Large-scale analysis of test–retest reliabilities of self-regulation measures

A. Zeynep Enkavi[a,1], Ian W. Eisenberg[a], Patrick G. Bissett[a], Gina L. Mazza[b], David P. MacKinnon[c], Lisa A. Marsch[d], and Russell A. Poldrack[a]

[a]Department of Psychology, Stanford University, Stanford, CA 94305; [b]Department of Health Sciences Research, Mayo Clinic, Scottsdale, AZ 85259; [c]Department of Psychology, Arizona State University, Tempe, AZ 85281; and [d]Geisel School of Medicine, Dartmouth College, Hanover, NH 03755

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved February 11, 2019 (received for review October 25, 2018)

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 (*i*) providing a comprehensive literature review on an extensive set of self-regulation measures and (*ii*) empirically evaluating test–retest reliability of this battery in a new sample. We find that dependent variables (DVs) from self-report surveys of self-regulation have high test–retest reliability, while DVs derived from behavioral tasks do not. This holds both in the literature and in our sample, although the test–retest reliability estimates in the literature are highly variable. We confirm that this is due to differences in between-subject variability. We also compare different types of task DVs (e.g., model parameters vs. raw response times) in their suitability as individual difference DVs, finding that certain model parameters are as stable as raw DVs. Our results provide greater psychometric footing for the study of self-regulation and provide guidance for future studies of individual differences in this domain.

self-regulation | retest reliability | individual differences

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), problem gambling (3–6), and overeating (7–9). 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), although its role as a moderator of behavior change has recently been challenged (12). 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 psychological mechanisms underlying self-regulatory functions. For example, behavioral tasks involving speeded choice responses are commonly used to compare conditions and isolate component processes. Within cognitive psychology and neuroscience, there has been particular interest in isolating mechanisms involved in "cognitive control" (13, 14). Candidate mechanisms include the ability to interrupt or preempt a particular behavior (response inhibition), the ability to rapidly switch between behavioral or task sets (set shifting or switching), and the ability to resist interference from irrelevant information (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 (delay discounting), which is thought to relate to numerous real-world outcomes (2, 15–17). Given that behavioral 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 (18–21).

One complicating factor in assessing the relation between behavioral task performance and self-report measures is potentially differing psychometric properties. Particularly, while the assessment of test–retest reliability (hereafter simply referred to as reliability) is a common aspect of survey development, 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% (22–24). Therefore, it is difficult to determine whether the weak relationship between different measures of self-regulation results from flawed theories or flawed operationalizations.

Here we report a large-scale examination of 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 to model the ontological structure of self-regulation (see refs. 20 and 25). 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

Significance

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 test–retest reliability of many of these measures is unclear. This paper reviews the literature on the test–retest reliability of self-regulation measures and characterizes long-term test–retest reliability in a large sample of individuals completing an extensive battery. The results show that while self-report measures have generally high test–retest reliability, behavioral task measures have substantially lower test–retest reliability, raising questions about their ability to serve as trait-like measures of individual differences.

These data were previously presented as a poster at the Society for Neuroeconomics Annual Conference, October 6–8 2017, Toronto, Canada, and the Society for Judgment and Decision Making Annual Conference, November 10–13, 2017, Vancouver, Canada.

Author contributions: D.P.M., L.A.M., and R.A.P. designed research; A.Z.E. and I.W.E. performed research; A.Z.E. analyzed data; and A.Z.E., I.W.E., P.G.B., G.L.M., D.P.M., L.A.M., and R.A.P. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The data used for this study have been deposited in GitHub, https://github.com/IanEisenberg/Self_Regulation_Ontology/tree/master/Data.

[1]To whom correspondence should be addressed. Email: zenkavi@stanford.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1818430116/-/DCSupplemental.

Published online March 6, 2019.

assess the relative reliability of data acquired online compared with laboratory samples. Although previous work suggested that data acquired online can exhibit high reliability (26–30), it did not encompass the breadth of measures relevant to self-regulation collected here. Additionally, the use of a relatively long retest delay (2–4 mo) placed the work on the timescale of many behavioral change studies, providing information on the stability of pre-/post intervention comparisons of self-regulatory function. Moreover, using the raw data allowed us to characterize the causes of systematic differences between measure types by isolating the sources of variance.

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 metrics 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 trade-offs), and nondecision time (perceptual and response execution process).

Another dimension of interest for the behavioral task dependent variables (DVs) was whether contrast DVs (subtraction of one condition from another) intended to isolate putative cognitive processes are suitable as trait DVs. This subtraction logic is a common strategy when using behavioral tasks for both raw DVs and model parameters. However, subtraction of random variables mathematically implies an increase in the contrast DVs' variance and therefore lower 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 self-regulation DV stability.

Results

Analysis of Prior Literature. Our literature review contained 171 DVs, 154 papers, 17,550 participants, and 583 data points on reliability (Fig. 1). Studies reporting reliability for surveys had,

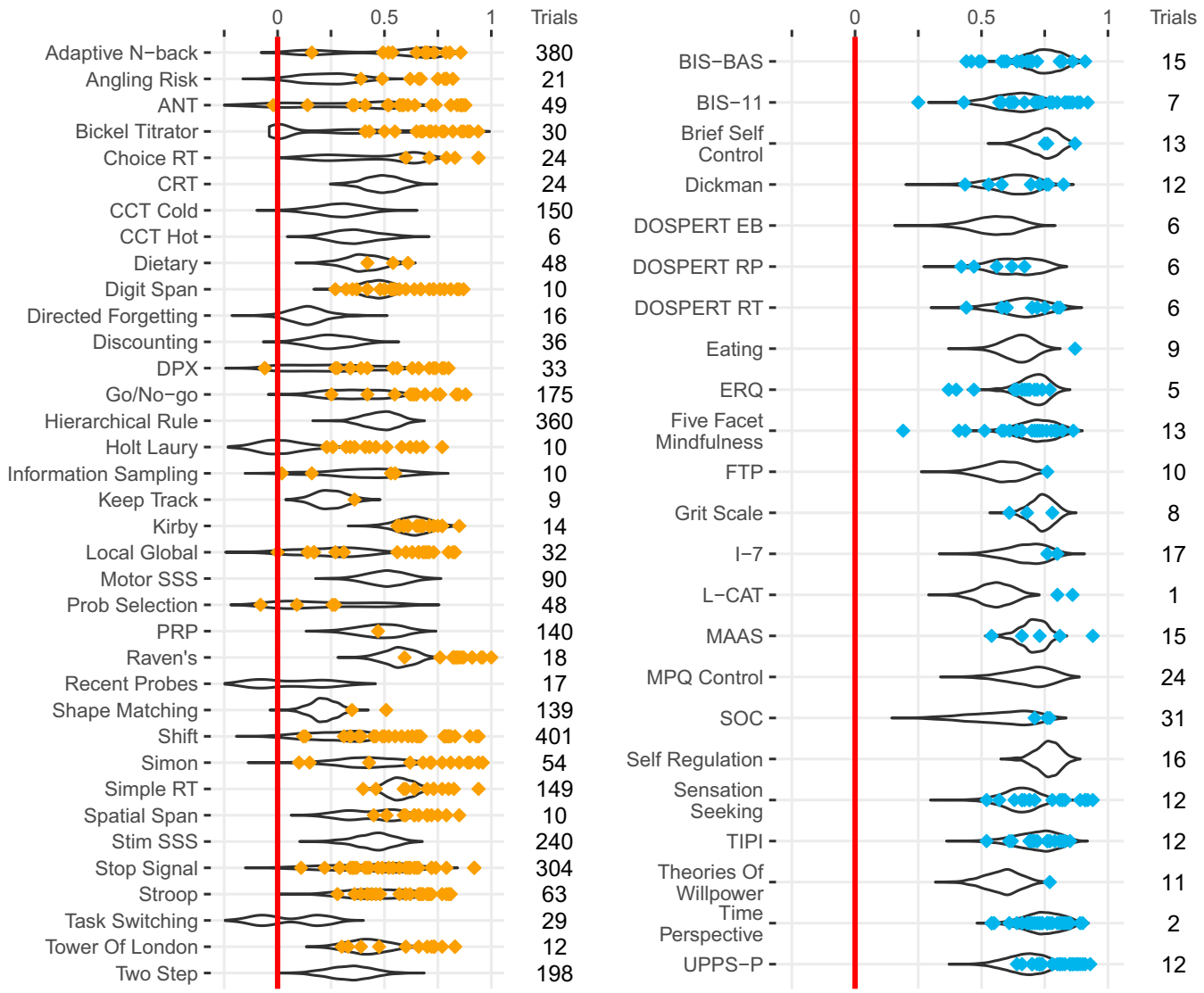

Fig. 1. Summary of the literature review and our new dataset for tasks (*Left*) and surveys (*Right*). Each point represents a study containing reliability data on an unspecified DV for a given task. Violin plots show bootstrapped reliability estimates for tasks (*Left*) and surveys (*Right*). We sampled 150 subjects with replacement 1,000 times to create a distribution of reliability estimates for each DV. DV reliability distributions are overlaid for each task, as shown in *SI Appendix*, Fig. S5. Vertical lines are zero reliability. Columns to the *Right* show mean number of trials for DVs in that task. See *SI Appendix* for abbreviated task names.

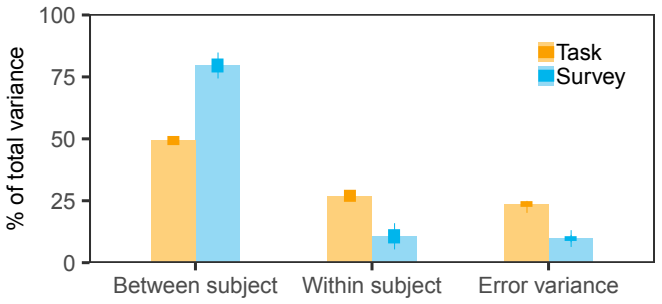

Fig. 2. Percentage of variance explained by the three sources of variance: between subjects, within subjects, and error variance for bootstrapped samples. Error bars are 95% CI.

on average, 50 more subjects than those reporting reliability for tasks (95% credible interval (CI) = [29, 70]). Controlling for sample size and retest delay, task DVs' reliability estimates were, on average, 0.139 lower compared with survey DVs' (95% CI = [−0.192, −0.084]; mean reliability for task DVs in the literature = 0.610 and for survey DVs = 0.716). Reliability decreased by 0.0001 for every additional participant in a study (95% CI for decrease = [−0.0002, −0.00001]). This negative relationship between sample size and DV reliability may reflect publication bias and/or variation in undocumented decisions taken by researchers, as discussed further later.

Analysis of New Dataset.

Data quality checks. To ensure data quality we conducted three tests (detailed further in *SI Appendix*): 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 these analyses raised concerns, and overall, they provided some assurance that the participants were real people and not automated machines [which is a concern since participants were recruited using Amazon Mechanical Turk and Experiment Factory (31)].

Survey and behavioral task reliability in new data. We calculated 372 DVs for behavioral tasks and 74 for surveys. Reliability for each DV was estimated using a nonparametric bootstrap (1,000 samples); statistics on these bootstrapped estimates are reported instead of point estimates. We report intraclass correlations [ICC (2,1)] as the main metric of reliability based on its ability to account for various sources of variance separately (*SI Appendix*, Table S1). The ICC, which ranges from −1 to 1, is a preferred metric for reliability and is not biased by sample size (32). While the ICC can have negative values, these are difficult to interpret as a proportion of variances. There were 21 variables that had negative point estimates of ICC. We repeated our analyses both replacing these negative values with zeros and removing these variables. None of our results change with either of these cleaning procedures. None of our conclusions change using other reliability metrics either. The correlation between different reliability metrics ranged from 0.932 to 0.998 (*SI Appendix*, Fig. S3).

Mirroring the literature, the average reliability of behavioral task DVs was 0.432 lower than the average reliability of survey DVs (95% CI for difference = [−0.482, −0.384]). While survey DVs had a median ICC of 0.674 (0.425 for the first quartile, 0.836 for the third quartile), behavioral task DVs had a median ICC of 0.311 (−0.091 for the first quartile, 0.665 for the third quartile).

A quantitative explanation for the difference in reliability estimates between surveys and tasks, as recently detailed by Hedge et al. (33), lies in the difference in sources of variance between these DVs. ICC is the ratio of between-subject variance versus total variance. Intuitively, DVs with high between-subject variance are better suited for individual difference analyses as they are more sensitive to the differences between the subjects in a sample. Conversely, as Hedge et al. noted, behavioral tasks are generally selected on the basis of reliable group effects, which select for DVs with low between-subject variance.

We find that 79.50% of survey DVs' variance is due to between-subject variability versus 49.30% of behavioral task DVs' (difference 95% CI = [26.10, 34.90]; Fig. 2). Conversely, 26.98% of behavioral tasks' variance is explained by within-subject variance compared with 10.8% of survey DVs' (systematic differences between sessions; difference 95% CI = [10.68, 21.84]). Task DVs also have higher percentages of residual variance (difference 95% CI = [11.46, 16.57]).

Comparison of literature and new data. To compare our findings to the literature, we sampled the same number of estimates from our bootstrapped results as we found in the literature for each DV and calculated the correlation between the sampled empirical (i.e., from our data) reliabilities and those in the literature. Repeating this 100 times, the mean correlation (Fig. 3) between empirical and literature reliabilities was 0.247 for behavioral task DVs (range = 0.176–0.297) and 0.063 for survey DVs (range = −0.024–0.164).

While these correlations seem weak, they must be interpreted in the context of the variability of reliability estimates in the literature. If individual studies in the literature have similarly weak relationships to the literature-wide reliability for a given DV (i.e., if the variance of the literature reliabilities for a given DV is large), this suggests a general issue of variability in reliability estimates across samples rather than a specific issue with our sample. Therefore, we compared two types of models: (*i*) One that predicted the literature reliability using an estimate sampled from the literature review. (*ii*) Another that predicted the literature reliability using the estimate from our new data.

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 (Fig. 4). However, the mean decrease in variance explained using our data is only 4.69% (95% CI of difference = [3.89%, 5.40%]), suggesting that published estimates of reliability in this domain are quite noisy. Hence, estimating reliability using an online sample does not substantively change conclusions compared with in-laboratory samples. (Although we did not limit our literature review to in-lab samples, all the papers we found that reported reliability were such.)

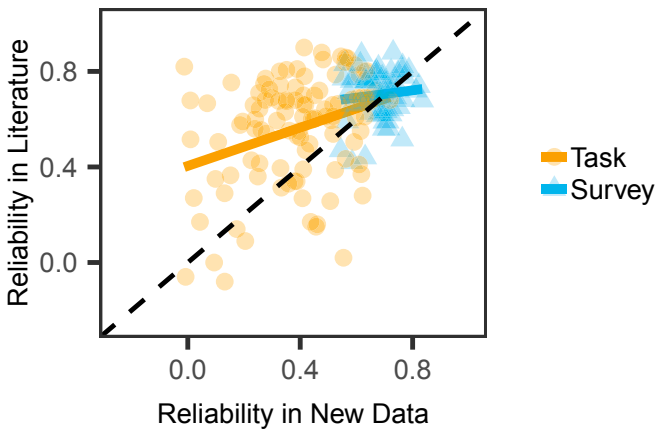

Fig. 3. Correlation between mean reliability estimates for each DV found in the literature and the mean reliability from our data.

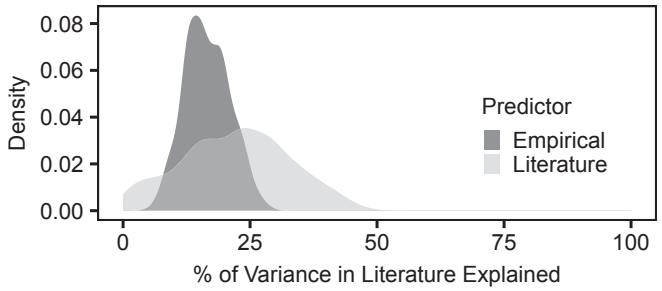

Fig. 4. Comparison of literature reliability predictability using literature vs. empirical reliabilities. Literature reliabilities are predicted using either a single reliability from the literature or the mean reliability from our new data as a predictor accounting for the sample size and measure modality.

Effect of task length on stability. To compare potential effects of task-specific attributes on reliability across tasks, we examined the relationship between the number of trials a task included and its reliability. Across non-DDM DVs, there was an insignificant 0.0002-point increase in reliability for each additional trial (95% CI = [−0.0002, 0.0006]). For DVs that were calculated using different numbers of trials for each subject due to time out or other exclusions we took the mean number of trials used for the DV across all subjects.

For tasks in which DVs are estimated using many trials, one can ask whether the same DV becomes less reliable if fewer trials are used to estimate its reliability, as this might suggest that low task reliability in our study is due to insufficient numbers of trials. The effect of task length on the stability of a DV is a largely open empirical question. Previous analyses (see SI Appendix of ref. 33) suggest both that some DVs require more trials than what is used in the literature for stable reliability estimates and also that the effects are highly DV dependent. We present only a brief exploration of this question; our data are openly available, so researchers can make more informed decisions when choosing the number of trials in other tasks.

We calculated DVs for six tasks of various lengths in our battery. Reliability increased by 0.119 when using 1/2 trials instead of 1/4 (95% CI = [0.082, 0.158]) and 0.040 when using 3/4 of trials compared with 1/2 (95% CI = [0.019, 0.063], *SI Appendix*, Fig. S6). These analyses can take into account differences in learning rates or practice effects between participants, but we defer from commenting further in this paper. However, there were nonnegligible differences between DVs. To identify patterns in these differences, we calculated a denser sample of reliability estimates for a single task with many trials. We found three patterns (*SI Appendix*, Fig. S7): (*i*) reaching acceptable reliability in many fewer trials than were used, (*ii*) increasing reliability with more trials and reaching acceptable levels at the end, and (*iii*) never reaching acceptable levels regardless of task length. Thus, a researcher might question whether to use a task for individual difference analyses, as many of the DVs that are usually of primary interest exhibit little or no reliability even after hundreds of trials. Alternatively, reliable results can be obtained with relatively few trials by using a more stable DV.

Comparison of task DV types. Data from any given behavioral task can be analyzed in various ways, yielding different types of DVs. We compare the reliability of raw DVs (response times and accuracies) to parameters of the DDM, a well-established model that addresses speed–accuracy trade-offs and offers interpretable latent variables (34, 35). We chose two approaches to parameter estimation: EZ-diffusion, so named to invoke the idea that is easier to estimate compared to its analytical competitors, (36) and hierarchical DDM (HDDM) (37). We compare raw DVs to DDM parameters in this paper as an example of a central approach in cognitive psychology: transforming performance DVs into interpretable metrics of putative constructs. However, these models are not equally appropriate for all of our tasks, nor do they fit equally well. Details of these model parameters and how they compare with raw performance DVs will be presented in further detail elsewhere.

The EZ-diffusion method is a set of closed-form expressions that transform mean response time (RT), variability of RT, and accuracy to drift rate, threshold, and nondecision time. The HDDM uses hierarchical Bayesian modeling to allow simultaneous estimation of both group and individual subject parameters. Both raw DVs and parameters can also be "contrast" and "noncontrast."

Cognitively interpretable parameter estimates are comparable in reliability to raw DVs of RT and accuracy [median ICC for noncontrast (contrast) raw DVs = 0.500 (0.174), for noncontrast EZ DVs = 0.471 (0.087), and for noncontrast HDDM DVs = 0.377 (0.232)]. (One can assume that DDM thresholds and nondecision times should not differ across conditions. This would imply that the contrasts of these parameters capture noise and therefore have low reliability. This assumption does not hold in our data. Threshold and nondecision time contrasts are systematically different from zero.) Reliability estimates of noncontrast DVs (Fig. 5) were, on average, 0.288 points higher than those of contrast DVs (95% CI = [0.249, 0.326]). This is not surprising given the summing of the variance in the difference score. Of concern is the fact that contrast DVs had low to no reliability (mean = 0.154, SD = 0.140) compared with the moderate to low reliability of the noncontrast DVs (mean = 0.442, SD = 0.152). This is particularly alarming given their common use in cognitive psychology as putative trait DVs of cognitive constructs and predictors of real-world outcomes.

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 reliability (95% CI = [−0.001, 0.004]), although, as with tasks, surveys could also be analyzed in more detail using item response theory or other models.

Reliability of latent variables. Although most individual DVs from tasks are not appropriate for individual difference analyses based

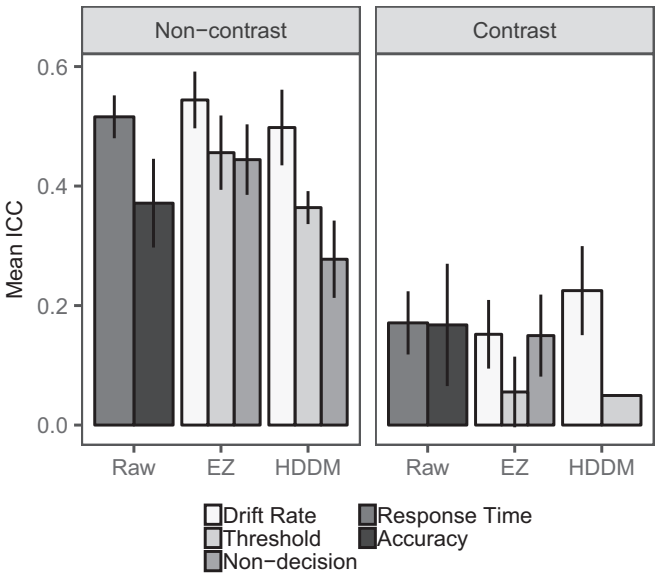

Fig. 5. Average reliability estimates comparing raw DVs and model parameters as well as contrast and noncontrast DVs for task DVs. Error bars indicate 95% CI. The rightmost bar depicts a single DV of difference between thresholds allowed to vary for conditions in a task.

on their low reliability, this does not preclude other ways of using them as trait DVs. One can use a data-driven approach to integrate them and extract scores that may be more stable. An example using the same dataset reported here is detailed in Eisenberg et al. (20): Factor scores computed at both time points using the same linear combination of DVs correlated highly with each other for five task factors ($M = 0.82$, min = 0.76, max = 0.85) and 12 survey factors ($M = 0.86$, min = 0.75, max = 0.95). However, 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 $R^2 = 0.10$), whereas tasks did not, neither as factors nor as separate DVs (average $R^2 = 0.01$).

Discussion

This report provides a systematic characterization of the reliability of self-report and behavioral task DVs of the construct of self-regulation. There is a broad set of theoretical approaches to the construct of self-regulation spanning different areas of psychological science, from social and personality psychology (38–40) to cognitive neuroscience (14, 41). We explicitly selected DVs that span the space of theories of self-regulation in psychology as broadly as possible to be relatively agnostic, particularly in light of evidence casting doubts on these previous conceptualizations (20, 38–40, 42, 43).

Findings from the Literature on Reliability of Self-Regulation Measures. We found that while psychometric studies of survey DVs have larger sample sizes than task DVs in the literature, reliability decreased with sample size. This might suggest that smaller studies afford researchers more control over their measurement, leading to higher reliability. On the other hand, larger sample sizes might be more reflective of the truly lower reliability of measures; Hopkins (32) suggests that studies of reliability with samples smaller than 50 should be treated as pilot studies for this reason. Studies with smaller samples are more prone to variable reliabilities. Coupled with publication bias, this may inflate the results in the literature. In our literature review 55.4% of the studies on tasks and 34.8% of the studies on surveys have sample sizes below 50. We had a larger sample size and found relatively low reliability of behavioral tasks, consistent with the literature.

We contextualized results from our battery of self-regulation measures with an extensive literature review, quantifying the variability of the literature's reliability estimates. This provided a "noise ceiling" for reliability studies, a reference point for the expected relationship between any two sets of reliability estimates. Because the literature reliability estimates lacked strong coherence for many DVs, their low correlations with our reliabilities led to a less than 5% decrease in the predictability of prior literature. Hence, the results reported for the present dataset are similar to what is expected from the literature.

Systematic Differences in the Reliability of Self-Regulation DVs. The literature and our data show that self-regulation DVs based on self-report surveys have higher reliability than behavioral task DVs due to higher between-subject variance of survey DVs. Thus, survey DVs are more appropriate for individual difference analyses. Whether this divergence of psychometric properties of self-regulation by measurement modality generalizes to other psychological constructs (e.g., working memory) and whether it reflects related cognitive processes from different timescales (e.g., state vs. trait) are important empirical questions for future study.

Exploratory analyses on task DVs suggested that while additional trials often lead to more stable DVs, task length has varying effects depending on the DV even within a task. On another note, the reliability of DDM parameters did not significantly differ from the reliability of raw DVs like response times and accuracy. Researchers may therefore prefer DDMs given their interpretability.

Revisiting a longstanding question on the reliability of contrast scores, we confirm that they are less reliable than their components. DVs of differences between conditions have lower reliabilities due to correlations between the two DVs used in calculating the difference score (44, 45) and the increase in the variance through subtraction. The concerning point is that behavioral task DVs of greatest interest in the self-regulation literature are contrast DVs, as they offer mechanistic insights psychologists seek, which have low to no reliability.

Implications of Low Reliability for Behavioral Task DVs. Although the unsuitability of task DVs for individual difference analyses of self-regulation might be disappointing, especially in the face of work showing correlations between these DVs and problematic real-world behaviors, it should not be surprising. As Hedge et al. (33) 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. This necessitates low reliability. For example, one might repeatedly find a significant Stroop effect (the difference in the response times between the congruent and incongruent conditions) in samples measured multiple times, even while the relative distributions 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 reliability. This does not invalidate the existence of the Stroop effect but does undermine its suitability as a trait DV. Detailed analyses of sources of variance provide researchers with a priori hypotheses on which DVs to expect significant changes in different experimental designs.

Despite psychometric shortcomings, task DVs can be integrated using data-driven approaches to extract more stable latent variables that are potentially more suitable for trait-like treatment. With this approach, we found more stable latent variables, although they were not more predictive of real-world behaviors (20). Notably, these latent variables included not only tasks that commonly appeared in theoretical frameworks but also tasks that are seldom considered within the self-regulation literature yet yield some of the more reliable task DVs (e.g., simple reaction time, hierarchical rule, and digit span).

On the other hand, different psychometric properties of DVs serve different purposes. For example, while high reliability is desirable for DVs that will be used in trait-like characteristic analyses, it is neither a necessary nor a sufficient condition for the responsiveness of a DV to capture change over time (46, 47). Although we provide practical guidelines for researchers interested in these DVs, we do not answer how these DVs relate to the construct of "self-regulation." While the reliability of a DV has consequences on the limits of its correlation with other DVs, specifically, for any two variables the correlation between them must be smaller than the square root of the reliability of each DV (44, 48, 49), the question of validity remains a separate one addressed in related work (20).

Conclusions

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

Materials and Methods

Sample. Participants were a subset from a larger study (25) conducted on Amazon Mechanical Turk. 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 *SI Appendix*, Table S2. The sample size was specified before 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, which addressed preferentially sampling the most motivated subjects who may systematically differ from the full sample. Each batch had a week to complete the battery. Retest data collection took place between November 2016 and March 2017. The mean number of days between the waves was 111 (median = 115; range = 60–228). Of the 242 participants invited, 175 participants started the battery, and 157 completed the battery. The 85 noncompleters were compared with the completers in their time 1 data. None of the DVs differed significantly between the groups (correcting for multiple comparisons), mitigating concerns of selection effects. The study was approved by the Stanford Institutional Review Board (IRB-34926). All participants clicked to confirm their agreement with an informed consent form before beginning the battery.

The details of the data collection platform, data analysis pipeline, including links to analysis scripts and interactive visualizations, descriptions of all measures, and the literature review steps are in the *SI Appendix*.

Data Availability. All versions of data and their deposition dates are available at: https://github.com/IanEisenberg/Self_Regulation_Ontology/tree/master/Data as described in ref. 20. The code for dependent measure calculation can be found at: https://github.com/IanEisenberg/expfactory-analysis/tree/master/expanalysis/experiments. All analysis code and data are available at: https://zenkavi.github.io/SRO_Retest_Analyses/output/reports/SRO_Retest_Analyses.nb.html.

ACKNOWLEDGMENTS. This work was supported by the NIH Science of Behavior Change Common Fund Program through an award administered by the National Institute for Drug Abuse (NIDA) (Grant UH2DA041713 to L.A.M. and R.A.P.). Additional support was provided by NIDA Grant P30DA029926.

1. Prochaska JO, DiClemente CC, Norcross JC (1992) In search of how people change. Applications to addictive behaviors. *Am Psychol* 47:1102–1114.
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* 128:78–87.
3. Alessi SM, Petry NM (2003) Pathological gambling severity is associated with impulsivity in a delay discounting procedure. *Behav Processes* 64:345–354.
4. Kertzman S, et al. (2008) Go-no-go performance in pathological gamblers. *Psychiatry Res* 161:1–10.
5. Lawrence AJ, Luty J, Bogdan NA, Sahakian BJ, Clark L (2009) Impulsivity and response inhibition in alcohol dependence and problem gambling. *Psychopharmacology (Berl)* 207:163–172.
6. Fuentes D, Tavares H, Artes R, Gorenstein C (2006) Self-reported and neuropsychological measures of impulsivity in pathological gambling. *J Int Neuropsychol Soc* 12:907–912.
7. Nederkoorn C, Smulders FTY, Havermans RC, Roefs A, Jansen A (2006) Impulsivity in obese women. *Appetite* 47:253–256.
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* 7:315–322.
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* 51:399–409.
10. Rozensky RH, Bellack AS (1974) Behavior change and individual differences in self-control. *Behav Res Ther* 12:267–268.
11. Bickel WK, Vuchinich RE (2000) *Reframing Health Behavior Change with Behavioral Economics* (Psychology Press, New York).
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* 12:157–178.
13. Miller EK, Cohen JD (2001) An integrative theory of prefrontal cortex function. *Annu Rev Neurosci* 24:167–202.
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. *Cognit Psychol* 41:49–100.
15. Mischel W, Shoda Y, Rodriguez MI (1989) Delay of gratification in children. *Science* 244:933–938.
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* 112:382–392.
17. Meier S, Sprenger CD (2012) Time discounting predicts creditworthiness. *Psychol Sci* 23:56–58.
18. Duckworth AL, Kern ML (2011) A meta-analysis of the convergent validity of self-control measures. *J Res Pers* 45:259–268.
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.
20. Eisenberg IW, et al. (December 12, 2018) Uncovering mental structure through data-driven ontology discovery, 10.31234/osf.io/fvqej.
21. Cyders MA, Coskunpinar A (2011) Measurement of constructs using self-report and behavioral lab tasks: Is there overlap in nomothetic span and construct representation for impulsivity? *Clin Psychol Rev* 31:965–982.
22. Cicchetti DV, Sparrow SA (1981) Developing criteria for establishing interrater reliability of specific items: Applications to assessment of adaptive behavior. *Am J Ment Defic* 86:127–137.
23. Fleiss JL, Levin B, Paik MC (2013) *Statistical Methods for Rates and Proportions* (Wiley, Hoboken, NJ).
24. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. *Biometrics* 33:159–174.
25. Eisenberg IW, et al. (2018) Applying novel technologies and methods to inform the ontology of self-regulation. *Behav Res Ther* 101:46–57.
26. Paolacci G, Chandler J, Ipeirotis PG (2010) Running experiments on Amazon Mechanical Turk. *Judgm Decision Mak* 5:411–419.
27. Horton JJ, Rand DG, Zeckhauser RJ (2011) The online laboratory: Conducting experiments in a real labor market. *Exp Econ* 14:399–425.
28. Buhrmester M, Kwang T, Gosling SD (2011) Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspect Psychol Sci* 6:3–5.
29. Behrend TS, Sharek DJ, Meade AW, Wiebe EN (2011) The viability of crowdsourcing for survey research. *Behav Res Methods* 43:800–813.
30. Crump MJC, McDonnell JV, Gureckis TM (2013) Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS One* 8:e57410.
31. Sochat VV, et al. (2016) The experiment factory: Standardizing behavioral experiments. *Front Psychol* 7:610.
32. Hopkins WG (2000) Measures of reliability in sports medicine and science. *Sports Med* 30:1–15.
33. Hedge C, Powell G, Sumner P (2018) The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behav Res Methods* 50:1166–1186.
34. Ratcliff R (1978) A theory of memory retrieval. *Psychol Rev* 85:59–108.
35. Ratcliff R, Smith PL, Brown SD, McKoon G (2016) Diffusion decision model: Current issues and history. *Trends Cogn Sci* 20:260–281.
36. Wagenmakers E-J, van der Maas HLJ, Grasman RPPP (2007) An EZ-diffusion model for response time and accuracy. *Psychon Bull Rev* 14:3–22.
37. Wiecki TV, Sofer I, Frank MJ (2013) HDDM: Hierarchical Bayesian estimation of the drift-diffusion model in Python. *Front Neuroinform* 7:14.
38. Stahl C, et al. (2014) Behavioral components of impulsivity. *J Exp Psychol Gen* 143:850–886.
39. Rey-Mermet A, Gade M, Oberauer K (2018) Should we stop thinking about inhibition? Searching for individual and age differences in inhibition ability. *J Exp Psychol Learn Mem Cogn* 44:501–526.
40. Saunders B, Milyavskaya M, Etz A, Randles D, Inzlicht M (2018) Reported self-control is not meaningfully associated with inhibition-related executive function: A Bayesian analysis. *Collabra Psychol* 4:39.
41. Hare TA, Camerer CF, Rangel A (2009) Self-control in decision-making involves modulation of the vmPFC valuation system. *Science* 324:646–648.
42. Karr JE, et al. (2018) The unity and diversity of executive functions: A systematic review and re-analysis of latent variable studies. *Psychol Bull* 144:1147–1185.
43. Sharma L, Markon KE, Clark LA (2014) Toward a theory of distinct types of "impulsive" behaviors: A meta-analysis of self-report and behavioral measures. *Psychol Bull* 140:374–408.
44. Salthouse TA, Hedden T (2002) Interpreting reaction time measures in between-group comparisons. *J Clin Exp Neuropsychol* 24:858–872.
45. Caruso JC (2004) A comparison of the reliabilities of four types of difference scores for five cognitive assessment batteries. *Eur J Psychol Assess* 20:166–171.
46. Guyatt G, Walter S, Norman G (1987) Measuring change over time: Assessing the usefulness of evaluative instruments. *J Chronic Dis* 40:171–178.
47. Tiplady B (1992) Continuous attention: Rationale and discriminant validation of a test designed for use in psychopharmacology. *Behav Res Methods Instrum Comput* 24:16–21.
48. 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.
49. Spearman C (1904) The proof and measurement of association between two things. *Am J Psychol* 15:72–101.