# A large-scale analysis of test-retest reliabilities of self-regulation measures

Running head: Retest reliabilities of self-regulation measures

A. Zeynep Enkavi1, Ian W. Eisenberg1, Patrick G. Bissett1, Gina L. Mazza2, David P. MacKinnon3, Lisa A. Marsch4, Russell A. Poldrack1

1Department of Psychology, Stanford University, Stanford, CA 94305

2Department of Health Sciences Research, Mayo Clinic, Scottsdale, AZ 85259

3Department of Psychology, Arizona State University, Tempe, AZ 85281

4 Geisel School of Medicine, Dartmouth College, Hanover, NH 03755

Author Note

These data have previously been presented as a poster at the Society for Neuroeconomics Annual meeting in Toronto (2017) and Society for Judgment and Decision Making Annual meeting in Vancouver (2017).

Correspondence concerning this article should be addressed to A. Zeynep Enkavi, Department of Psychology, Stanford University, Stanford, CA 94305.

Phone: 215-760-9290

Email: [zenkavi@stanford.edu](mailto:zenkavi@stanford.edu)

Classification: Social Sciences > Psychological and Cognitive Sciences

# Abstract

The ability to regulate behavior in service of long-term goals is a widely studied psychological construct known as self-regulation. This wide interest is in part due to the putative relations between self-regulation and a range of real-world behaviors. Self-regulation is generally viewed as a trait, and individual differences are quantified using a diverse set of measures including self-report surveys and behavioral tasks. Accurate characterization of individual differences requires measurement reliability, a property frequently characterized in self-report surveys, but rarely assessed in behavioral tasks. We remedy this gap by (1) providing a comprehensive literature review on an extensive set of self-regulation measures, and (2) empirically evaluating retest reliability in this battery of measures in a new sample. We find that self-report survey measures of self-regulation have high test-retest reliability while measures derived from behavioral tasks do not. This holds both in the literature and in our sample, though the reliability estimates in the literature are highly variable. We confirm that this is due to differences in between-subjects variability. We also compare different types of task measures (e.g., model parameters vs. raw response times) in their suitability as individual difference measures, finding that certain model parameters are as stable as raw measures. Our results provide greater psychometric footing for the study of self-regulation and provide guidance for future studies of individual differences in this domain.

Keywords: self-regulation, retest reliability, individual differences

# Significance Statement

Self-regulation is a psychological construct that is characterized using a broad set of measures and is thought to be related to a number of real-world outcomes. However, the reliability of many of these measures is unclear. This paper reviews the literature on reliability of self-regulation measures, and characterizes long-term retest reliability in a large sample of individuals completing an extensive battery of measures. The results show that while self-report measures have generally high reliability, behavioral task measures have substantially lower reliability, raising questions about their ability to serve as trait-like measures of individual differences.

# Author Contributions

LAM and RAP designed research, AZE and IWE performed research, AZE analyzed data, AZE, IWE, PGB, GLM, DPM, RAP wrote the paper

# Introduction

The ability to control behavior in service of goals, known as *self-regulation*, is a fundamental aspect of adaptive behavior and central to theories in nearly every area of psychology. Individual differences in self-regulatory ability are thought to be associated with a number of maladaptive behaviors in the real world, including drug abuse [(1, 2)](https://paperpile.com/c/mC6vv9/OVqtp+uS6V5), problem gambling [(3–6)](https://paperpile.com/c/mC6vv9/dZdJ6+xIuxT+SrOKu+hvLdT), and overeating [(7–9)](https://paperpile.com/c/mC6vv9/M1gvw+szOc4+2cTDy). Self-regulation is also thought to play a critical role in behavior change, bolstering the individual against temptations to revert to older behaviors [(1, 10, 11)](https://paperpile.com/c/mC6vv9/OVqtp+bSDtw+J8tAU), though its role as a moderator of behavior change has recently been challenged [(12)](https://paperpile.com/c/mC6vv9/Z8AN6). Self-regulation, when conceptualized as a personality trait, has generally been measured using self-report surveys that focus on various aspects of naturalistic behavior including impulsivity, sensation-seeking, goal-directedness, and risk-taking.

A central challenge for psychological science is to identify the psychological mechanisms that underlie self-regulatory functions. For example, behavioral tasks, often involving speeded choice responses, are commonly used to compare conditions and isolate component processes. Within cognitive psychology and cognitive neuroscience, there has been particular interest in isolating mechanisms involved in “cognitive control” [(13, 14)](https://paperpile.com/c/mC6vv9/jSLVA+2H13g). Candidate mechanisms include the ability to interrupt or preempt a particular behavior (know as *response inhibition*), the ability to rapidly switch between behavioral or task sets (known as *set shifting* or *switching*), and the ability to resist interference from irrelevant information (known as *resistance to distractor interference*). Similarly, researchers in the domain of decision making have focused on the ability to delay gratification in service of larger rewards in the future (known as *delay discounting*), which is thought to relate to a number of real world outcomes [(2, 15–17)](https://paperpile.com/c/mC6vv9/LC1TG+uS6V5+JnGY9+s3EUR). Given that tasks are intended to capture the mechanisms underlying self-regulation, they would be expected to relate to self-report surveys of self-regulation, but the evidence is mixed. [(18–20)](https://paperpile.com/c/mC6vv9/aJuYi+UlVQg+To7It).

One potential complicating factor in assessing the relation between behavioral task performance and self-report measures is that their psychometric features may differ. In particular, whereas the assessment of retest reliability is a nearly ubiquitous aspect of the development of survey measures, it is rarely assessed in the development of novel behavioral tasks. Further, when assessed in behavioral tasks, it has often been found to fall far short of the common criterion of 75% [(21–23)](https://paperpile.com/c/mC6vv9/mQCJW+GFLWY+1dIN2). Therefore, it is difficult to determine whether the weak relationship between different measures of self-regulation result from flawed theories or flawed operationalizations of self-regulation.

Here we report a large-scale examination of retest reliability across a broad set of self-report and behavioral task measures relevant to self-regulation and related psychological constructs. We collected retest data on a large battery of measures from 150 participants. These participants comprised a subset of a larger sample acquired in order to model the ontological structure of self-regulation (see [20, 24)](https://paperpile.com/c/mC6vv9/dEnge+To7It). We bolstered our dataset with an extensive analysis of the relevant literature for each measure. This allowed us to both compare our data to the literature and assess the relative reliability of data acquired online compared to in-lab samples. Although previous work suggested that data acquired online can exhibit high reliability [(25–29)](https://paperpile.com/c/mC6vv9/141d4+GWDhp+LYm8V+nhBvi+fLvbm), it has not encompassed the breadth of measures relevant to self-regulation collected here. Additionally, the use of a relatively long retest delay (2-4 months) placed the work on the timescale of many studies of behavioral change, providing information relevant to the stability of pre-post intervention comparisons of self-regulatory function. Moreover, using the raw data allowed us to characterize the underlying causes of systematic differences between measure types by isolating the sources of variance for each measure.

With our new dataset we first compared differences between measure modalities (surveys vs. tasks) and recapitulated effects we found in the literature. Then we expanded our analyses to novel comparisons. For example, we compared relative reliability of performance measures quantified using raw variables versus model-based decompositions. We fit the drift-diffusion model (DDM), which transforms raw reaction times and accuracies to the more interpretable latent variables of drift rate (processing speed), threshold (caution that captures speed-accuracy tradeoffs), and non-decision time (perceptual and response execution process).

Another dimension of interest for the behavioral task measures was whether contrast measures (subtraction of one condition from another) intended to isolate putative cognitive processes also served as good trait measures. This subtraction logic is a common strategy when using behavioral tasks, both for raw measures and model parameters. Yet, subtraction of random variables mathematically implies an increase in the contrast measures’ error variance and therefore lower retest reliability. We empirically assessed the severity of this decreased reliability for common task contrasts.

By combining an analysis of the literature with a new large dataset involving the largest battery of self-regulation measures to date, we provide a comprehensive picture of the stability of measures of self-regulation.

# Results

## Analysis of prior literature

Our literature review contained 171 dependent measures, 154 papers, 17550 participants and 583 data points on retest reliability (Fig. 1). We first tested for systematic differences between dependent measures from tasks and surveys in the literature. Studies reporting retest reliability for surveys had on average 48 more subjects than those reporting retest reliabilities for tasks (95% credible interval for difference = [28, 70]). We then examined whether sample size and retest delay (see Methods) were associated with the retest reliability of a measure. Using a model including the sample size along with an indicator variable for task vs. survey measures, we found that task measures’ reliability estimates were on average 0.139 lower compared to survey measures’ (95 % credible interval of difference = [-0.192, -0.088]; mean retest reliability for task measures in the literature = 0.610, for survey measures = 0.716) and that retest reliability decreased by 0.0001 for every additional participant in a study (95% credible interval for decrease = [-0.0002, -0.00001]). To our knowledge, this is the first documentation of such a bias in the literature with respect to reliability measures,which may reflect publication bias and/or variation in undocumented decisions taken by researchers, as discussed further below.

**Fig. 1.** Summary of the literature review for tasks (left) and surveys (right). Each point represents a study containing test-retest reliability data on an unspecified dependent measure for a given task. The size of the point depends on the sample size of the study and the shape depends on the metric that was used to estimate reliability. Each vertical red line indicates 0 reliability.



## Analysis of new dataset

### Data Quality Checks

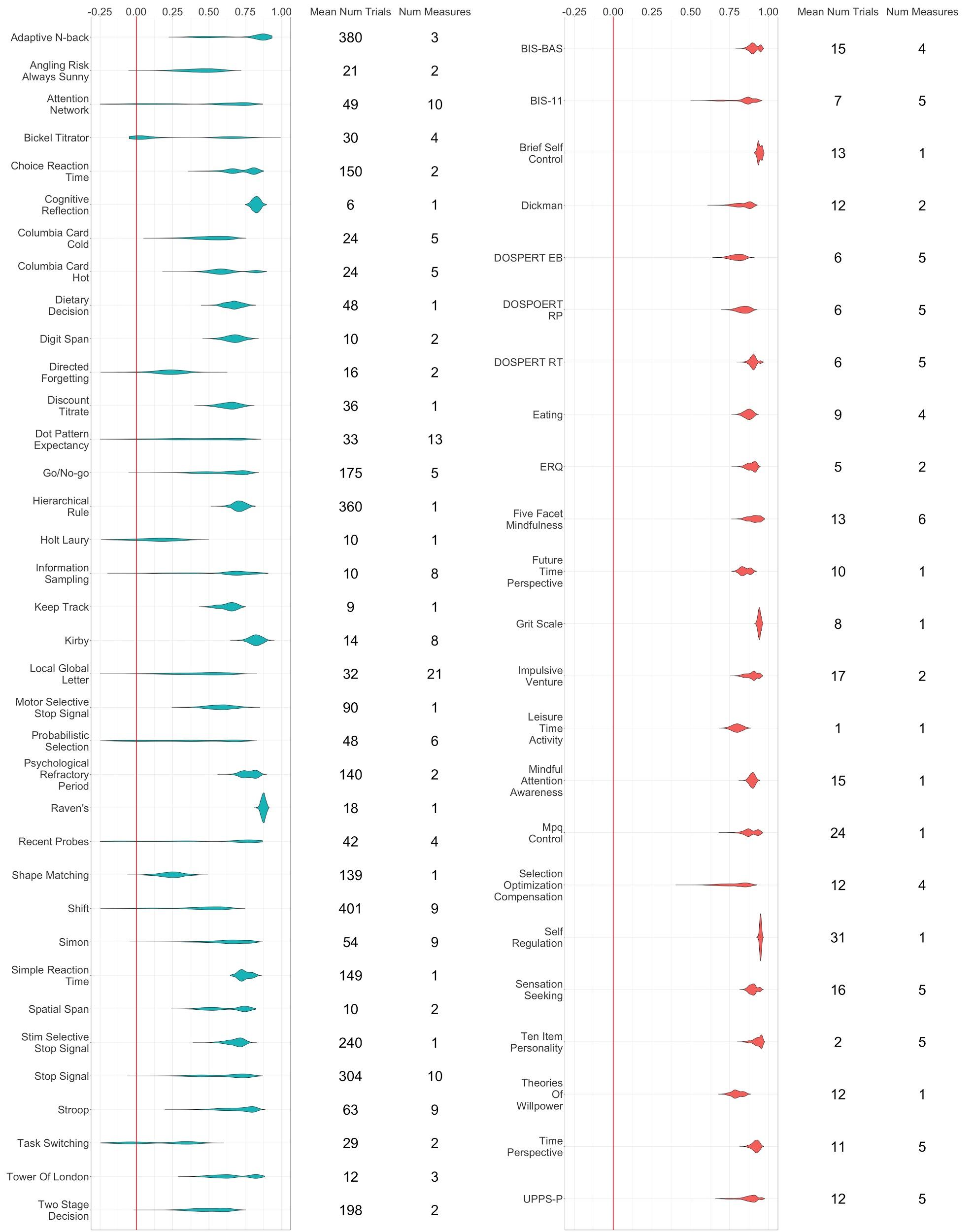
To ensure data quality we conducted three tests that are described in more detailed in SI: We checked the reliability of the demographic items in our battery, the effect of retest delay on change of subject scores and the correlation between similar survey items. None of our analyses raised concerns and overall provided some degree of assurance that the participants were real people and not automated machines (which is a concern given that the data were collected using Amazon Mechanical Turk).

### Survey and behavioral task reliability in new data

We calculated 372 dependent measures for behavioral tasks and 74 for surveys. Retest reliabilities for each measure were estimated using a nonparametric bootstrap (1000 samples); statistics on these bootstrapped estimates are reported instead of point estimates. We report ICC(3,k) as the main metric of retest reliability, based on its ability to account for various sources of variance separately as outlined in the Methods[[1]](#footnote-1). The ICC, which ranges from -1 to 1, is a preferred metric for retest reliability and is not biased by sample size [(30)](https://paperpile.com/c/mC6vv9/3YkY6). Larger values mean that the two scores of a subject for a given measure are more similar to each other than they are to the scores of other subjects. None of our conclusions change using other reliability metrics. The correlation between point estimates of the different reliability metrics for each measure ranged from 0.932 to 0.980 (see Fig. S3 for scatter plots of different reliability metrics).

Mirroring the results in the literature, the average behavioral task measure reliability was 0.391 lower than the average survey measure reliability (95% credible interval for difference = [-0.451, -0.332]). While survey measures had a median ICC of 0.886 (first quartile 0.708, third quartile 0.958), cognitive measures had a median of 0.544 (first quartile -0.140, third quartile 0.843).

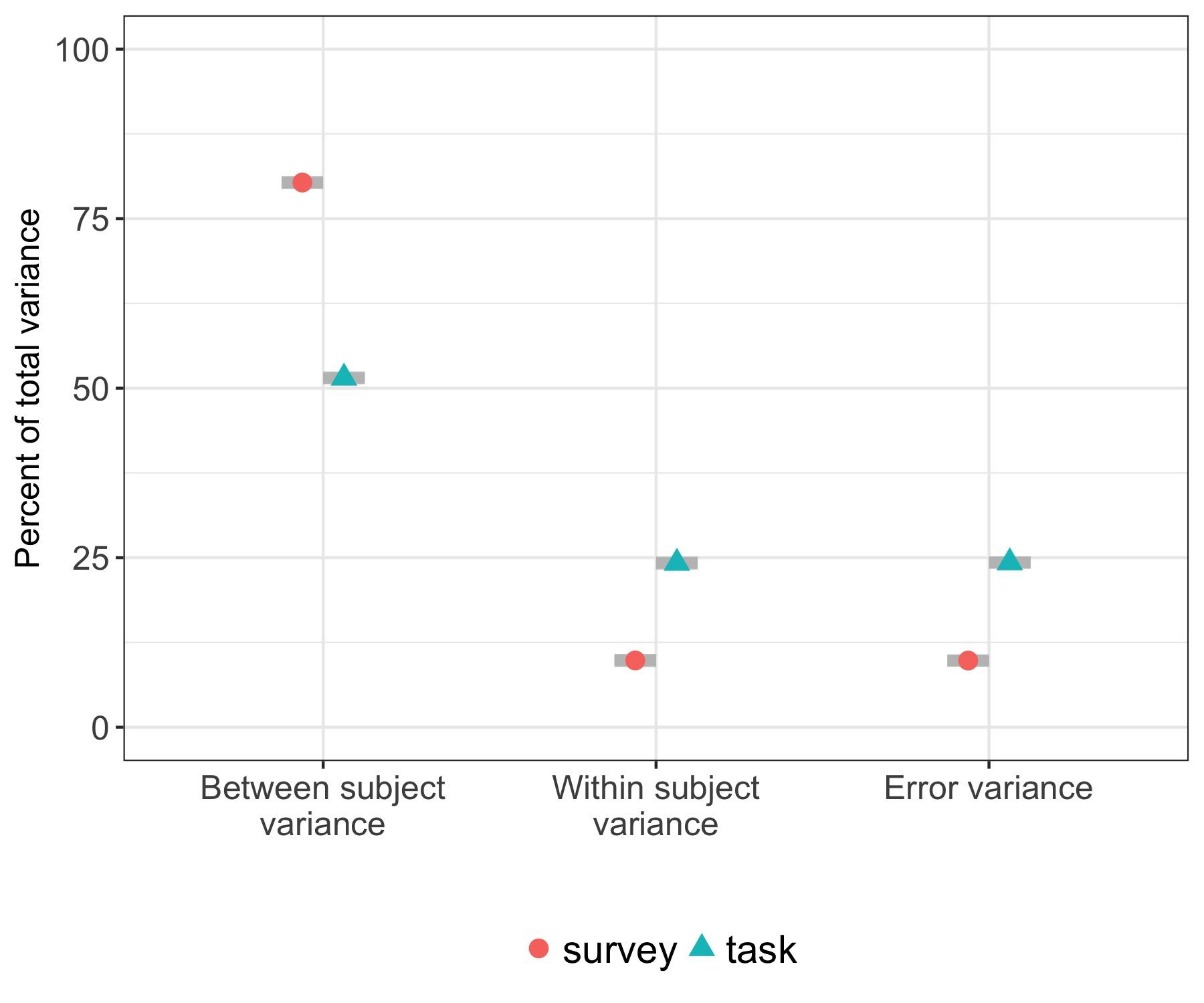
**Fig. 2.** Summary of bootstrapped reliability estimates for tasks (left) and surveys (right). We sampled 150 subjects with replacement 1000 times to create a distribution of bootstrapped reliability estimates for each measure. Reliability measure depicted is ICC. Each vertical red line indicates 0 reliability. Columns following each graph present the mean number of trials used for dependent measures in that task and the number of ‘meaningful measures,’ dependent measures used in the literature from each task.



A quantitative explanation for the difference in reliability estimates between surveys and tasks, as recently detailed by Hedge et al. (2017), lies in the difference in sources of variance between these measures. Specifically, the ICC is calculated as the ratio of between-subjects variance versus total variance. Intuitively, measures with high between-subjects variance are better suited for individual difference analyses as they will be sensitive to the differences between the subjects in a sample. Conversely, as Hedge et al. note, behavioral tasks are generally selected on the basis of reliable group effects, which systematically selects for measures with low between-subject variance.

We find that on average 83.36% of survey measures’ variance is due to between subjects variability compared to 52.55% of behavioral task measures’ (95% credible interval of difference = [25.2, 32.1]; Fig. 5). Conversely, 18.07% of behavioral tasks variance is explained by within-subject variance compared to 5.38% of survey measures (systematic differences between sessions; 95% credible interval of difference = [10.55, 18.36]) and 22.82% by residual variance compared to 8.976% for survey measures (95% credible interval difference = [12.29, 16.69]; model includes random effects for each measure).

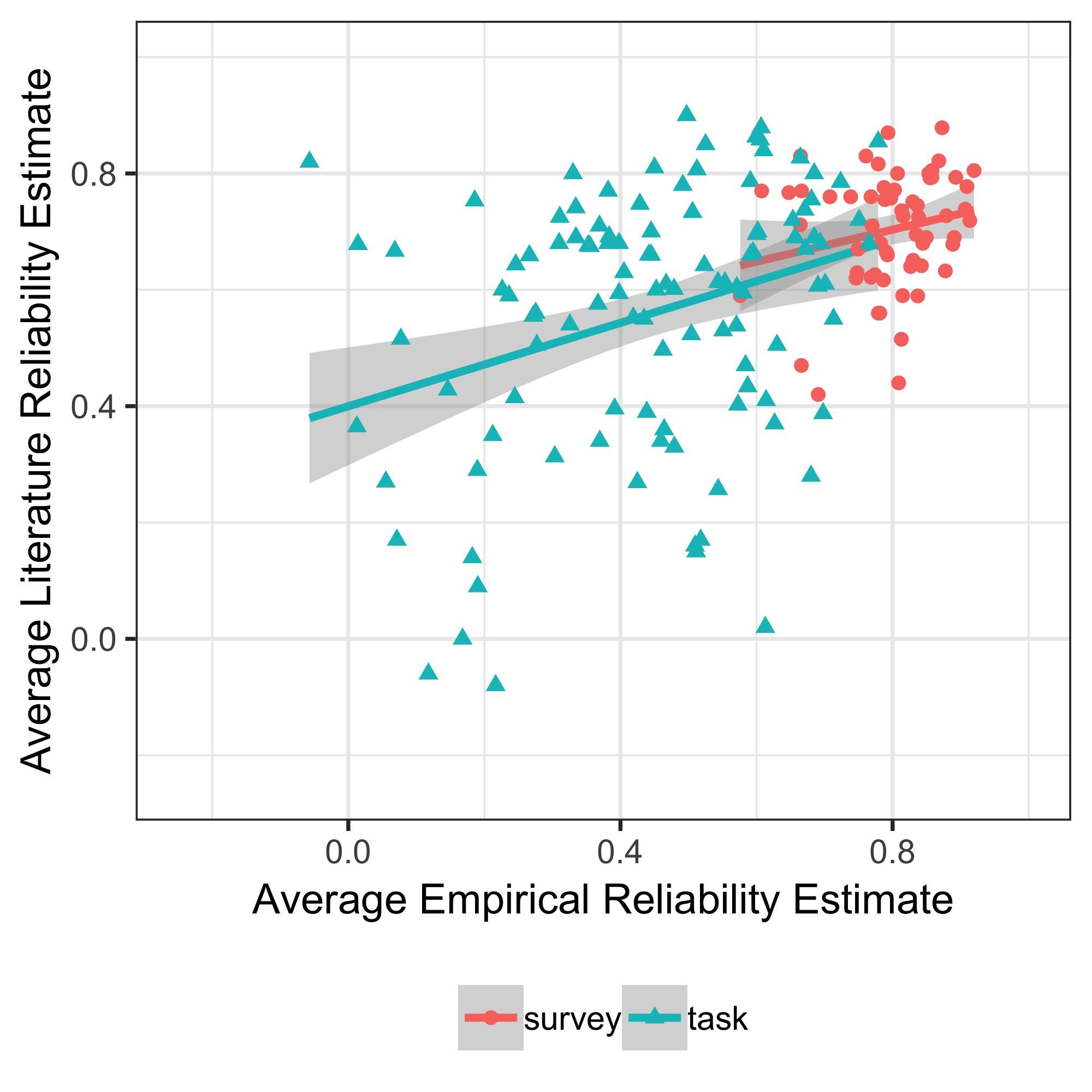
**Fig. 3.** Percentage of variance explained by each of the three sources of variance: between subjects, within subjects (between sessions) and error variance for 1000 bootstrapped samples. Error bars are 95% confidence intervals. The confidence intervals for the bootstrapped samples are quiet small and thus appear like lines compared to the larger markers depicting the means.



### Comparison of literature and new data

To compare our findings to the literature, we first sampled the same number of estimates from our bootstrapped results as we found in the literature for each measure and calculated the correlation between the sampled empirical (i.e. from our data) reliability estimates to those found in the literature. Repeating this 100 times we found that the mean correlation (Fig 2) between our empirical reliability estimates and those based on the literature was 0.274 for behavioral task measures (range = 0.235 - 0.323) and 0.138 for survey measures (range = 0.038 - 0.249).

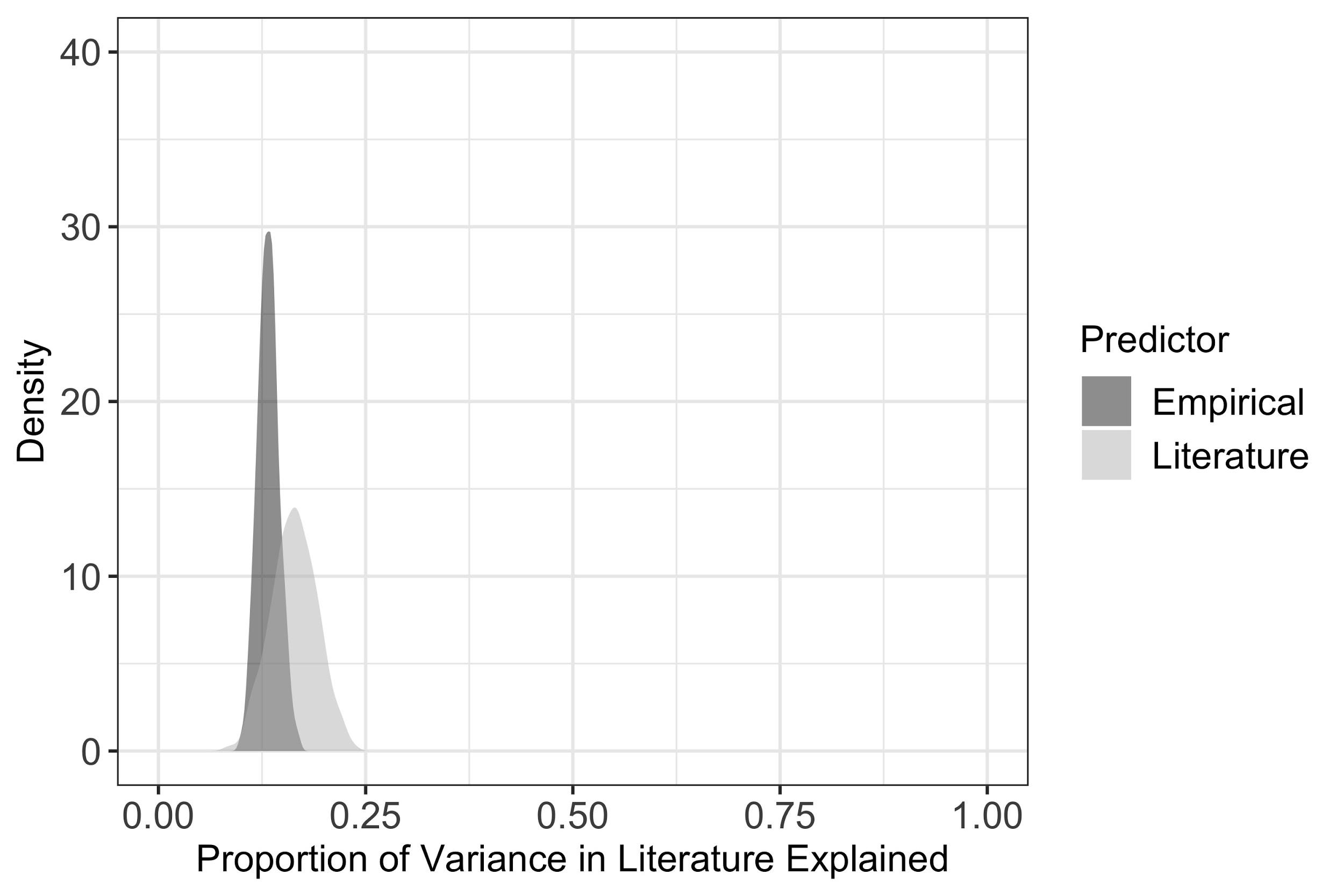
**Fig. 4.** Correlation between mean reliability estimates for each measure found in the literature with the mean reliability from our data.



While the relationship between the empirical and literature-derived reliability estimates seems weak, this must be contextualized by evaluating the variability of retest reliability estimates in the literature. If individual studies in the literature have similarly weak relationships to the literature-wide retest reliability for a given measure (i.e. if the variance of the reliability estimates reported in the literature for a given measure is large), this suggests a general issue of variability in retest reliability estimates across samples and not a specific issue with our sample. Therefore, we compared two types of models: (1) One where we predicted the literature retest reliability using an estimate sampled from the literature review. (2) Another where we predicted the literature retest reliability using the estimate from the new data we collected.

Fig. 3 shows that models using an estimate from the literature to predict the remaining reliability estimates from the literature are systematically better than models using the estimate from our sample. However, the decrease in variance explained using our data is only 2.95% (95% credible interval of difference = [2.78% - 3.13%]) on average, suggesting that published estimates of retest reliability in this domain are relatively noisy. This also suggests that estimating reliability using an online sample does not change conclusions compared to in-lab samples.

**Fig. 5.** Noise ceiling for comparing empirical retest reliability estimates to reliability estimates from the literature. Data come from sampling a single reliability estimate from the literature and using that as a predictor of the remaining retest reliability estimates from the literature versus using the mean estimate from our empirical results as a predictor. Models account for the effect of sample size in the literature and whether the measure is a task or survey variables. The literature samples are significantly better at predicting the rest of the literature than our empirical averages, but the distributions of variance explained across samples are highly overlapping.



### Effect of task length on stability

To compare potential effects of task-specific attributes on retest reliability across tasks, we examined the relationship between the number of trials a task included and its stability. Across non-DDM[[2]](#footnote-2) measures, there was an insignificant 0.023 point change increase in reliability for each additional trial[[3]](#footnote-3) (95% credible interval of increase = [-0.011, 0.061]).

For tasks for which dependent measures are estimated using many trials one can ask whether the same measure becomes less reliable if fewer trials are used to estimate its reliability. Such analyses would provide a detailed examination of how to extract the most reliable individual difference measure from tasks with measures that have low retest reliabilities, and would address the concern that reliability might be underestimated in the present data due to insufficient numbers of trials. It could also guide researchers in choosing number of trials for long tasks in an informed manner. We provide an example of this approach in the SI; given the open access nature of the data, investigators interested in other tasks can perform similar analyses on those.

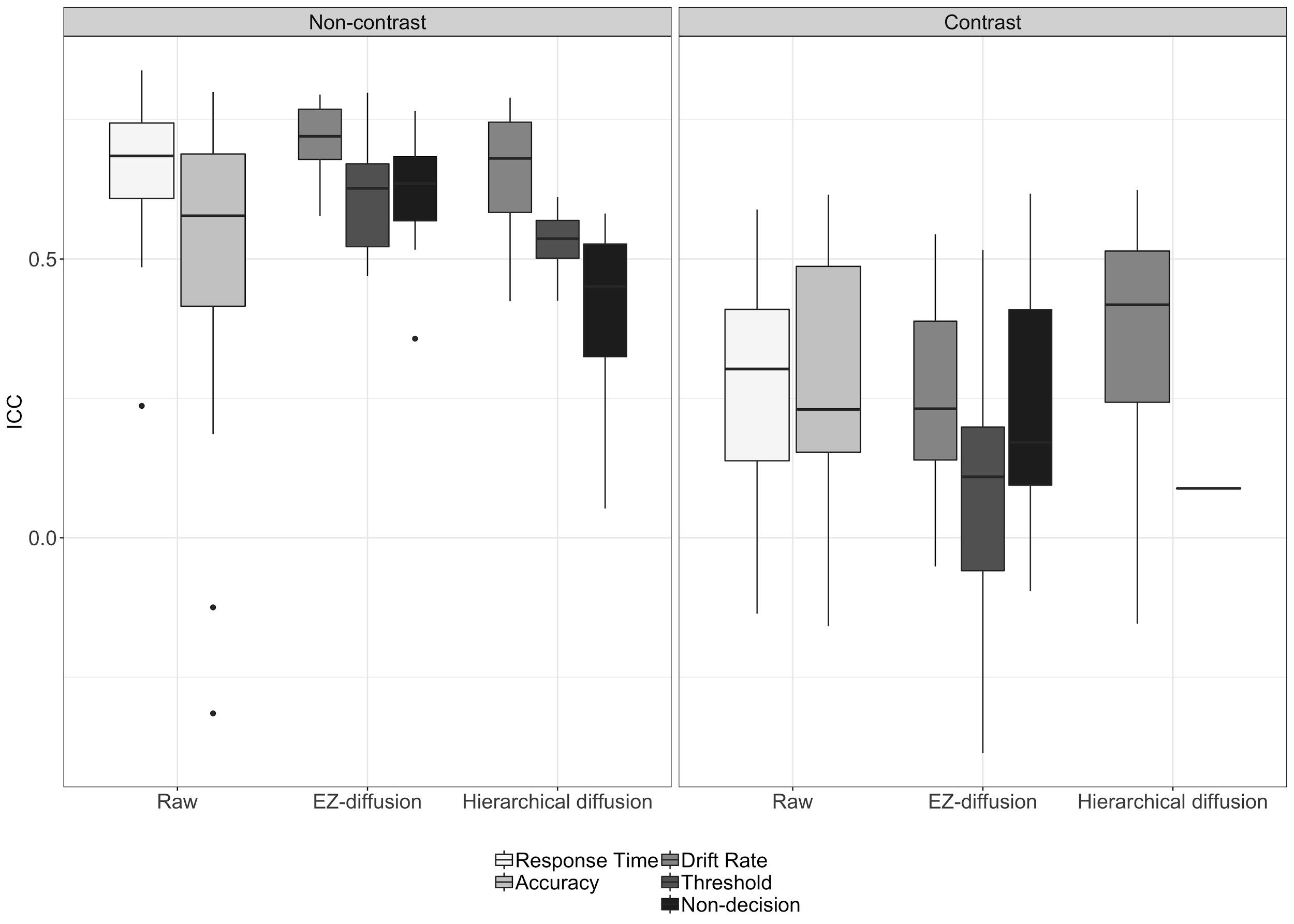
### Comparison of task measure types

Data from any given behavioral task can often be analyzed in various ways, yielding different types of dependent measures. We compare raw measures of response times and accuracies to parameters of the drift diffusion model, which addresses speed-accuracy tradeoffs and offers more interpretable latent variables from a rich literature in mathematical and experimental psychology [(31, 32, 33)](https://paperpile.com/c/mC6vv9/hvdGf+JtOPi). There are different fitting procedures within this class of models. We chose two approaches to DDM: EZ-diffusion [(34)](https://paperpile.com/c/mC6vv9/jtWtJ) and hierarchical drift diffusion model (HDDM) [(35)](https://paperpile.com/c/mC6vv9/HIOq8).

The EZ-diffusion method is a set of closed-form expressions that transform mean RT, variability of RT, and accuracy to drift rate, threshold, and non-decision time. The HDDM uses hierarchical Bayesian modeling to allow simultaneous estimation of both group and individual subject parameters. We compare the reliability estimates of raw measures (response times and accuracies) to DDM parameter estimates. Both types of measures can also be ‘contrast’ and ‘non-contrast.’

We found that estimates of non-contrast measures (Fig. 6) are on average 0.356 points more reliable than estimates of contrast measures (95% credible interval of difference = [0.311, 0.406]). This is not surprising given the summing of the variance in the difference score. Of concern, however, is the fact that contrast measures had low to no reliability (mean = 0.245, SD = 0.242) compared to the moderate reliability of the non-contrast measures (mean = 0.601, SD = 0.183). This is particularly alarming given their common use in cognitive psychology as putative trait measures of cognitive constructs and predictors of real-world outcomes.

**Fig. 6.** Average bootstrapped reliability estimates per measure comparing raw measures and model parameters as well as contrast and non-contrast measures for task measures.



Although individual measures from tasks are not appropriate for individual difference analyses this does not preclude other ways of using them as trait measures. One way to do this is to use a data-driven approach to integrate them and extract scores that may be more stable though not necessarily more predictive of real-world behaviors. An example of this using the same dataset reported in this paper is detailed in Eisenberg et al. [(20)](https://paperpile.com/c/mC6vv9/To7It/?noauthor=1): Factor scores computed at both time points using the same linear combination of dependent measures correlated highly with each other for 5 task factors (*M* = .81, min = .76, max = .86) and 12 survey factors (*M* = .86, min = .75, max = .95). Yet despite adequate reliability for both task and survey factors, only surveys predicted a significant amount of variance in real-world behaviors out of sample (average R2 = .10) whereas tasks did not, either as factors or as separate dependent measures (average R2 = .02)

### Effect of survey length on stability

Mirroring the task analysis, we examined the relationship between the number of items in a survey and its stability. Each additional item used in the calculation of a subscale was associated with an insignificant 0.001 increase in retest reliability (95% credible interval = [-0.001, 0.004]) though as with tasks, surveys could also be analyzed in more detail using item response theory or more sophisticated models given the open availability of the data.

# Discussion

This report provides a systematic characterization of the reliability of self-report and behavioral task measures of the construct of self-regulation. We first summarized the prior literature on the retest reliability of different types of self-regulation measures. We found that while psychometric studies of survey measures have larger sample sizes than task measures, reliability estimates generally decreased with sample size. On the one hand, this might suggest that smaller studies afford researcher's more control over their measurement, leading to higher reliability. On the other hand, larger sample sizes might be more reflective of the truly lower stability of measures; Hopkins [(30)](https://paperpile.com/c/mC6vv9/3YkY6/?noauthor=1) suggests that studies of retest reliability with samples smaller than 50 should be treated as pilot studies for this reason. Studies with smaller samples are more prone to yield variable reliability estimates and coupled with publication bias might inflate the results in the literature. The majority of the results in the literature, particularly those on tasks, have sample sizes in the <50 range. Our new data acquisition had a large sample size and found relatively low reliability of behavioral tasks, consistent with the results from the literature.

Second, we contextualized new results from our battery of self-regulation measures using an extensive literature review. We estimated the general reliability of the literature's own estimates of the retest reliability of self-regulation. This provided a sense of the "noise ceiling" of reliability studies, and a reference point for the expected relationship between any two sets of reliability estimates. Because the literature reliability estimates lacked a strong coherence for many measures, their low correlations with our reliability estimates led to a less than 3% decrease in the predictability of prior literature, suggesting that the results reported on the new dataset here are not far outside what one would expect from the literature.

## Systematic differences in the reliability of self-regulation measures

Literature analysis and our data show that measures of self-regulation that are based on self-report surveys have higher retest reliability than behavioral task measures because of higher between subject variance for survey measures compared to task measures. This suggests that survey measures are more appropriate as trait-level measures suitable for individual difference analyses. Exploratory analyses on task measures suggested that the reliability of DDM parameters did not significantly differ from the reliability of raw measures like response times and accuracy. Researchers may therefore prefer drift diffusion measures given their interpretability.

Revisiting a longstanding question on the reliability of contrast scores, we confirm that they are less reliable than their components in this study. Measures of differences in response times between conditions have lower reliabilities due to correlations between the two measures used in the creation of the difference score [(36, 37)](https://paperpile.com/c/mC6vv9/ZEwGX+h6Inn) and the increase in the variance through subtraction. The novel and concerning point of this finding is that many cognitive measures of interest in the self-regulation literature are contrast measures that have low to no reliability.

## Implications of low reliability for behavioral task measures

Although our conclusion that task measures of self-regulation are less suitable for individual difference analyses might be disappointing, especially in the face of many lines of work showing correlations between these measures and problematic real-world behaviors, it should not be surprising. As Hedge and colleagues [(38)](https://paperpile.com/c/mC6vv9/4DBuR/?noauthor=1) argue, behavioral tasks designed with the subtraction logic to isolate specific cognitive processes become well-established in the literature precisely for their low between-subject variability, which necessitates that they will have low retest reliability. For example, one might repeatedly find a significant Stroop effect (difference in the response times between the congruent and incongruent conditions) in samples measured multiple times, even while the relative distribution of individual response times for the subjects differ. In other words, the task might have low between subject variability and high within subject (between session) variability resulting in low test-retest reliability. This does not invalidate the existence of the Stroop effect but does undermine its suitability as a trait measure. Detailed analyses of sources of variance (within versus between subjects) provides researchers with a priori hypotheses on which measures to expect significant changes in different experimental designs. MacLeod et al. [(39)](https://paperpile.com/c/mC6vv9/PMR3p/?noauthor=1) provide an example where they hypothesize that one of three attentional networks from the ANT task is best suited for detecting significant changes in within-subjects designs due to its low within-subject variance but least suited for detecting significant changes in between-subject designs due to its high between-subject variance. Furthermore, task measures can be integrated using data-driven approaches to extract factor scores that are more stable and potentially more suitable for trait-like treatment. Using this approach, we indeed found more stable dependent measures, though they were not more predictive of real-world behaviors [(20)](https://paperpile.com/c/mC6vv9/To7It).

On the other hand, different psychometric properties of measures serve different purposes. For example, while high retest reliability is desirable for measures that will be used in trait-like characteristic analyses, it is neither a necessary nor a sufficient condition for the responsiveness of a measure to capture change over time [(40)](https://paperpile.com/c/mC6vv9/IXrnc). Although our results provide practical guidelines for researchers interested in these measures they do not answer how these measures relate to the construct of “self-regulation.” While the retest reliability of a measure has consequences on the limits of its correlation with other measures, specifically for any two variables the correlation between them must be smaller than the square root of the reliability of each measure [(36, 41, 42)](https://paperpile.com/c/mC6vv9/ZEwGX+CKaZb+Iht5n), the question of validity remains a separate one that we address in related work [(20)](https://paperpile.com/c/mC6vv9/To7It).

## Conclusions

Self-regulation is a central construct in many theories of behavior and is often targeted by interventions to reduce or control problem behavior. Our study of self-report and task measures of self-regulation suggests stability in many self-report measures and less stability in behavioral task measures. We hope that these analyses and open data provide guidance for future individual difference work in self-regulation.

## Materials & Methods

## Sample

Participants in this study were a subset from a larger study of self-regulation [(24)](https://paperpile.com/c/mC6vv9/dEnge) conducted on Amazon Mechanical Turk (MTurk). Invitations were sent to 242 of 522 participants (52% female, age: mean = 34.1, median = 33, range = 21-60) who had satisfactorily completed the first wave of data collection between July and September 2016. The final sample for the retest study consisted of 150 participants (52.7% female, age: mean = 34.5, median = 33, range = 21-60) whose data passed basic quality checks as described in Table S1. The sample size was specified prior to data collection based on financial constraints. Instead of inviting all 522 eligible participants at once we invited randomly selected subsets of participants in small batches. This addressed preferentially sampling the most motivated, prompt subjects who may systematically differ from the full sample. Each batch was given a week to complete the battery. Data collection for the second wave took place between November 2016 and March 2017. The mean number of days between the two waves was 111 days (median = 115 days ; range = 60-228 days). Of the 242 participants invited 175 participants started the battery and 157 completed the battery. This study was approved by the Stanford Institutional Review Board (protocol IRB-34926).

The data collection platform as well as the details of the data analysis pipeline including links to analysis scripts and interactive visualizations are listed in the SI.

## Literature review

The literature review was conducted on Google Scholar, which was chosen for its breadth. Our strategy consisted of the following steps: (1) Manually check the reference article for a given task or survey (i.e. the article that described the task or survey for the first time) for retest reliability data. (2) Search within the full text of the articles that cite the reference article for the term ‘retest.’ (3) Examine each of these resulting articles up to the first 100 results ordered by the number of times they have been cited. (4) Scan the abstract and the methods sections of each article to determine whether the article reports original empirical retest results. (5) Extract the empirical results from the Results section making sure to include the following information: (a) the type of retest reliability statistic, (b) the magnitude of the statistic, (c) the dependent measure the retest reliability data pertains to, (d) delay between the two measurements, (e) sample size, (f) any differences from the procedure used in our battery, and (g) Article reference. (6) If the resulting article cites other articles with retest reliability for the measure, then find and examine them for retest reliability data using the same method as above. (7) If the reference article describes a version of the task that is modified for specific purposes (e.g. the Shift Task is a modified version of the older Wisconsin Card Sorting task) then find the reference article for the parent task and apply the same search routine for retest reliability on the parent task.

Detailed descriptions of all behavioral tasks and surveys as well as the findings on retest reliability for all of them are listed in the SI.

# Acknowledgements

This work was supported by the National Institutes of Health (NIH) Science of Behavior Change Common Fund Program through an award administered by the National Institute for Drug Abuse (NIDA) (UH2DA041713; PIs: Marsch, LA & Poldrack, RA). Additional support was provided by NIDA P30DA029926.

# 

# References

1. [Prochaska JO, DiClemente CC, Norcross JC (1992) In search of how people change. Applications to addictive behaviors. *Am Psychol* 47(9):1102–1114.](http://paperpile.com/b/mC6vv9/OVqtp)

2. [Kirby KN, Petry NM, Bickel WK (1999) Heroin addicts have higher discount rates for delayed rewards than non-drug-using controls. *J Exp Psychol Gen* 128(1):78.](http://paperpile.com/b/mC6vv9/uS6V5)

3. [Alessi SM, Petry NM (2003) Pathological gambling severity is associated with impulsivity in a delay discounting procedure. *Behav Processes* 64(3):345–354.](http://paperpile.com/b/mC6vv9/dZdJ6)

4. [Kertzman S, et al. (2008) Go--no-go performance in pathological gamblers. *Psychiatry Res* 161(1):1–10.](http://paperpile.com/b/mC6vv9/xIuxT)

5. [Lawrence AJ, Luty J, Bogdan NA, Sahakian BJ, Clark L (2009) Impulsivity and response inhibition in alcohol dependence and problem gambling. *Psychopharmacology*  207(1):163–172.](http://paperpile.com/b/mC6vv9/SrOKu)

6. [Fuentes D, Tavares H, Artes R, Gorenstein C (2006) Self-reported and neuropsychological measures of impulsivity in pathological gambling. *J Int Neuropsychol Soc* 12(6):907–912.](http://paperpile.com/b/mC6vv9/hvLdT)

7. [Nederkoorn C, Smulders FTY, Havermans RC, Roefs A, Jansen A (2006) Impulsivity in obese women. *Appetite* 47(2):253–256.](http://paperpile.com/b/mC6vv9/M1gvw)

8. [Nederkoorn C, Braet C, Van Eijs Y, Tanghe A, Jansen A (2006) Why obese children cannot resist food: the role of impulsivity. *Eat Behav* 7(4):315–322.](http://paperpile.com/b/mC6vv9/szOc4)

9. [Hendrickson KL, Rasmussen EB (2013) Effects of mindful eating training on delay and probability discounting for food and money in obese and healthy-weight individuals. *Behav Res Ther* 51(7):399–409.](http://paperpile.com/b/mC6vv9/2cTDy)

10. [Rozensky RH, Bellack AS (1974) Behavior change and individual differences in self-control. *Behav Res Ther* 12(3):267–268.](http://paperpile.com/b/mC6vv9/bSDtw)

11. [Bickel WK, Vuchinich RE (2000) *Reframing Health Behavior Change With Behavioral Economics* (Psychology Press).](http://paperpile.com/b/mC6vv9/J8tAU)

12. [Stautz K, Zupan Z, Field M, Marteau TM (2018) Does self-control modify the impact of interventions to change alcohol, tobacco, and food consumption? A systematic review. *Health Psychol Rev* 12(2):157–178.](http://paperpile.com/b/mC6vv9/Z8AN6)

13. [Miller EK, Cohen JD (2001) An integrative theory of prefrontal cortex function. *Annu Rev Neurosci* 24:167–202.](http://paperpile.com/b/mC6vv9/jSLVA)

14. [Miyake A, et al. (2000) The Unity and Diversity of Executive Functions and Their Contributions to Complex “Frontal Lobe” Tasks: A Latent Variable Analysis. *Cogn Psychol* 41(1):49–100.](http://paperpile.com/b/mC6vv9/2H13g)

15. [Mischel W, Shoda Y, Rodriguez MI (1989) Delay of gratification in children. *Science* 244(4907):933–938.](http://paperpile.com/b/mC6vv9/LC1TG)

16. [Baker F, Johnson MW, Bickel WK (2003) Delay discounting in current and never-before cigarette smokers: similarities and differences across commodity, sign, and magnitude. *J Abnorm Psychol* 112(3):382.](http://paperpile.com/b/mC6vv9/JnGY9)

17. [Meier S, Sprenger CD (2012) Time discounting predicts creditworthiness. *Psychol Sci* 23(1):56–58.](http://paperpile.com/b/mC6vv9/s3EUR)

18. [Duckworth AL, Kern ML (2011) A Meta-Analysis of the Convergent Validity of Self-Control Measures. *J Res Pers* 45(3):259–268.](http://paperpile.com/b/mC6vv9/aJuYi)

19. [Nęcka E, Gruszka A, Orzechowski J, Nowak M, Wójcik N (2018) The (In)significance of Executive Functions for the Trait of Self-Control: A Psychometric Study. *Front Psychol* 9:1139.](http://paperpile.com/b/mC6vv9/UlVQg)

20. [Eisenberg IW, et al. (2018) Uncovering mental structure through data-driven ontology discovery. Available at:](http://paperpile.com/b/mC6vv9/To7It) <https://psyarxiv.com/fvqej/>[.](http://paperpile.com/b/mC6vv9/To7It)

21. [Cicchetti DV, Sparrow SA (1981) Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. *Am J Ment Defic* 86(2):127–137.](http://paperpile.com/b/mC6vv9/mQCJW)

22. [Fleiss JL, Levin B, Paik MC (2013) *Statistical Methods for Rates and Proportions* (John Wiley & Sons).](http://paperpile.com/b/mC6vv9/GFLWY)

23. [Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. *Biometrics* 33(1):159–174.](http://paperpile.com/b/mC6vv9/1dIN2)

24. [Eisenberg IW, et al. (2017) Applying novel technologies and methods to inform the ontology of self-regulation. *Behav Res Ther*.](http://paperpile.com/b/mC6vv9/dEnge)

25. [Paolacci G, Chandler J, Ipeirotis PG (2010) Running Experiments on Amazon Mechanical Turk. Available at:](http://paperpile.com/b/mC6vv9/141d4) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1626226> [[Accessed May 3, 2018].](http://paperpile.com/b/mC6vv9/141d4)

26. [Horton JJ, Rand DG, Zeckhauser RJ (2011) The online laboratory: conducting experiments in a real labor market. *Exp Econ* 14(3):399–425.](http://paperpile.com/b/mC6vv9/GWDhp)

27. [Buhrmester M, Kwang T, Gosling SD (2011) Amazon’s Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspect Psychol Sci* 6(1):3–5.](http://paperpile.com/b/mC6vv9/LYm8V)

28. [Behrend TS, Sharek DJ, Meade AW, Wiebe EN (2011) The viability of crowdsourcing for survey research. *Behav Res Methods* 43(3):800–813.](http://paperpile.com/b/mC6vv9/nhBvi)

29. [Crump MJC, McDonnell JV, Gureckis TM (2013) Evaluating Amazon’s Mechanical Turk as a tool for experimental behavioral research. *PLoS One* 8(3):e57410.](http://paperpile.com/b/mC6vv9/fLvbm)

30. [Hopkins WG (2000) Measures of reliability in sports medicine and science. *Sports Med* 30(1):1–15.](http://paperpile.com/b/mC6vv9/3YkY6)

31. [Wickelgren WA (1977) Speed-accuracy tradeoff and information processing dynamics. *Acta Psychol*  41(1):67–85.](http://paperpile.com/b/mC6vv9/icFRQ)

32. [Ratcliff R (1978) A theory of memory retrieval. *Psychol Rev* 85(2):59.](http://paperpile.com/b/mC6vv9/hvdGf)

33. [Ratcliff R, Smith PL, Brown SD, McKoon G (2016) Diffusion Decision Model: Current Issues and History. *Trends Cogn Sci* 20(4):260–281.](http://paperpile.com/b/mC6vv9/JtOPi)

34. [Wagenmakers E-J, van der Maas HLJ, Grasman RPPP (2007) An EZ-diffusion model for response time and accuracy. *Psychon Bull Rev* 14(1):3–22.](http://paperpile.com/b/mC6vv9/jtWtJ)

35. [Wiecki TV, Sofer I, Frank MJ (2013) HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. *Front Neuroinform* 7:14.](http://paperpile.com/b/mC6vv9/HIOq8)

36. [Salthouse TA, Hedden T (2002) Interpreting reaction time measures in between-group comparisons. *J Clin Exp Neuropsychol* 24(7):858–872.](http://paperpile.com/b/mC6vv9/ZEwGX)

37. [Caruso JC (2004) A Comparison of the Reliabilities of Four Types of Difference Scores for Five Cognitive Assessment Batteries. *Eur J Psychol Assess* 20(3):166–171.](http://paperpile.com/b/mC6vv9/h6Inn)

38. [Hedge C, Powell G, Sumner P (2017) The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behav Res Methods*:1–21.](http://paperpile.com/b/mC6vv9/4DBuR)

39. [MacLeod JW, et al. (2010) Appraising the ANT: Psychometric and theoretical considerations of the Attention Network Test. *Neuropsychology* 24(5):637.](http://paperpile.com/b/mC6vv9/PMR3p)

40. [Guyatt G, Walter S, Norman G (1987) Measuring change over time: assessing the usefulness of evaluative instruments. *J Chronic Dis* 40(2):171–178.](http://paperpile.com/b/mC6vv9/IXrnc)

41. [Paap KR, Sawi O (2016) The role of test-retest reliability in measuring individual and group differences in executive functioning. *J Neurosci Methods* 274:81–93.](http://paperpile.com/b/mC6vv9/CKaZb)

42. [Spearman C (1904) The Proof and Measurement of Association between Two Things. *Am J Psychol* 15(1):72–101.](http://paperpile.com/b/mC6vv9/Iht5n)

1. In the remainder of this article we will use *r* to denote Pearson’s correlation, *⍴* to denote Spearman correlation, *ICC* to denote intraclass correlation and *𝜏* to denote Kendall’s correlation. [↑](#footnote-ref-1)
2. A detailed analysis of the DDM parameter estimates will be reported elsewhere. [↑](#footnote-ref-2)
3. For measures that were calculated using different numbers of trials for each subject due to timing out or other exclusions we took the mean number of trials used for the measure across all subjects. [↑](#footnote-ref-3)