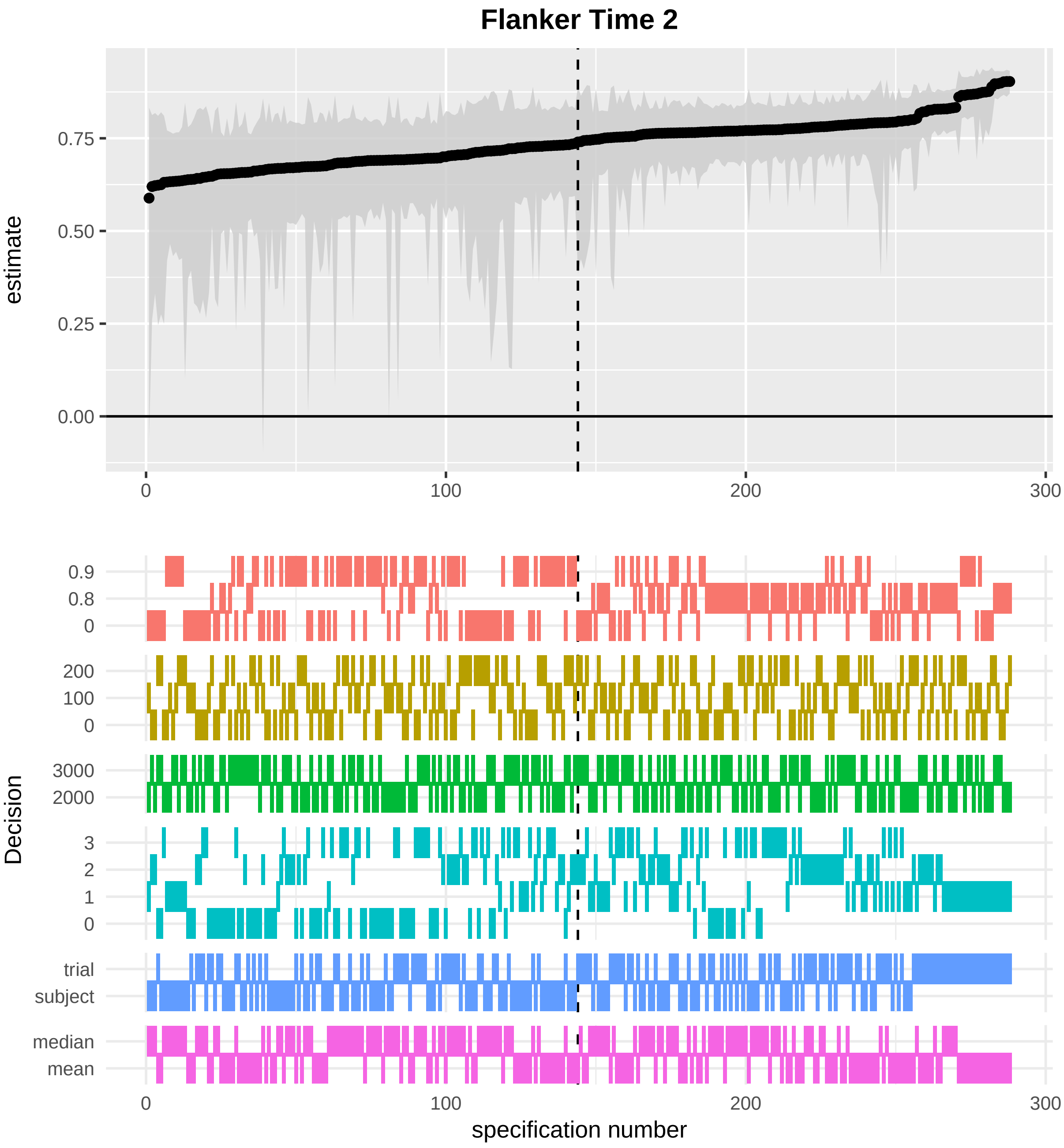

Figure 5. Internal consistency reliability multiverse for Flanker RT cost at time 2

Flanker: test-retest. The median reliability estimate was 0.55, 95% CI [0.30,0.69]. Estimates ranged from 0.28 to 0.72. 2% of the reliability estimates were > 0.7.

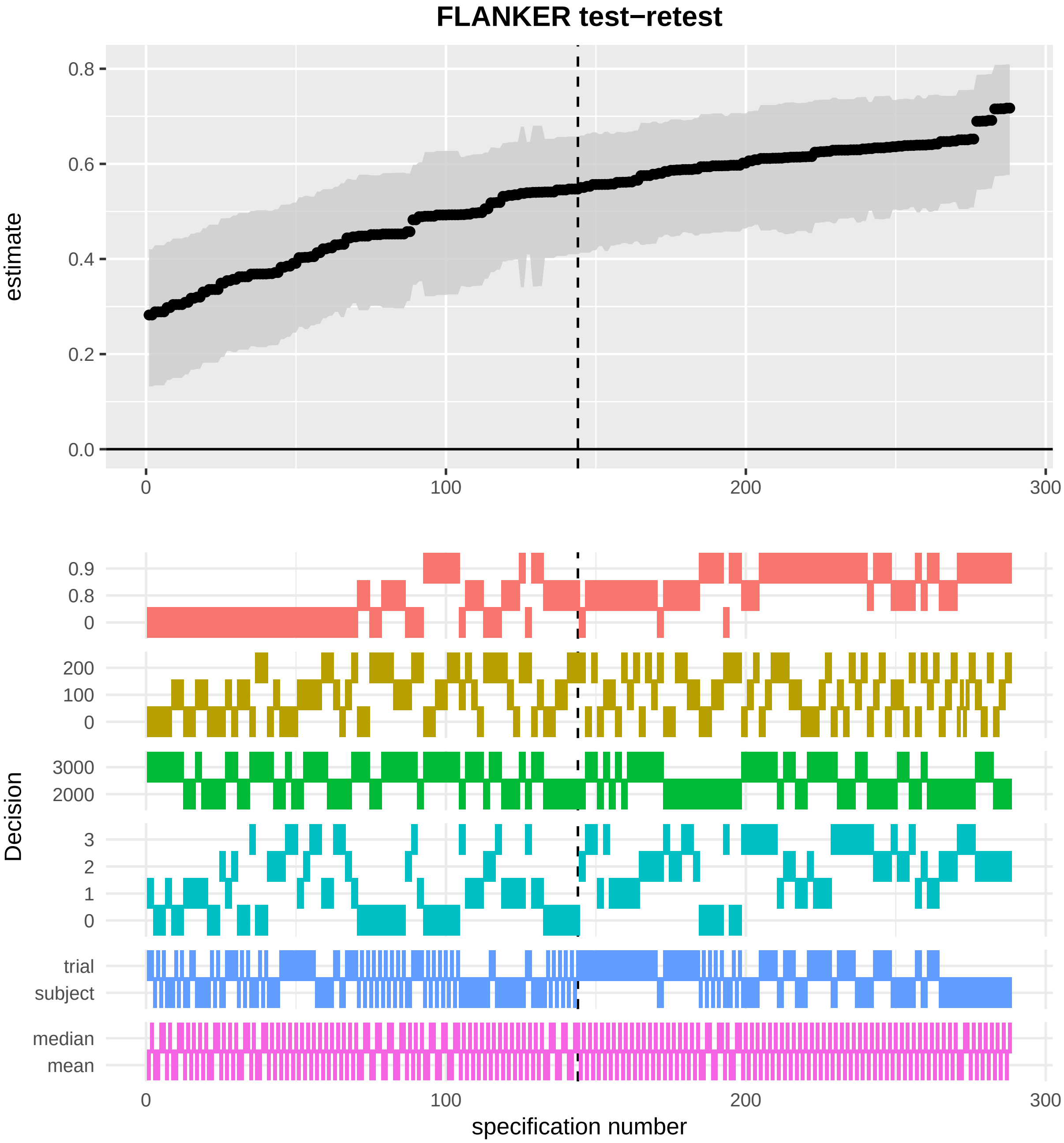

Figure 6. Test-retest reliability multiverse for Flanker RT cost

Overlapping time 1 and time 2 multiverses. In the final two figures I overlap the time 1 and time 2 multiverses, separately for the Stroop and Flanker data. The specifications are ordered by the reliability estimates at time 1 for each measure (Figures 1 and 3). These figures allow us to compare the patterns of reliability estimates following the same data processing decisions.

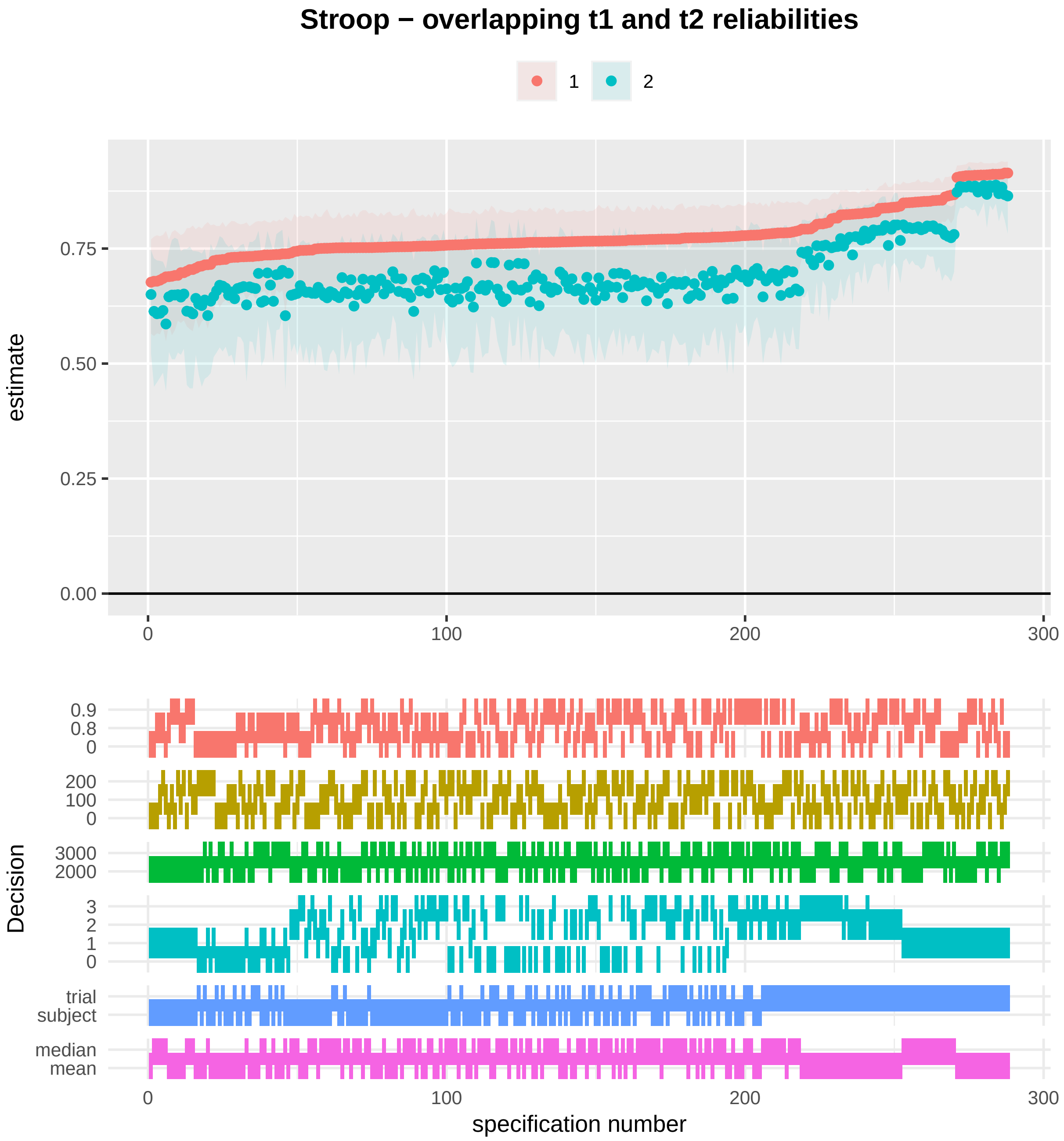

Figure 7. Overlapped internal consistency reliability multiverse for Stroop RT cost at times 1 and 2

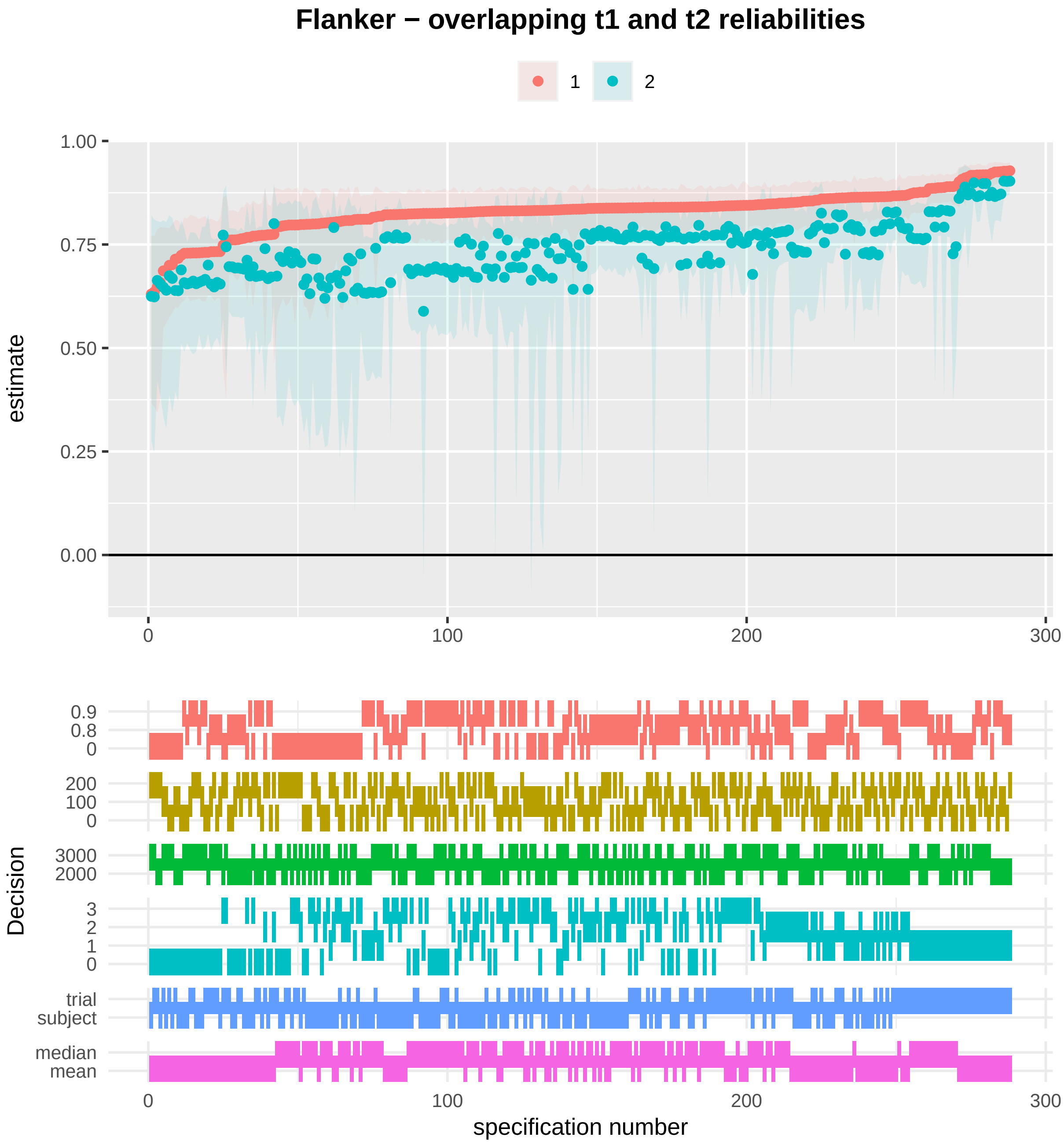

Figure 8. Overlapped internal consistency reliability multiverse for Flanker RT cost at times 1 and 2

Discussion

In the results section I presented visualisations of six multiverse analyses and an additional two visualisations overlapping the internal consistency multiverses. The range of reliability estimates was 0.24-0.31 for internal consistency and 0.16-0.43 for test-retest. In the introduction, I reminded the reader that reliability estimates are a product of the sample and the population they are drawn from, the task (including any differences in implementation), and the circumstances in which the measurement was obtained; i.e. reliability is not an inherent quality of the task itself. The first conclusion to take from these multiverse analyses is that data processing specifications are also an integral part of this list.

At the onset of this project, I thought it reasonable to assume that a particular feature of the data processing path might result in consistently higher (and lower) reliability estimates. The clearest indication we can take from these analyses is that there is no single set of data processing specifications, or combination of data processing decisions, that lead to improved reliability. The wide ranges of estimates are an additional cause for concern. Seemingly arbitrary data processing decisions can lead to differences of more than .3 in the reliability of a measure. These decisions are equally reasonable and logical choices, and we should not expect them to have meaningful impact on the theoretical questions being asked of the data. The reliability multiverse analyses presented here demonstrate this using data from a Stroop and a Flanker task. As well as across tasks, overlapping the time 1 and time 2 multiverses for both tasks highlights that even the same set of specifications does not lead to directly comparable internal consistency reliability estimates over time. data processing decisions appear to be extremely important contributors to measure reliability, but their influence is unpredictable and arbitrary.

Despite this pessimistic outlook, there is a trend across the multiverse analyses that inspires some small hope. There is a trend for increased reliability in specifications that exclude trials based on a standard deviation cutoff away from the trial mean RT, rather than away from the participant's grand average RT. The highest reliability estimates are also where a relative RT cut-off of one standard deviation from the mean is used. However, I have yet to read a paper that used such a conservative approach to removing outlier trials (approximately 30% of trials). I am not convinced that these small trends offer strong insight into how we should process task data to maximise reliability; further exploration is undoubtedly needed. In the core of this discussion I raise several open questions and suggest some plausible actions that could be taken to mitigate some of the risk reliability heterogeneity poses.

How do we guard against reliability heterogeneity?

Low reliability attenuates effect sizes estimated from tests in which the measure is used. It is therefore important to take reliability heterogeneity into account when comparing effect sizes (for several clear examples, see Cooper, Gonthier, Barch, & Braver, 2017). It is plausible that some studies may have obtained smaller or larger effect sizes than others based, in part, on the reliability of the measurements taken. Similarly, identical observed effect sizes may represent very different "true" effect sizes, if reliability is taken into account. Recently, Wiernik and Dahlke (2020) made a strong case for correcting for measurement error in meta-analyses, and provide the necessary formula and code for doing so. There are several actions we can take to begin to account for reliability heterogeneity.

Two simple recommendations. To briefly reiterate two recommendations I and my colleagues have made previously: report all data processing steps taken, and report the reliability of measures analysed (Parsons et al., 2019). These recommendations will not "fix" potential psychometric issues within one's study, or reliability heterogeniety across studies. However, complete reporting of data processing will assist in the computational reproducibility of one's results. Reporting psychometric information will assist in the interpretation of results, including comparisons of effect sizes, as well as provide useful information about the utility of a task in studies of individual differences.

Adopt a modelling approach. Incorporating trial level variation into our analyses with hierarchical modelling approaches will likely be a vital step in protecting us against reliability heterogeneity. Psychological effects are often heterogeneous across individuals (Bolger, Zee, Rossignac-Milon, & Hassin, 2019), and factors within tasks have important effects (e.g. stimuli differences; DeBruine & Barr, 2019). It follows that our models should take trial-level variation into account. Using the Stroop and Flanker data from Hedge et al. (2018) Rouder, Kumar, and Haaf (2019; also see Rouder & Haaf, 2018) demonstrated that hierarchical models should be used to account for error in measurement (for additional guidance on applying this modelling, see Haines, 2019). Adopting this approach would have the benefit of "correcting" the effect size estimate (and standard error) for measurement error as part of the model. Rouder and colleagues demonstrate that this is also a more effective approach than "correcting" the effect size estimate using e.g. Spearman's correction for attenuation formula (Spearman, 1904). Yet, even better corrections cannot fully save us from measurement error.

The benefits of adopting multilevel approaches could be demonstrated empirically with an extension of the multiverse approach adopted here. If we were to extract outcome data from the multiverse and correlate it with another variable (held constant across the multiverses), we would expect the correlation effect size to increase with increased reliability estimates. I expect that incorporating hierarchical modelling into the multiverse would lead to reduced heterogeneity across the effect size estimates as error has been accounted for in the model. Though, we should expect the confidence interval around the estimate to reduce in size with increasing reliability – reducing error is almost always a good thing. I believe this would be a strong argument for more widespread use of hierarchical models in the analysis of behavioural measures.

Limitations and room for expansion

One potential limitation of this study is the focus on only two tasks. It is possible that data from other tasks tend to yield more or less consistent patterns of reliability estimates across data processing specifications. Similarly, I have only examined RT costs (i.e. a difference score between two trial types) as the outcome measure. The analyses could have examined accuracy rates, RT averages, signal detection, and a wide variety of outcome measures. It is very possible that other outcome indices would be more or less consistently reliable across the range of data processing specifications. I opted for brevity in this paper by selecting only two tasks from Hedge et al. (2018) and looked at only RT costs as the outcome; I welcome future work seeking to examine a wider range of tasks and outcome indices.

There is a paradox in measurement reliability (see 2018): Experimental effects that are highly replicable (for example, the Stroop effect) may also show low reliability. Homogeneity within groups or experimental conditions allows for larger and more robust effects; researchers can opt to develop tasks that capitalise on homogeneity. Unfortunately, reliability requires robust individual differences (and vice versa). Highly reliable measures by necessity show consistent, potentially large, individual differences and would not be suitable for group differences or experimental research.

As a result, measures tend to be more appropriate for questions of a) assessing differences between groups or experimental conditions, or b) correlational or individual differences. I was primarily concerned with the use of these measures in individual differences research - hence the focus on reliability. Yet, it would be overly simplistic to assert that the discussions in this paper do not also relate to experimental differences questions. Indeed, the data processing specifications that maximise the measure's utility in individual differences analyses will also hinder the measure's utility in experimental quesitons. Further research would be needed to quantify the relative influences on correlational vs experimental analyses. Yet, large fluctuations in relative between-subjects vs within-subjects variance, due to data processing, holds importance for any research question.

What about validity?

Others have previously demonstrated that measures are often used ad hoc or with little reported validation efforts (e.g. Flake et al., 2017; Hussey & Hughes, 2018). This study cannot begin to assess the influence of data processing flexibility on measure validity – nor did this paper attempt to address this question. Reliability is only one piece of evidence needed to demonstrate the validity of a measure. Yet, it is an important piece of evidence as "reliability provides an upper bound for validity" (Zuo, Xu, & Milham, 2019, p. 3). While we cannot directly conclude that flexibility in data processing influences measure validity, we should look to further research to investigate. One possibility would be to conduct a validity multiverse analysis similar to the "Many Analysts, One Data Set" project (Silberzahn et al., 2018). In this project, 29 teams (61 analysts total) analysed the same dataset. The teams adopted a number of different analytic approaches which resulted in a range of results. The authors concluded that, "Uncertainty in interpreting research results is therefore not just a function of statistical power or the use of questionable research practices; it is also a function of the many reasonable decisions that researchers must make in order to conduct the research" (page 354).

Returning to the garden

My intention for this project was to provide some indication about the influence of data processing pathways on the reliability of our cognitive measurements. The influence can be profound; the multiverse analyses show large differences between the highest and lowest reliability estimates. Yet, we see little consistency in the pattern of decisions leading to higher, or lower, estimates. We have the worst of both worlds: data processing decisions are largely arbitrary, yet can have a large – relatively unpredictable – impact on the resulting reliability estimates. Briefly returning to the garden of forking paths metaphor; I imagined that this project would help illuminate the point in which our hypothetical researcher would enter the garden, based on their data processing decisions. But, our investigation has uncovered an unfortunate secret: Our researcher's forking paths are almost entirely arbitrary and interwoven. Each path diverges wildly, leading to almost anywhere in the garden. It is as if our researcher is simply spinning in dizzy circles until they stumble somewhere along the fence of reliability. I discussed several actions researchers can take collectively to help with the issue. But, by no means were these remedies to our reliability issues, nor would they directly help issues with the validity of our measurements.

I am concerned that we sit on the precipice of a measurement crisis. The so-called replication crisis shook much of our field into widespread and ongoing reforms. Yet, much of the focus has been on improving methodological and statistical practices. This is undoubtedly worthwhile, but largely omits discussion of reliability and validity of our measurements – despite our measurements forming the basis of any outcome or inference. This oversight feels like repairing a damaged wall at the same time as ignoring the shifting foundations under it. I hope that this paper, and similar work, highlights the issue and encourages researchers to place more emphasis on quality measurement. As a field, we can orchestrate a measurement revolution (cf. the "credibility revolution", Vazire, 2018) in which the quality and validity of our measurements is placed an order of importance above obtaining desired results. If the reader takes home a single message from this paper, let it be "measurement matters".

References

Auguie, B. (2017). *GridExtra: Miscellaneous functions for "grid" graphics.* Retrieved from https://CRAN.R-project.org/package=gridExtra

Aust, F., & Barth, M. (2020). *papaja: Create APA manuscripts with R Markdown.* Retrieved from https://github.com/crsh/papaja

Bolger, N., Zee, K. S., Rossignac-Milon, M., & Hassin, R. R. (2019). Causal processes in psychology are heterogeneous. *Journal of Experimental Psychology: General*, *148*(4), 601–618. https://doi.org/10.1037/xge0000558

Cooper, S. R., Gonthier, C., Barch, D. M., & Braver, T. S. (2017). The role of psychometrics in individual differences research in cognition: A case study of the AX-CPT. *Frontiers in Psychology*, *8*, 1–16. https://doi.org/10.3389/fpsyg.2017.01482

DeBruine, L. M., & Barr, D. J. (2019). *Understanding mixed effects models through data simulation* (preprint). PsyArXiv. https://doi.org/10.31234/osf.io/xp5cy

Flake, J. K., & Fried, E. I. (2019). *Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them* (preprint). PsyArXiv. https://doi.org/10.31234/osf.io/hs7wm

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct Validation in Social and Personality Research: Current Practice and Recommendations. *Social Psychological and Personality Science*, *8*(4), 370–378. https://doi.org/10.1177/1948550617693063

Fried, E. I., & Flake, J. K. (2018). Measurement matters. *Observer.* Retrieved from https://www.psychologicalscience.org/observer/measurement-matters

Gawronski, B., Deutsch, R., & Banse, R. (2011). Response interference tasks as indirect measures of automatic associations. In *Cognitive methods in social psychology.* (pp.

78–123). New York, NY, US: The Guilford Press.

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time, 17. Retrieved from https://doi.org/dx.doi.org/10.1037/a0037714

Haines, N. (2019). Thinking generatively: Why do we use atheoretical statistical models to test substantive psychological theories? Retrieved from http://haines-lab.com/post/thinking-generatively-why-do-we-use-atheoretical-statistical-models-to-test-substantive-psychological-theories/

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Henry, L., & Wickham, H. (2019). *Purrr: Functional programming tools.* Retrieved from https://CRAN.R-project.org/package=purrr

Hussey, I., & Hughes, S. (2018). Hidden invalidity among fifteen commonly used measures in social and personality psychology. https://doi.org/10.31234/osf.io/7rbfp

Jones, A., Christiansen, P., & Field, M. (2018). Failed attempts to improve the reliability of the alcohol visual probe task following empirical recommendations. *Psychology of Addictive Behaviors*, *32*(8), 922–932. https://doi.org/10.1037/adb0000414

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., … Zwaan, R. A. (2018). Justify your alpha. *Nature Human Behaviour*, *2*(3), 168–171. https://doi.org/10.1038/s41562-018-0311-x

Leek, J. T., & Peng, R. D. (2015). P values are just the tip of the iceberg. *Nature*, *520*, 612.

https://doi.org/10.1038/520612a

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, *355*(6325), 584–585. https://doi.org/10.1126/science.aal3618

Müller, K., & Wickham, H. (2019). *Tibble: Simple data frames.* Retrieved from https://CRAN.R-project.org/package=tibble

Orben, A., & Przybylski, A. K. (2019). The association between adolescent well-being and digital technology use. *Nature Human Behaviour*, *3*(2), 173–182. https://doi.org/10.1038/s41562-018-0506-1

Parsons, S. (2019). *Splithalf: Robust estimates of split half reliability.* Retrieved from https://doi.org/10.6084/m9.figshare.5559175.v5

Parsons, S. (2020). Splithalf; robust estimates of split half reliability. Retrieved from https://doi.org/10.6084/m9.figshare.5559175.v5

Parsons, S., Kruijt, A.-W., & Fox, E. (2019). Psychological Science Needs a Standard Practice of Reporting the Reliability of Cognitive-Behavioral Measurements. *Advances in Methods and Practices in Psychological Science*, *2*(4), 378–395. https://doi.org/10.1177/2515245919879695

Pedersen, T. L. (2019). *Patchwork: The composer of plots.* Retrieved from https://CRAN.R-project.org/package=patchwork

Price, R. B., Kuckertz, J. M., Siegle, G. J., Ladouceur, C. D., Silk, J. S., Ryan, N. D., … Amir, N. (2015). Empirical recommendations for improving the stability of the dot-probe task in clinical research. *Psychological Assessment*, *27*(2), 365–376. https://doi.org/10.1037/pas0000036

Quintana, D. S., & Heathers, J. (2019). A GPS in the Garden of Forking Paths (with Amy

Orben). Retrieved from 10.17605/OSF.IO/38KPE

R Core Team. (2018). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Revelle, W. (2017). *Psych: Procedures for Personality and Psychological Research.* Northwestern University, Evanston, Illinois, USA.

Revelle, W. (2019). *Psych: Procedures for psychological, psychometric, and personality research.* Evanston, Illinois: Northwestern University. Retrieved from https://CRAN.R-project.org/package=psych

Rohrer, J. M., Egloff, B., & Schmukle, S. C. (2017). Probing Birth-Order Effects on Narrow Traits Using Specification-Curve Analysis. *Psychological Science*, *28*(12), 1821–1832. https://doi.org/10.1177/0956797617723726

Rouder, J., & Haaf, J. M. (2018). A Psychometrics of Individual Differences in Experimental Tasks. https://doi.org/10.31234/osf.io/f3h2k

Rouder, J., Kumar, A., & Haaf, J. M. (2019). Why Most Studies of Individual Differences With Inhibition Tasks Are Bound To Fail. https://doi.org/10.31234/osf.io/3cjr5

Schmukle, S. C. (2005). Unreliability of the dot probe task. *European Journal of Personality*, *19*(7), 595–605. https://doi.org/10.1002/per.554

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., … Nosek, B. A. (2018). Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356. https://doi.org/10.1177/2515245917747646

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology:

Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Specification Curve: Descriptive and Inferential Statistics on All Reasonable Specifications. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2694998

Spearman, C. (1904). The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, *15*(1), 72. https://doi.org/10.2307/1412159

Staugaard, S. R. (2009). Reliability of two versions of the dot-probe task using photographic faces. *Psychology Science Quarterly*, *51*(3), 339–350.

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science*, *11*(5), 702–712. https://doi.org/10.1177/1745691616658637

Urbanek, S., & Horner, J. (2019). *Cairo: R graphics device using cairo graphics library for creating high-quality bitmap (png, jpeg, tiff), vector (pdf, svg, postscript) and display (x11 and win32) output*. Retrieved from https://CRAN.R-project.org/package=Cairo

Vazire, S. (2018). Implications of the Credibility Revolution for Productivity, Creativity, and Progress. *Perspectives on Psychological Science*, *13*(4), 411–417. https://doi.org/https://doi.org/10.1177/1745691617751884

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Wickham, H. (2019a). *Forcats: Tools for working with categorical variables (factors)*. Retrieved from https://CRAN.R-project.org/package=forcats

Wickham, H. (2019b). *Stringr: Simple, consistent wrappers for common string operations.* Retrieved from https://CRAN.R-project.org/package=stringr

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., François, R., Henry, L., & Müller, K. (2019). *Dplyr: A grammar of data manipulation.* Retrieved from https://CRAN.R-project.org/package=dplyr

Wickham, H., & Henry, L. (2019). *Tidyr: Tidy messy data.* Retrieved from https://CRAN.R-project.org/package=tidyr

Wickham, H., Hester, J., & Francois, R. (2018). *Readr: Read rectangular text data.* Retrieved from https://CRAN.R-project.org/package=readr

Wiernik, B. M., & Dahlke, J. A. (2020). Obtaining Unbiased Results in Meta-Analysis: The Importance of Correcting for Statistical Artifacts. *Advances in Methods and Practices in Psychological Science.* https://doi.org/10.1177/2515245919885611

Zuo, X.-N., Xu, T., & Milham, M. P. (2019). Harnessing reliability for neuroscience research. *Nature Human Behaviour.* https://doi.org/10.1038/s41562-019-0655-x