Learning from the Reliability Paradox: How Theoretically Informed Generative Models Can Advance the Social, Behavioral, and Brain Sciences

Nathaniel Haines*
The Ohio State University, Department of Psychology,
1835 Neil Ave., Columbus, OH., 43210
Email: haines.175@osu.edu; Website: http://haines-lab.com/

Peter D. Kvam
University of Florida, Department of Psychology,
945 Center Dr, Gainesville, FL, 32611
Email: pkvam@ufl.edu; Website: https://peterkvam.com/

Louis Irving
University of Florida, Department of Psychology,
945 Center Dr, Gainesville, FL, 32611
Email: louis.irving@ufl.edu; Website: https://theapclab.wordpress.com/people/

Colin Tucker Smith
University of Florida, Department of Psychology,
945 Center Dr, Gainesville, FL, 32611
Email: colinsmith@ufl.edu; Website: https://theapclab.wordpress.com/

Theodore P. Beauchaine
The Ohio State University, Department of Psychology,
1835 Neil Ave., Columbus, OH., 43210
Email: beauchaine.1@osu.edu; Website: https://tpb.psy.ohio-state.edu/LAP/people.html

Mark A. Pitt
The Ohio State University, Department of Psychology,
1835 Neil Ave., Columbus, OH., 43210
Email: pitt.2@osu.edu; Website: https://u.osu.edu/markpitt/

Woo-Young Ahn
Seoul National University, Department of Psychology,
1 Gwanak-ro, Gwanak-gu, Seoul, South Korea
Email: wahn55@snu.ac.kr; Website: https://ccs-lab.github.io/

Brandon M. Turner*
The Ohio State University, Department of Psychology,
1835 Neil Ave., Columbus, OH., 43210
Email: turner.826@gmail.com; Website: https://turner-mbcn.com/

*Co-corresponding authors

Word counts: Abstract 249; main text 13,814; references 3,334; entire text: 17,397.

Abstract: Short

The reliability paradox implies that popular statistical modeling tools (general linear model) are not well-suited for advancing theories of individual differences. Such tools only provide superficial summary descriptions of observed data, and by extension they are atheoretical with respect to the psychological mechanisms that generate behavior. Further, they lack the flexibility needed to develop and test increasingly complex theories of behavior. We argue that generative modeling fills this theory-description gap, and demonstrate its superiority in a reanalysis of data. Generative models produce higher test-retest reliability and more theoretically informative parameter estimates than do traditional methods.

Abstract: Long

Behavioral tasks (e.g., Stroop task) that produce replicable group-level effects (e.g., Stroop effect) often fail to reliably capture individual differences between participants (e.g., low test-retest reliability). This "reliability paradox" has led many researchers to conclude that most behavioral tasks cannot be used to develop and advance theories of individual differences. However, these conclusions are derived from statistical models that provide only superficial summary descriptions of behavioral data, thereby ignoring theoretically-relevant data-generating mechanisms that underly individual-level behavior. More generally, such descriptive methods lack the flexibility to test and develop increasingly complex theories of individual differences. To resolve this theory-description gap, we present *generative modeling* approaches, which involve using background knowledge to specify how behavior is generated at the individual level, and in turn how the distributions of individual-level mechanisms are characterized at the group level—all in a single joint model. Generative modeling shifts our focus away from estimating descriptive statistical "effects" toward estimating psychologically meaningful parameters, while simultaneously accounting for measurement error that would otherwise attenuate individual difference correlations. Using simulations and empirical data from the Implicit Association Test and Stroop, Flanker, Posner Cueing, and Delay Discounting tasks, we demonstrate how generative models yield (1) higher test-retest reliability estimates, and (2) more theoretically informative parameter estimates relative to traditional statistical approaches. Our results reclaim optimism regarding the utility of behavioral paradigms for testing and advancing theories of individual differences, and emphasize the importance of formally specifying and checking model assumptions to reduce theory-description gaps and facilitate principled theory development.

Keywords: Bayesian analysis, implicit attitudes, impulsivity, individual differences, generative modeling, measurement error, reliability, self-control, theory development

1. Introduction

A primary aim of social, behavioral, and brain sciences is to develop *explanations* that answer questions of *why* or *how* observed psychological phenomena occur (Hempel & Oppenheim, 1948). *Explanatory theories* are indispensable for making valid causal inferences and for determining how to successfully intervene on psychological processes. However, developing useful yet accurate explanations of complex psychological processes is a serious challenge. Explanation requires (a) a theory encoding core causal assumptions about the phenomenon of interest, (b) experimental tasks or data sources that capture the key theoretical phenomenon, and often (c) statistical models that test theoretical principles against observed data while accounting for uncertainty (Guest & Martin, 2020; Kellen, 2019; Suppes, 1966). In the social, behavioral, and brain sciences, theories may be specified verbally, conceptually, or (less commonly) mathematically. Typically, inference then proceeds using some combination of summary statistics (e.g., summed items on a questionnaire, average response times, etc.) and a descriptive statistical model as a medium for performing null hypothesis significance testing (Tong, 2019). For example, researchers might apply a *t*-test or multiple regression to summary statistics (e.g., means), yielding a *p*-value that is subsequently interpreted with respect to the substantive theory.

Despite their popularity, summary statistics followed by significance tests may misalign with the objectives of explanatory theories. Specifically, when we default to summary statistics when analyzing behavioral data, we are assuming that summary statistics adequately capture the underlying data-generating mental process of interest. As we will demonstrate, this use of summary statistics both restricts theory development and produces suboptimal measurement precision, thereby constraining the explanatory power of our theories and the implementation of

corresponding models. In other words, the typical approach to explaining psychological phenomena using experimental design and hypothesis testing leads to a theory-description gap: researchers invest in verbal or conceptual theories to account for new statistical effects (i.e., data), but historically invest less in amending their statistical models in ways that best embody their theories (e.g., Beauchaine & Hinshaw, 2020; Michell, 2008; Szollosi & Donkin, 2019). For example, a typical approach to theory development in the behavioral sciences is to assume a verbal theory, make directional predictions based on that theory, design an experiment that can produce the predicted effect, and then use a standard statistical model (e.g., *t*-test, multiple regression) and corresponding statistical test to determine if in-sample directional effects can plausibly be attributed to the population of interest. The theory-description gap then arises as we continue to refine our verbal theories to the point at which they are no longer amenable to simple experimental designs, summary statistics, and standard inferential modeling[1]. Because theories evolve in the presence of new data, we argue that statistical models should do the same, thereby providing the needed quantitative precision and hypothesized explanation to fill the gap. Fortunately, such theory-description gaps can be addressed by explicating assumptions of both the descriptive and theoretical models, and by iteratively refining them in a mutually constraining fashion (Guest & Martin, 2020; Kellen, 2019; Suppes, 1966).

Iterative approaches to theory development and testing emphasize a shift away from pure empiricism and deductive inference and toward more principled abduction of explanatory models, along with subsequent model comparison and refinement (Navarro, 2018; van Rooij & Baggio, 2020). The key tenet of iterative, abductive inference is that we, as scientists, approach

[1] We provide examples of this phenomenon in section 3.2.

research questions with substantial background knowledge. Even in the absence of empirical data, we can use our background knowledge to instantiate potential explanatory mechanisms within competing statistical models, thereby producing *explanatory models*. In this way, background knowledge imposes considerable constraint on statistical inference that is not afforded by traditional summary statistic approaches. The role of empirical data is then secondary—we use data and experiments to arbitrate among or refine competing explanatory models. For example, if a theory predicts that a task manipulation should cause an increase in mean response times, there are presumably multiple cognitive mechanisms that could cause such an increase. Without building these mechanisms into the statistical model, it is unclear how the resulting statistical model estimates relate back to the theory (i.e., a theory-description gap) and, consequently, there is risk of misinterpreting even a well-fitting statistical model (see also Roberts & Pashler, 2000). By not filling gaps, theories remain abstract and vague, divorced from details in the data that require explanation, slowing advancement in the field.

In this article, we introduce *generative modeling* as a general solution to the theory-description gap. As a motivating example, we focus our attention on a current and vexing theory-description gap problem: the reliability paradox (Hedge et al., 2017). The reliability paradox, as we discuss in section 2, implies that many behavioral paradigms that are otherwise robust at the group-level (e.g., those that produce highly replicable condition- or group-wise differences) are unsuited for testing and building theories of individual differences due to low test-retest reliability. As we will demonstrate, these conclusions are derived from statistical models that do not include our background knowledge of behavior that, once built into a generative model, resolves issues of

low reliability and reveals the rich individual differences in behavior that can be used to advance explanatory theories.

The remainder of this article is organized as follows. First, we discuss the reliability paradox and its traditional interpretation as a motivating example. Second, we outline atheoretical assumptions researchers often make when analyzing behavioral data (i.e., the theory-description gap). These assumptions can lead to inappropriate conclusions about individual differences and the apparent reliability paradox. Third, as an alternative, we introduce generative modeling—a theoretically driven approach to statistical modeling that involves making explicit assumptions about data-generating mechanisms as a tool for filling the gap. We then re-analyze data collected from several common tasks used in psychology, neuroscience, and behavioral economics to show how generative modeling can improve precision of individual-level inferences from task data, thereby filling the theoretical gap. Finally, we discuss implications of our findings and provide actionable steps on how researchers can use generative modeling to advance the study of human behavior.

2. The Reliability Paradox

Paradigms such as the Implicit Association Test (IAT: Greenwald et al., 1998) and the Stroop (1935), Flanker (Eriksen & Eriksen, 1974), Posner Cueing (Posner, 1980), and Delayed Discounting Tasks (Green & Myerson, 2004; Mazur, 1987) consistently produce robust group effects using simple behavioral summary statistics and traditional statistical tests. For example, the Stroop effect is traditionally quantified as the mean difference in response times between incongruent (“red” colored blue) and congruent (“red” colored red) word-color pairs. Longer

response times on incongruent trials are interpreted as psychological interference resulting from competition between stimulus features. Since 1935, the basic Stroop effect has been replicated countless times (MacLeod, 1991), and is among the most well-known and easy to reproduce effects in behavioral science. Indeed, it appears that "everybody Stroops" (Haaf & Rouder, 2017).

Despite its replicability, Hedge et al. (2017) concluded that the Stroop effect is unreliable because of its low test-retest correlations within participants. In two separate studies, they found three-week test-retest intraclass correlation coefficients (ICCs) of .60 and .66. Similarly, ICCs for the Flanker and Posner Cueing tasks ranged from .4 to .7. Such findings have since been replicated and extended to wide range of self-control tasks (Enkavi et al., 2019). Other behavioral tasks that are used widely throughout the behavioral sciences, including the IAT, also show similarly low test-retest correlations (r = .01-.72; average of $r \approx .4$) across versions and timepoints (Gawronski et al., 2016; Klein, 2020). In the brain sciences, similarly low intraclass correlation coefficients were found in a meta-analysis of 90 experiments (mean ICC=0.397), and poor reliability of activity in regions of interest of brain regions across 11 common tasks used within the Human Connectome Project and the Dunedin Study (ICCs=0.067-0.485; Elliott et al., 2020). Unfortunately, such low test-retest reliability is not limited to task-based fMRI measures—both resting state and functional connectivity measures show comparably low reliability (e.g., Chen et al., 2015; Noble et al., 2019).

Low test-retest reliability across brain and behavioral tasks has led to considerable, justified concerns about using these tasks to test and develop theories of individual differences (Dang et

al., 2020; Elliott et al, 2020; Schimmack, 2019; Wennerhold & Friese, 2020). When there are concerns about reliability, sample sizes required to overcome measurement error must increase to compensate. For example, the sample size needed to detect a true medium effect (r = .3, with 80% power at α = .05) between two measures with perfect test-retest reliabilities is 82, whereas two measures with test-retest reliabilities of .6 requires a sample size of 239 (Hedge et al., 2017). The implication for studies relating behavioral measures to BOLD responses in fMRI studies is quite sobering. In an optimistic setting where brain and behavioral measures have test-retest reliabilities of .6, assuming that 1 hour of fMRI scanning is $500 (USD), a study powered to detect a true effect of r = .3 costs $500 $\times$ 239 $\approx$ $120,000 for data collection alone. Otherwise, if low test-retest reliabilities are combined with small samples sizes, effects inferred from null hypothesis significance tests (when many tests are conducted) are often spurious and may be of much greater magnitude or in the wrong direction compared to true underlying effects (Gelman & Carlin, 2014).

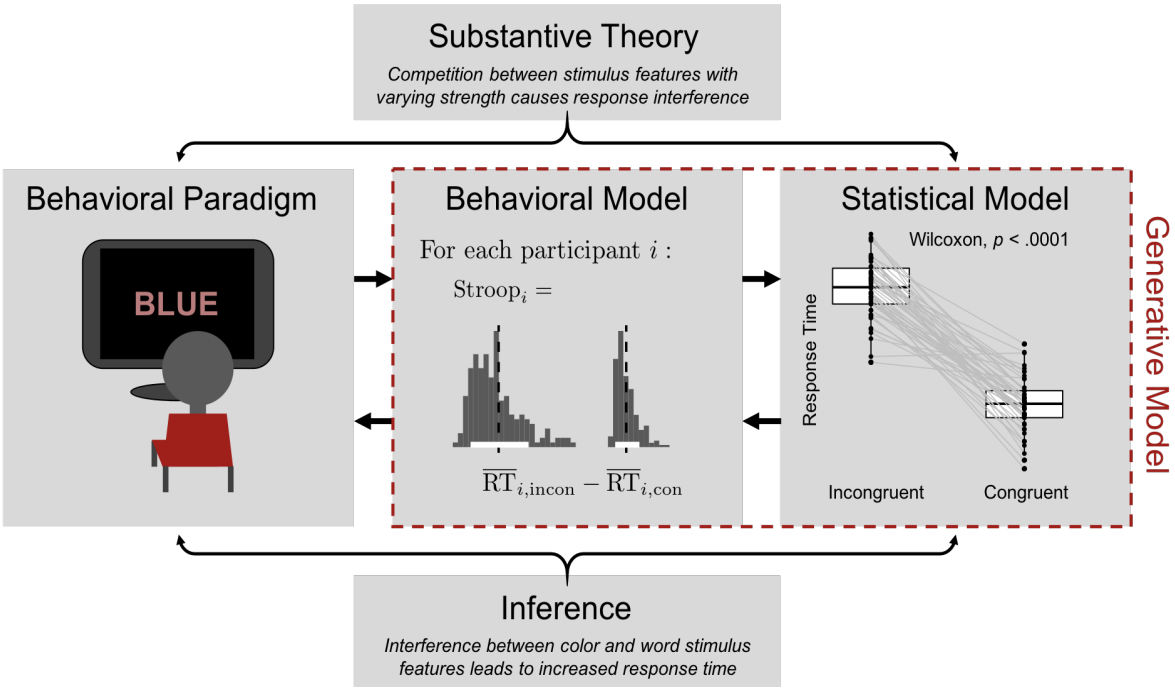
In sum, low test-retest reliabilities of popular behavioral paradigms limit their usefulness for testing and therefore developing explanatory theories of individual differences. Few substantive proposals have emerged to address this problem. Existing suggestions include abandoning unreliable measures altogether, recruiting far more participants, and increasing the number of trials collected. These work-arounds limit areas of research that have sample size constraints (e.g. neuroimaging research and clinical studies of difficult-to-recruit populations) or that have theoretical underpinnings necessitating measurement using behavioral tasks (e.g., implicit social cognition). In the sections that follow, we demonstrate how simply enhancing the approach to analyzing data can improve the precision of individual differences measures, thereby improving

statistical inference and opening the doors to more principled explanatory theory development with behavioral tasks.

3. Breaking Down Inference

In addition to the behavioral paradigm within which data are collected, inferences about behavioral phenomena depend on both a *behavioral model* assumed to generate data from the task and a *statistical model* used to estimate parameters from the behavioral model. Figure 1 represents the interplay between these components, which we describe in more detail below.

Figure 1. Pathway from theory to inference with behavioral data. Behavioral tasks are designed to elicit behaviors that test the substantive theory. Behavioral models formally relate the theory to features of the observed behavior. Here we show the "behavioral model" often assumed when analyzing Stroop data. Finally, the statistical model is used to calibrate uncertainty in estimates from the behavioral model. Such data are traditionally analyzed using a two-stage approach, whereby point-estimates of behavior are entered into a secondary statistical model. By contrast, with generative modeling we construct a single model that integrates the entire data generating process, spanning trial-by-trial response times to the group-level effects (e.g., test-retest reliability, individual differences, etc.).



3.1 The Behavioral Paradigm

We define the behavioral paradigm by the stimuli, design space, response options, and other contextual features afforded to participants by a behavioral task. The Stroop task includes various word-color pair stimuli, response options for each possible color (e.g., blue, red, yellow, green), instructions about how to respond (e.g., based on colors of text), and the number of behavioral observations (trials) collected across conditions within the task (e.g., numbers of congruent versus incongruent trials). One challenge with implementing behavioral tasks is that, unlike standardized questionnaires where participants complete the same items, specific stimuli and numbers of trials often vary across studies (e.g., Judd et al., 2012; Wolsiefer et al., 2017), and sometimes even across individuals within studies. Traditional estimates derived from such tasks therefore vary as a function of stimuli used and numbers of observations, making them non-portable from the perspective of classical test theory (Rouder & Haaf, 2019). Non-portability means that statistical effects estimated from behavioral measures vary as a function of task properties (e.g., numbers of trials per participant), which can have adverse effects on psychometric properties such as test-retest reliability as well as predictive validity and convergent validity. For any given study, non-portability can lead to attenuated and overconfident estimates. For example, the test-retest correlations and confidence intervals can be shifted downward toward zero. By contrast, when averaging across studies as in meta-analyses, non-portability can render estimates unstable and altogether uninterpretable—especially when different labs use variations of the task.

Theoretical issues also arise when comparing different behavioral tasks that are intended to measure the same phenomenon. For example, the Stroop task can be viewed as one instantiation

of a potentially infinite set of alternative tasks for testing the verbal theory claim that "*competition between stimulus features causes response interference*". One alternative instantiation is the Flanker task, in which stimuli are directed arrowheads rather than conflicting word-color pairs. Interference is induced by changing the orientation of "distractor" arrows relevant to a "target" arrow to be congruent (e.g., < < < < <) or incongruent (e.g., < < > < <). In principle, both tasks include key design elements (i.e., variation in congruency) necessary to examine interference effects. Nevertheless, the two tasks have distinct task demands that evoke different behaviors, and these differences in task demands must be accounted for to meaningfully compare performance across the two tasks. In other words, a theory must consider the task itself, because the data, which the theory imparts meaning on, are generated by the task. This is the job of a *behavioral model*.

3.2 The Behavioral Model

Behavioral models formally represent relevant aspects of the data that relate to psychological theory. Although often overlooked, the behavioral models that are assumed to generate effects of interest may be more important than the paradigm itself. In the Stroop paradigm (see Figure 1), the behavioral model traditionally specifies the effect of interference as the difference in mean response times across the two types of stimuli (i.e., congruent and incongruent):

$$\text{Stroop}_i = \overline{\text{RT}}_{i,\text{incongruent}} - \overline{\text{RT}}_{i,\text{congruent}} \tag{1}$$

Equation 1 is indexed by i, indicating that the Stroop effect is calculated for each individual participant, where a positive Stroop effect indicates that average response time is longer for incongruent than congruent trials. By comparing the summary statistics associated with stimuli thought to instantiate different levels of interference, the behavioral model mathematically

embodies the overarching substantive theory (see Figure 1). After the mean differences are computed, resulting "Stroop effects" are then used to make statistical inferences such as between-groups or individual-level comparisons. Although Equation 1 is typically not interpreted as a behavioral model, it implicitly assumes a specific data-generating model (we expand on this point in section 6.2). We therefore consider its widespread use atheoretical.

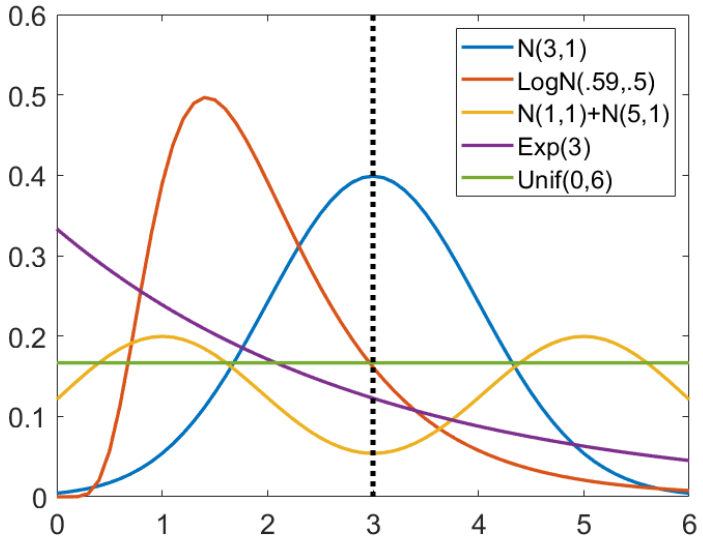
Interference in the Flanker task is typically estimated in the same way as the Stroop effect (Equation 1). Despite both tasks being designed to measure the same phenomenon, correlations of individual differences on the two measures are consistently small (Hedge et al., 2017). More broadly, low convergence across behavioral measures designed to capture the same construct is the rule rather than the exception, with similarly weak effects emerging across measures of self-control and implicit self-esteem (Cyders & Coskunpinar, 2011; Duckworth & Kern, 2011; Bosson et al., 2000). Although low convergence is partly due to low reliability and non-portability as described above, *we argue that the use of atheoretical behavioral models play a key role in producing low convergence because they fail to dissociate the psychological phenomenon of interest from auxiliary psychological processes*. That is, their lack of specificity creates ambiguity in interpretation. Although Equation 1 details the main effect of interest, it is not well constrained with information about each individual's pattern of responding across many different types of stimuli. Even at a very high level, the behavioral model in Equation 1 neglects the variance of the individual's pattern of response times, and so it is incapable of answering even the simplest of questions about what the mean difference in response times actually means in the context of a set of response times. The situation is far bleaker when one considers that the behavioral model does not bear in mind the distribution of interference, which could be

established by changing the task demands (i.e., the base rate). For example, the original Stroop task in 1935 included a "word response" condition in which participants were asked to verbally respond to the stimulus word rather than the color (e.g., saying "red" out loud when the word "Red" is shown in the color blue; MacLeod, 1991). In these conditions, interference technically still exists because the stimuli involve two properties (i.e., the word and the color) which can be either congruent or incongruent, yet the interference effects as measured by the behavioral model are far weaker than the "color response" condition counterparts. Because the behavioral model is not equipped to capture both of these theoretically relevant stimulus effects within a single task, we have little reason to believe it should precisely capture the same interference phenomenon in a different task using different stimuli such as the Flanker task. Nevertheless, low convergence across behavioral tasks is often interpreted to as a difference in constructs rather than incomplete behavioral models (e.g., Cyders & Coskunpinar, 2011; Duckworth & Kern, 2011; Bosson et al., 2000).

More generally, use of descriptive summary statistics such as mean differences limits inferences about mechanisms underlying various patterns of behavior produced by a given task. As demonstrated in Figure 2, many different distributions—which could imply different data-generating mechanisms—can yield the same mean. This is important because, once we collect behavioral data from participants, we are left with distributions of responses (e.g., choices, response times) for each individual. How we summarize these distributions has strong implications on resulting inference. When we limit ourselves to summary statistics, we can miss theoretically relevant aspects of our data such as variance (Johnson & Busemeyer, 2005), bimodality (Kvam, 2019a), or skew (Kvam & Busemeyer, 2020; Leth-Steensen et al., 2000).

Without employing a behavioral model that captures such characteristics, we can and often will draw inappropriate conclusions. For example, observed response time distributions in behavioral tasks such as the IAT, Stroop, Flanker, and Posner Cueing tasks are often heavily right-skewed (e.g., Hockley & Corballis, 1982; Whelan, 2008). In the Stroop task, both ignoring and removing skew results in incorrect conclusions: mean contrasts fail to uncover instances where congruent text color and color words *facilitate* performance, a phenomenon that can only be detected with a more theoretically informed behavioral model (i.e. a right-skewed ex-Gaussian distribution; Heathcote et al., 1991).

Figure 2. Qualitatively different distributions with the same mean. These distributions include a typical normal distribution (N, blue), a lognormal distribution (LogN, red), a sum of two normal distributions (yellow), an exponential distribution (Exp, purple), and a uniform distribution (Unif, green). All of these distributions have exactly the same mean and would therefore produce the same conclusions if analyzed with the behavioral model from Equation 1, regardless of how different their data-generating process may be.



Problems with behavioral summary statistics are not specific to response time data. Rotello et al. (2014) showed that using the ratio of correct to incorrect classifications as a metric for eye-witness detection accuracy led researchers to mistakenly infer that sequential lineups (i.e., suspects shown one at a time) are superior to simultaneous lineups (i.e., suspects all shown at once). This behavioral model, however, does not account for differences between conditions in participants' unwillingness to choose a suspect. The model therefore fails to capture the intended effect because the difference in detection accuracy is caused simply by participants being less likely to choose *any* suspect in the sequential lineups. When data are instead analyzed using a signal-detection theory model, the effect reverses (see also Kellen, 2019; Ross et al., 2020).

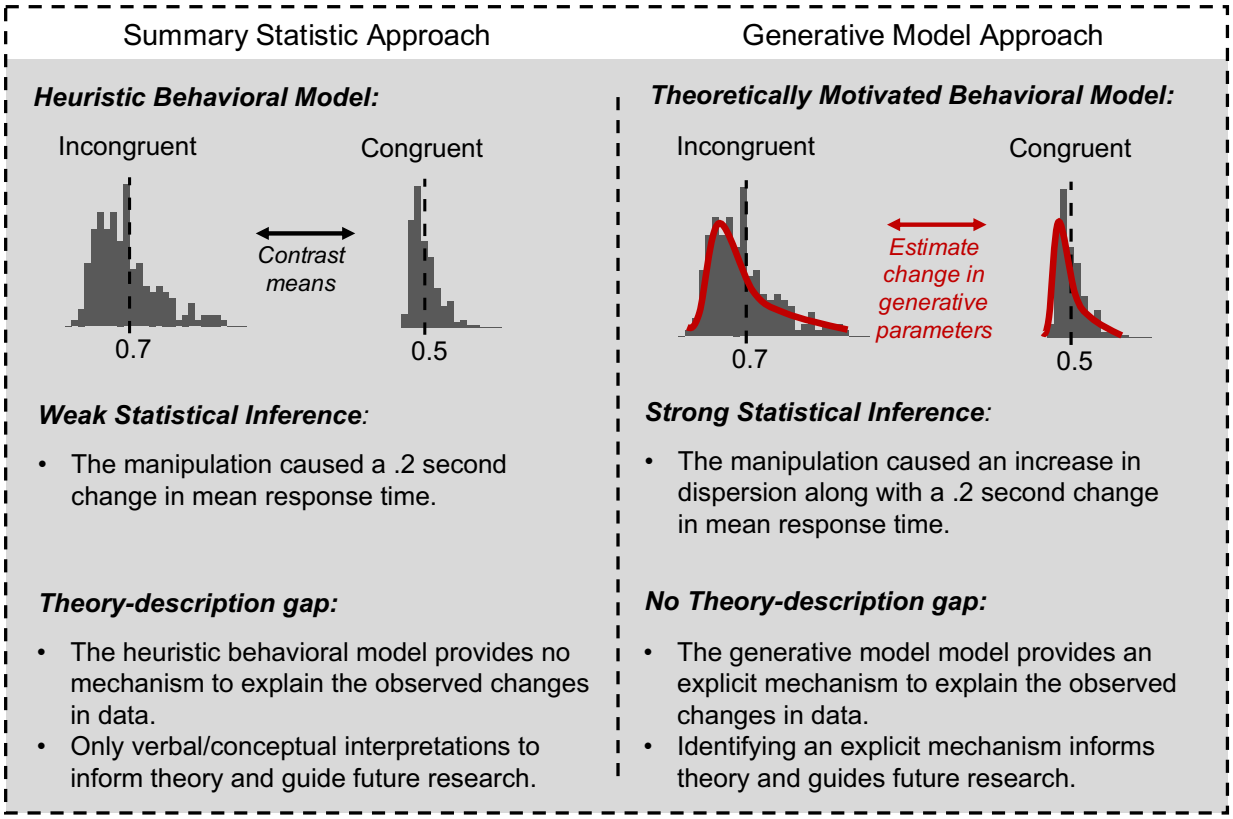
There are many other examples that demonstrate how the unquestioned use of behavioral summary statistics can obscure proper explanations of phenomena, leading to strong conclusions that clash with theory-informed approaches. It is important to note that drawing theoretically inappropriate conclusions will occur *even when heuristic approaches produce highly replicable results* (Devezer et al., 2019; 2020). Despite repeated warnings going back decades (e.g., Meehl, 1967), unchecked use of summary statistics as opposed to theoretically informed behavioral models continues to impede scientific progress. As stated by Regenwetter and Robinson (2017), "*No amount of replication would provide a theoretical foundation for such methods. What is needed is a theoretically sound process of deriving accurate predictions from concise assumptions*" (p. 540). It is critical that social, behavioral, and brain scientists work toward constructing models that *reproduce* theoretically relevant aspects of empirical data. Otherwise, we risk perpetuating the theory-description gap, using and misinterpreting models that fail to capture the intended behavioral mechanisms. Echoing broader discussions throughout the social,

behavioral, and brain sciences, a paradigm shift is called for in theory development, using tools made possible by advances in statistical computing.

Fortunately, frameworks to characterize behavioral data more precisely and thoroughly are available across disciplines, including mathematical psychology (Navarro, 2020; Townsend, 2008), neuroeconomics/value-based decision-making (Rangel et al., 2008; Busemeyer et al., 2019), computational psychiatry (Ahn & Busemeyer, 2016; Friston et al., 2014; Huys et al., 2016; Montague et al., 2012; Wiecki et al., 2015), neuroscience (Turner et al., 2013; Turner et al., 2017; Bahg et al., in press), and other areas throughout behavioral and cognitive science more broadly (Guest & Martin, 2020; Wilson & Collins, 2019). These frameworks use theoretically informed mechanisms to develop *generative models* of behavior that can be compared based on explanatory power. We define generative models of behavior as those that simulate data consistent with true behavioral observations *at the level of individual participants*[2]. Thus, mean contrasts do not qualify as generative models because they reduce individual-level data to a single estimate that cannot capture a full distribution of behavior (Equation 1). Figure 3 illustrates the difference between traditional and generative modeling approaches, using inference based on response times as an example. In the remainder of this article, we refer to this approach toward modeling behavior-generating processes as the *generative perspective*. In Section 4 (below), we present simulations to provide a concrete example of why this approach is useful.

[2]Although many computational models are developed with the goal of neurobiological plausibility or to estimate parameters with definite psychological interpretations, we note that neither is strictly necessary by our definition of generative modeling. More detailed delineations among models can, however, be disentangled according to stricter criteria (Jarecki et al., 2020).

Figure 3. Interpretations of summary statistic versus generative approaches to inferring between-condition changes in response times. The summary statistic approach is often used by default and chosen without reference to an underlying theory. By contrast, the generative approach begins with a model of behavior at the individual level (e.g., a lognormal distribution), and inferences are made by interpreting changes in model parameters across conditions, individuals, or other units of analysis. For example, if the response time distributions pertain to the Stroop task or IAT, the summary statistic approach simply infers a mean difference. The generative modeling approach infers a change in evidence dispersion, but not stimulus difficulty (we depict these parameters in section 5). Notably, increased dispersion produces a higher mean response time, but also a higher number of rapid response times. There are strong implications for our theory—what does it mean for stimulus interference or implicit bias to produce dispersed response times?



3.3 The Statistical Model

As shown in Figure 1, the statistical model is used to make inferences in the face of uncertainty, using parameters estimated from the behavioral model. Traditionally, summary statistics are estimated from behavioral data (e.g., percent correct, difference scores) and then entered into a secondary statistical model (e.g., linear regression). Group differences, correlations with other measures, and other theoretically relevant effects are then explored. With the Stroop and IAT, mean response time contrasts are used to estimate effects for each participant, and a linear model is used to determine if individual differences correlate with other variables, such as attention, self-control, or attitudes (Gawronski et al., 2016; Hedge et al., 2017). This two-stage approach—whereby effects are computed for each participant then used in a secondary statistical model—makes a strong assumption that when unmet contributes to poor test-retest reliability and low validity more generally (Ly et al., 2017; Rouder & Haaf, 2019; Turner et al., 2017). Specifically, it ignores uncertainty (i.e., measurement error) associated with each participant's summary score. In Figure 1, white bars in the middle panel represent confidence intervals for means of each response time distribution, and therefore depict uncertainty around "true" mean values. Ignoring this uncertainty is mathematically equivalent to assuming that individual-level Stroop effects are estimated with infinite precision (i.e., no error), or that we have an infinite number of trials for each participant. There are many examples of how averaging across individuals while ignoring this uncertainty leads to faulty inferences (e.g., Davis-Stober et al., 2016; Estes, 1956; Heathcote, et al., 2000; Liew, Howe, & Little, 2016; Pagan, 1984; Vandekerckhove, 2014; Turner et al., 2018), and in fact this inadequate treatment of individual-level uncertainty is directly responsible for making estimates from behavioral tasks non-portable (see Rouder & Haaf, 2019). By contrast, using statistical models that account for individual-level uncertainty leads to more

powerful group- and individual-level inferences (e.g., Haines et al., 2020; Romeu et al. 2019), as is shown next.

Many readers will have anticipated that hierarchical (mixed effects, random effects, multilevel) modeling is one framework that can account for uncertainty in behavioral data at both individual and group levels. Hierarchical modeling is already common practice in some fields (Gelman & Hill, 2007), and it is a natural solution to traditional designs where trials/observations are nested within individuals who are themselves nested within groups, as well as designs where amounts of individual-level data are limited. Key for our purposes, hierarchical Bayesian analysis solves the issues of non-portability in behavioral paradigms because it specifies a single model that *jointly* captures individual- and group-level uncertainty. Further, it allows us to specify arbitrarily complex models that best meet our generative assumptions (i.e., properties of the underlying mechanism), which is not necessarily true of other approaches that accommodate measurement error such as structural equation modeling or classical attenuation corrections (e.g., Kurdi et al., 2019; Westfall & Yarkoni, 2016)[3]. By specifying a hierarchical model over individual-level parameters of the behavioral model, we are building a full generative model spanning from within-person trial-level variation to between-person group-level effects/trends of interest. Variants of this model can be constructed, each with different assumptions, and then compared against the data. These models and their evaluation are presented in section 6.

[3] Although we cannot provide a detailed explanation here, Rouder and Haaf (2019) provide a comprehensive account of how hierarchical Bayesian models address psychometric issues such as non-portability, and both limitations to and future directions for hierarchical approaches (Rouder et al., 2019). Applied examples that demonstrate advantages of hierarchical Bayesian modeling over traditional two-stage approaches include Haines et al. (2020) and Romeu et al. (2019). For more general discussions, we refer interested readers to the extensive literature on hierarchical Bayesian modeling and related approaches (e.g., Craigmile et al., 2010; Kruschke, 2015; Lee, 2011; Ly et al., 2017; Rouder & Lu, 2005; Shiffrin et al., 2008).

Our central premise is that atheoretical behavioral models that rely on summary statistics (i.e., the *summary statistic* approach) and the two-stage approach described above produce an impoverished and incomplete view of rich individual differences underlying behavioral data. We argue further that generative modeling is better suited to detect and understand individual differences in behavioral data compared to traditional approaches. Here, we focus our attention on how generative modeling affects test-retest reliability, but the same logic applies to any correlation measured between two constructs. Given that Rouder and Haaf (2019) already provide a thorough account of how hierarchical models yield higher test-retest reliabilities than the traditional two-stage approach, we focus on the choice of a behavioral model in the simulations presented below.

4. Simulated Demonstration

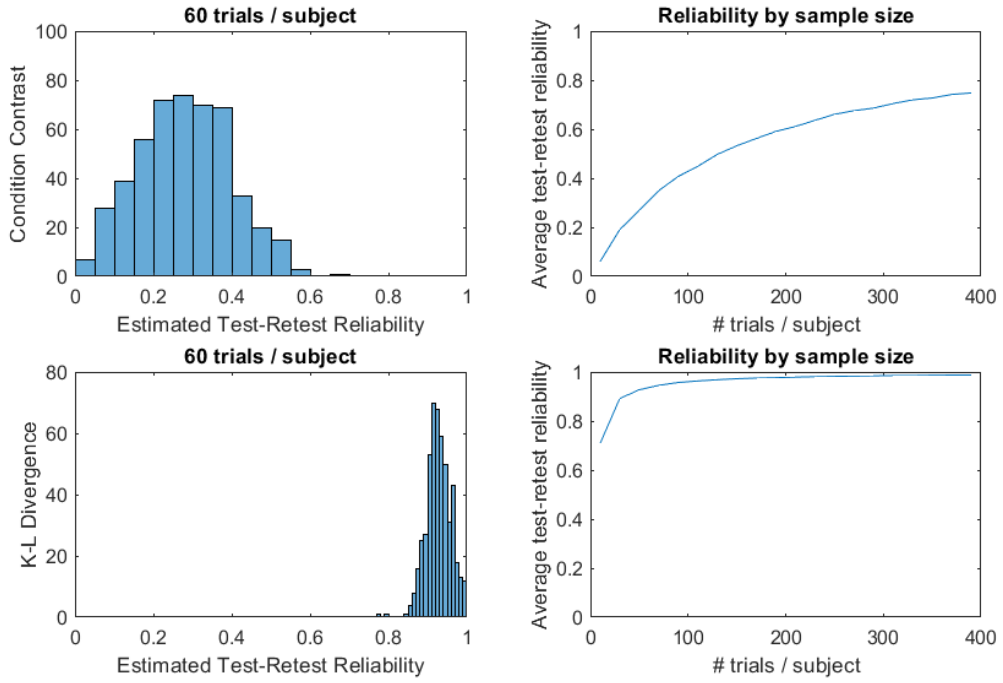
Using simulated response time data, we compare the following two "behavioral models" for estimating reliability: (1) the traditional two-stage summary statistic method (compute means, take the difference, and compute a test-retest correlation), and (2) a method that contrasts the distributions holistically (as articulated below) before computing a test-retest correlation. To generate simulated data, we drew response times from a lognormal distribution (right-skewed as in most response time tasks; Figure 2) for each participant and condition, then compared test-retest correlations across the approaches. We simulated 150 "participants", each of whom completed the response time task at two different sessions, with an artificial "congruent" and "incongruent" condition at each timepoint. Critically, the parameters used to generate response time data at each timepoint were *exactly the same for each participant*—the generative

parameters had true test-retest correlations of $r = 1.0$. The procedure produced right-skewed response time distributions with 80% of draws between 300 and 2000 milliseconds (see the online supplement for additional details).

For each simulated participant, we conducted two reliability tests. The first simulated a traditional reliability analysis of performance (e.g., test-retest reliability of mean response time difference between congruent and incongruent trials in the Stroop task), with knowledge that the true generating parameters were unchanged across test and retest. For each of the two sessions, we computed the mean difference between each participant's "incongruent" and "congruent" response time distributions. Next, we estimated Pearson correlations between the Session 1 and Session 2 mean differences across participants as an index of test-retest reliability. We repeated this procedure 1,000 times at sample sizes ranging from 10 to 400 per participant.

Figure 4 shows results of this analysis. The top left panel shows an example distribution of inferred test-retest estimates across 1,000 repetitions for a sample size of 60 trials. These test-retest reliabilities of mean contrasts ranged from close to $r = 0$ to $r = .5$ (middle-left panel, Figure 4). Test-retest reliability improved substantially with more trials for each participant, to around $r = .8$ at 400 trials (middle-right panel).

Figure 4. Test-retest reliability simulations comparing the mean difference between two conditions (top), and contrasting distributions using K-L divergence (bottom). The left panels show estimated reliabilities for sample sizes of 60 response times per participant (a typical size for the IAT) across 1,000 simulations. The right panels show how average reliability of these contrasts changes across sample sizes.



Taking the mean of a distribution is only one way to characterize the distribution, and mean contrasts are therefore only one way to represent our substantive psychological theory within the behavioral model (Figure 1). Given that means alone are often imprecise when characterizing entire distributions (Figure 2), a behavioral model that captures the entire shape of participants' individual response time distributions may yield very different inferences. To demonstrate how important distributional information can be, we performed a second reliability analysis which used Kullback-Leibler (K-L) divergence to quantify the relative difference between each participant's response time distributions across trials within conditions. K-L divergence makes no assumptions about the shape of response time distributions. However, it is not directly interpretable in the sense of a mean contrast (see the online Supplement for K-L divergence details). Nevertheless, it is useful to demonstrate the importance of distributional information for recovering individual differences. We estimated test-retest reliability as the Pearson correlation of the K-L divergence measure, as opposed to a mean contrast, across the simulated sessions for each of 1,000 repetitions. Results appear in the bottom panels of Figure 4. Most test-retest reliabilities based on K-L divergence between congruent and incongruent trials were between r = .85 and 1.0. Use of a distribution-informed metric was therefore much more successful in recovering the true test-retest of reliability (r = 1.0).

4.1 Empirical and Theoretical Implications

Results from our test-retest reliability simulations have both empirical and theoretical implications. Empirically, achieving desirable psychometric properties such as high test-retest reliabilities requires many behavioral observations (trials) from each participant—particularly when relying on traditional behavioral models (e.g., mean contrasts). Indeed, the reliability of the

mean contrasts only began to approach $r = .8$ after 400 trials per participant per condition, which is far beyond the typical number of trials used in such tasks. Theoretically, the implications are much broader. Psychometric properties of behavioral paradigms are highly dependent on underlying behavioral models (e.g., mean contrasts versus K-L divergence). Accordingly, models that are sensitive to the entire distribution of individual-level behavior are better suited for recovering individual differences. For response times, this necessitates behavioral models that capture full distributions of response times across trials, and the right-skewed nature often observed for such distributions (e.g., Heathcote et al., 1991; Hockley & Corballis, 1982; Kvam & Busemeyer, 2020; Leth-Steensen et al., 2000; Whelan, 2008). For dichotomous or categorical data, as we will demonstrate with the delay discounting task, this requires models that produce probabilities that represent how likely participants are to select each of the possible responses.

5. Developing Generative Behavioral Models

In this section, we illustrate how generative models can be built up from very primitive assumptions to fully characterize data. As you will see, even a very simple generative models can improve upon the problems with the behavioral model in Equation 1.

5.1 The Normal Model

To characterize response time data, a generative model must obey some very simple properties. First, response times are never negative. Second, response times typically have some spread or variance around a central tendency. Third, the variance is not spread evenly: the variance typically increases linearly with the mean of the response time (Wagenmakers & Brown, 2007), and so there is typically larger spread on the right side of the distribution than the left, which is

often called “right skew.” Fourth, there is typically some linear shift associated with response times, such that they are usually not near the lower bound of zero. As we build our generative model, we will bear these simple properties in mind.

Perhaps the simplest behavioral model that can generate a full distribution of response times is the normal (Gaussian) distribution. For now, the normal distribution will not capture many of the aforementioned properties, but it can still be useful for exemplifying the shift away from the behavioral model in Equation 1 and the generative perspective. At the very least, the normal distribution characterizes both the central tendency and the variance or spread of the response time distribution.

Using the Stroop task as a running example, each individual’s set of response times can be conceptualized as arising from a separate normal distribution. Parameters from each distribution (e.g., means/standard deviations) are specific to each person within each task condition. Similar to the K-L divergence test-retest simulation, the Stroop effect can be characterized by within-participant changes in the shape of each individual’s response time distribution across trials within conditions. When using a normal generative distribution, the shape of the response time distribution is characterized by changes in the mean and standard deviation parameters across congruent and incongruent condition trials for each participant. We can write the normal generative model as

$$\mathbf{RT}_{i,c,t} \sim \mathcal{N}(\mu_{i,c,t}, \sigma_{i,c,t}) \tag{2}$$

where $\mathrm{RT}_{i,c,t}$ contains the set of response times for participant i in condition c during experimental session t. The notation $\mathrm{RT} \sim N(a, b)$ signifies that the response times are drawn

from a generative process of a normal distribution (N) with mean a and standard deviation b. In Equation 2, the collection of response times in each block of our experiment are separately characterized by a specific mean ($\mu_{i,c,t}$) and standard deviation ($\sigma_{i,c,t}$).

To facilitate interpretation, we will introduce a relabeling of the terms in Equation 2 based on the conditions they correspond to. First, we label the congruent condition (i.e., the first condition $c = 1$) as a baseline condition, where $\mathrm{RT}_{i,1,t} = \mathrm{RT}_{i,base,t}$, characterized by a baseline mean $\mu_{i,1,t} = \mu_{i,base,t}$ and baseline standard deviation $\sigma_{i,1,t} = \exp(\sigma_{i,base,t})$.[4] To isolate the effects of interference, or Stroop effects, we labeled a parameter Δ to signify the change from the baseline condition to the condition of interest (e.g., incongruent condition). This means that $\mathrm{RT}_{i,2,t}$ is characterized by a mean $\mu_{i,2,t} = \mu_{i,base,t} + \mu_{i,\Delta,t}$ and standard deviation $\sigma_{i,2,t} = \exp(\sigma_{i,base,t} + \sigma_{i,\Delta,t})$. Hence, whereas the behavioral model in Equation 1 reduces the response time data into a single summary statistic per condition, the behavioral model in Equation 2 will reduce the data into two parameters per condition, parameters which, as we discuss below, can be assessed in terms of their own mean and variance (Williams et al., 2019).

5.2 The Lognormal Model

Although the normal generative model provides a better characterization of distributional changes in response times across conditions than Equation 1, the model is limited in the sense that it is not flexible enough to obey all the simple properties of response time we outlined

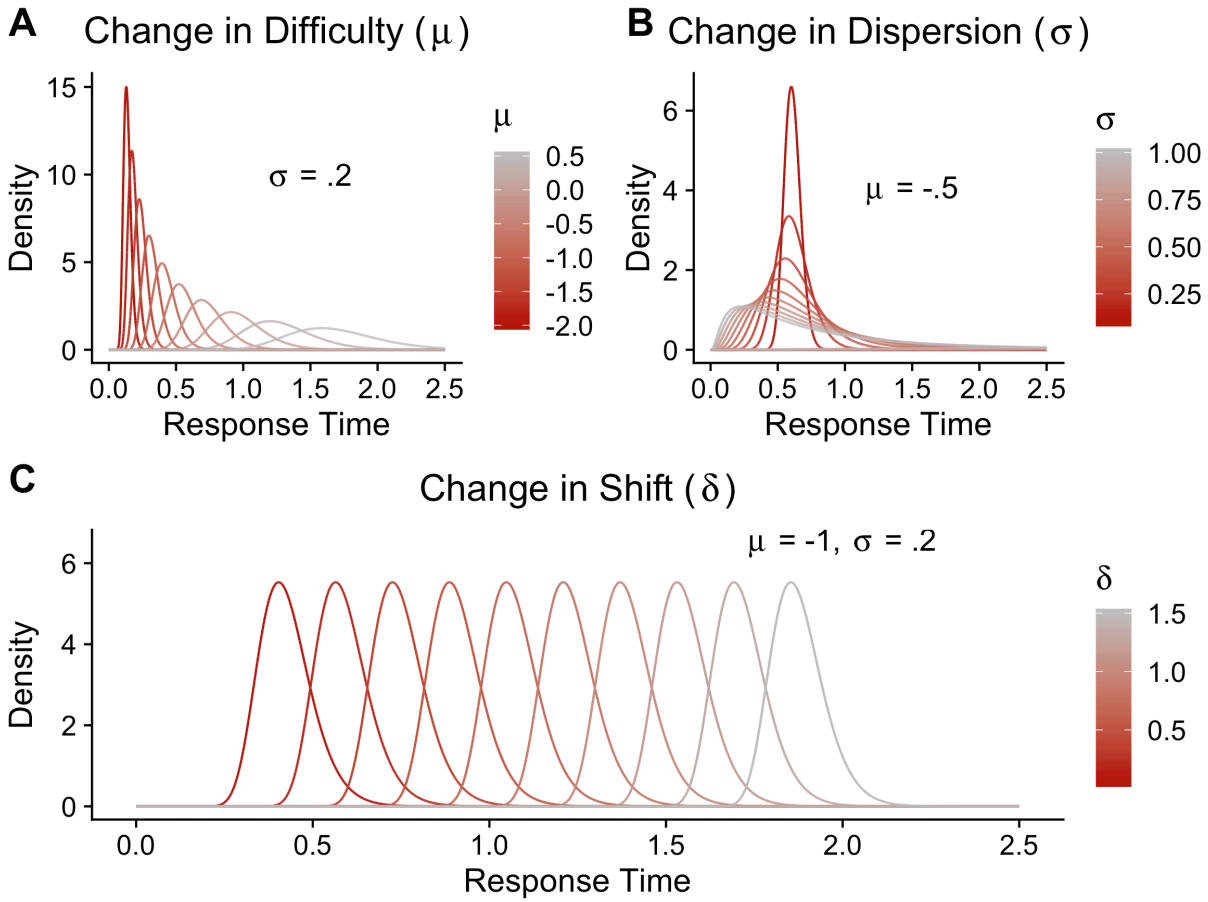
[4] Note that we estimate the *base* and Δ standard deviation parameters on the log scale and exponentially transform them to ensure they are greater than 0. Therefore, the test-retest correlation for the Δ standard deviation parameters indicates their correlation on the log scale. See the online supplement for details.

above. In particular, the normal model (1) can produce negative response times, and (2) cannot capture asymmetric variance with respect to the mean (i.e., right skew). One simple adjustment we can make is to logarithmically transform the response time data, and assume a normal model on this transformed data. This process is equivalent to assuming that the response time data come from a different generative model called the lognormal distribution. Given this equivalence, we can specify a more theoretically consistent generative model as

$$\mathbf{RT}_{i,c,t} \sim \text{Lognormal}(\mu_{i,c,t}, \sigma_{i,c,t}) \tag{3}$$

With this small adjustment, parameters $\mu_{i,c,t}$ and $\sigma_{i,c,t}$ will have very different abilities when characterizing the many shapes of response time distributions. The lognormal model has a very helpful property in how the mean and standard deviation parameters interact (the *law of response time*; Wagenmakers & Brown, 2007): an increase in either parameter, holding the other constant, produces an increase in both the mean and standard deviation of the response times predicted by the model. As illustrations, Figures 5A and 5B show how changes in either parameter change the shape of the predicted response time data. Each possible distribution shape can be viewed as a prediction about how each participant's response time data should look, where the possible shapes are constrained by our commitments (or hypotheses) regarding the data-generating process (i.e. the lognormal model).

Figure 5. Lognormal and shifted-lognormal generative distributions. (A) For the lognormal distribution, changes in the μ parameter (interpreted as "stimulus difficulty") produce changes in both means and variances of response time distributions (see Heathcote & Love, 2012; Rouder et al., 2014). (B) The σ parameter controls dispersion; changes in σ affect means and ranges of likely response times, but medians remain constant. (C) The shift parameter δ translates the distribution forward in time without changing the shape of the response time distribution.



5.3 The Shifted Lognormal Model

Although the lognormal model is an improvement over the normal model, it still misses one important property of response time data. It is well established that different response modalities (e.g., responding with a key press versus mouse, versus verbal response) can produce shifts in response time distributions, even when the task demands and underlying evidence accumulation dynamics are identical (e.g., Gomez et al., 2015). Typically, this extra time taken to interact with the stimuli and apparatus is not considered part of the decision process, and is often referred to as "non-decision" time to make this theoretical position clear. Although non-decision factors seem unimportant, their presence may compromise our ability to accurately characterize response time data. For example, suppose a person completes a Stroop task in two conditions, one in which they are asked to respond verbally, and one in which they are asked to manually select an option. Even when we can assume that the individual will follow the same decision process in identifying the color of the word (i.e., they have the same $\mu_{i,c,t}$ parameter), there are likely to be differences in executing the response across conditions. For example, if it took longer to manually select a response compared to the verbal condition, we would expect the response times to be shifted relative to the verbal condition. In this case, fitting the lognormal distribution to the observed response times would lead to different estimates for $\mu_{i,c,t}$ across the two conditions because the simple lognormal is not specified correctly relative to the demands of the experiment. Consequently, having different estimates for $\mu_{i,c,t}$ might result in different interpretations about cognitive factors across the two contexts, when in reality, the factors were related to the influence of non-decision factors.

A simple solution is to adjust the lognormal distribution by introducing an additional parameter δ to move the distribution a distance of δ away from zero. Figure 5C illustrates the effect of δ on a specific lognormal distribution. To impose some theoretical constraints, we could assume δ is specific to each person, and that it is unlikely to change between conditions within a behavioral task. In our example above, this assumption would be inappropriate, but for the analyses we perform in later sections of the paper, such assumptions are justified by the manner in which the data were collected. With this new shift parameter and imposed theoretical constraints, we can now specify a shifted-lognormal model as

$$\mathbf{RT}_{i,c,t} \sim \text{Shifted-Lognormal}(\delta_{i,t}, \mu_{i,c,t}, \sigma_{i,c,t}) \tag{4}$$

where $\mu_{i,c,t}$ and $\sigma_{i,c,t}$ have the same interpretations as described for Equation 3, and $\delta_{i,t}$ indicates the amount of shift or "non-decision time" specific to each individual at each of the two experimental sessions.

Development of a generative model involves iteratively refining the model's assumptions until they are as consistent with existing domain knowledge as possible (e.g., developing a models that adheres to the simple properties of response time we defined above). The appropriateness of the model is then evaluated by fitting it to empirical data. In the next section, we use data from previous studies to demonstrate how generative models can help us understand the reliability of individual differences in tasks that are popular across social, behavioral, and brain sciences. As mentioned in the Introduction, recent work has called into question the reliability of behavioral tasks for measuring individual differences (Hedge et al, 2017). Our goal in choosing these tasks is to demonstrate the generality of generative modeling across both different content areas of research and across different types of data (response times versus choice data in the delay

discounting task). By demonstrating the consistent increase in test-retest reliability afforded by generative modeling, regardless of the task or data type, we hope to convince readers that rich theories of individual differences can in fact be developed based on behavioral data, but that it requires a shift in focus toward modeling data-generating mechanisms. Details pertaining to each dataset and task appear below.

6. Method

6.1 Datasets and Behavioral Paradigms

In total, we re-analyzed data from three different studies. First, we analyzed data from Hedge et al. (2017), who collected data on the Stroop, Flanker, and Posner Cueing tasks. Second, we analyzed data from Gawronski et al. (2017), who collected data on the Self-Concept (introversion/extraversion) and Race (Black/White) versions of the Implicit Association Test (IAT). Lastly, we analyzed data from Ahn et al. (2020), who collected data on the delay discounting task. Individually, each of these behavioral tasks has produced a deep body of literature—the Stroop, Flanker, and Posner Cueing tasks have been used extensively to develop theories of attention and inhibitory control, the IAT has been used to develop theories of implicit cognition and evaluations, and the delay discounting task has been used to develop theories of impulsivity and self-control. On Google Scholar alone (as of August 2020), the collective citation count of the original research pertaining to these tasks is over 54,000 (Eriksen & Eriksen, 1974; Green & Myerson, 2004; Greenwald et al., 1998; Mazur, 1987; Posner, 1980; Stroop, 1935). Further, these tasks cover areas of research spanning from psychology and neuroscience to behavioral economics.

Given that the Stroop task has served as the running example throughout this article, we describe the details of the Stroop task from Hedge et al. (2017) below. We provide details of all other tasks and datasets in the online supplement.

For the Stroop task, two sets of participants (n = 47, n = 60 for Studies 1 and 2, as reported in the original work) performed the task in two separate sessions separated by three weeks. The main effect of interest is the contrast between congruent and incongruent conditions. Specifically, participants responded to the color of a word, which could be red, blue, green, or yellow. The word could be the same as the font color (e.g., the word "red" colored in red font; congruent condition or c = 1 [see online supplementary text]), a non-color word (e.g., "ship"; neutral condition), or a color word mapping onto another response option (e.g., the word "red" colored blue, green, or yellow; incongruent condition or c = 2). Participants completed 240 trials in each of the three conditions.

6.2 Data Analysis

6.2.1 Data Preprocessing

For all tasks involving response times, we removed trials for which response times were recorded as < 0, assuming that such trials could not be part of the data-generating process[5]. For the delay discounting task, we did not remove trials. We used these liberal inclusion criteria primarily to keep our models consistent with the goals of generative modeling, but also to demonstrate the utility of hierarchical modeling. By keeping all trials (except negative response

[5] RTs < 0 were only found for 8 trials in total across 4 participants in the Posner Cueing task. We assume these RTs were recorded as less than 0 due to experimenter error (e.g. keyboard responses not being flushed before stimulus presentation), and therefore we removed them.

times), we can identify regions of model misfit that offer insights into cognitive mechanisms that would otherwise be obscured by oversimplified preprocessing choices (e.g., removing trials with response times less than 100 milliseconds). Such heuristic preprocessing choices tend to have strong, unpredictable effects on inference (Parsons, 2020).

6.2.2 Two-Stage Summary Statistic as Behavioral Model Approach

The two-stage approach proceeds by reducing behavior within each participant to a point estimate before entering the resulting point estimates into a secondary statistical model to make inference. Below, we describe its implementation for each task.

6.2.2.1 Response Time Tasks.

For the IAT, Stroop, Flanker, and Posner Cueing tasks, our first analysis followed the two-stage approach as described in the simulation study above. We computed mean contrasts across task conditions for each participant using Equation 1[6]. In addition, we computed standard deviation contrasts for comparison with the generative models (i.e., standard deviations of incongruent condition response times minus standard deviations of congruent condition response times). To estimate test-retest reliabilities, we computed Pearson correlations across participants for the mean and standard deviation contrasts.

6.2.2.2 Delay Discounting Task

[6] We recognize that the IAT is typically scored using the D-score, which is a mean contrast divided by the pooled standard deviation (Greenwald et al., 2003). However, the D-score also uses multiple empirically-derived preprocessing steps, including removing response times > 10,000 ms, removing participants with > 10% trials with response times < 300 ms, and replacing response times for all incorrect response trials with the mean response time of correct responses + 600 ms. We therefore used the simple mean contrast to maintain consistency across tasks and to facilitate comparison of summary statistic versus generative modeling approaches.

We used maximum likelihood estimation to estimate discounting rates (k) and choice sensitivity parameters (c) from a hyperbolic model for each participant and session, followed by Pearson correlations across participant to estimate test-retest reliabilities of model parameter point estimates (see online supplementary text for details)[7]. We compare these estimates to a hierarchical Bayesian estimation approach described below.

6.2.3 Generative Modeling Approach

If the goal is to make group-level inferences, hierarchical models allow us to appropriately account for individual-level uncertainty (see section 3.3). Further, hierarchical models can increase precision of parameter estimates at the individual level. Below, we extend the concept of generative modeling from individual- to group-level model parameters.

6.2.3.1 Response Time Models

We have now defined generative models of individual-level behavior for both response time tasks (normal, lognormal, and shifted lognormal models) and the delay discounting task (hyperbolic model). The next step toward building full generative models of test-retest reliability is to define group-level probability distributions for individual-level parameters. Starting with the three response time models, we assume that all i individual-level parameters in the congruent

[7]The sample mean and standard deviation contrast approach used for response time models is equivalent to assuming that response times are generated by normal distributions within participants (as in generative models), wherein the sample mean and standard deviation are maximum likelihood estimators for the normal generative distribution mean and standard deviation. The contrasts can therefore be thought of as contrasts between maximum likelihood estimates of normal generative models. This correspondence motivates our use of maximum likelihood estimation for the delay discounting model to show that benefits of generative modeling generalize beyond response time measures (see online supplementary text for details).

task condition at each of the two sessions t are drawn from a normal group-level distributions with unknown means and standard deviations[8]:

$$\begin{aligned}\mu_{i,\text{base},t} &\sim \mathcal{N}(\mu_{\text{mean},\text{base},t}, \mu_{\text{sd},\text{base},t}) \\ \sigma_{i,\text{base},t} &\sim \mathcal{N}(\sigma_{\text{mean},\text{base},t}, \sigma_{\text{sd},\text{base},t})\end{aligned} \tag{7}$$

The group-level normal distributions here are considered prior models (or prior distributions) on the individual-level parameters. Estimating group-level parameters from prior models allows for information to be pooled across participants such that each individual-level estimate influences its corresponding group-level mean and standard deviation estimates, which in turn influence all other individual-level estimates. This interplay between the individual- and group-level parameters produces regression of individual-level estimates toward the group mean (also referred to as *hierarchical pooling*, *shrinkage*, or *regularization*), which increases precision of individual-level estimates (Gelman et al., 2014). Note that the normal distribution functions similarly for individual-level latent parameters in Equation 7 as they do for observed response times in Equation 2. The assumption in both cases is that a normal distribution at one level of analysis generates observed or unobserved data at another level (e.g., observed response times are generated by normal distributions within participants, with unobserved means and standard deviations generated from normal group-level distributions). *This joint specification of relations between parameters over all levels of analysis embodies the generative perspective*. It allows for group- and individual-level model parameters to be estimated simultaneously (we illustrate the effect of these generative assumptions on individual-level parameters in section 7.6). Although we do not demonstrate it here, the group-level model (i.e., Equation 7) can be extended to

[8] As described in section 5.1, individual-level standard deviations were exponentially transformed such that $\sigma_{i,1,t} = \exp(\sigma_{i,base,t})$. Therefore, the normal group-level distribution on $\sigma_{i,base,t}$ corresponds to a lognormal distribution on $\sigma_{i,1,t}$.

estimate relations between personality traits and decision mechanisms (e.g., Haines et al., 2020), or to generalize parameter estimates beyond non-representative samples (Kennedy & Gelman, 2019).

To estimate test-retest reliability, we can assume that individual-level change parameters (e.g., $\mu_{i,\Delta,t}$ and $\sigma_{i,\Delta,t}$) are correlated across sessions. Staying true to the generative perspective, we can estimate this correlation by assuming scores are drawn from a multivariate normal distributions rather than independent normal distributions as in Equation 7:

$$\begin{bmatrix}\mu_{i,\Delta,1}\\ \mu_{i,\Delta,2}\end{bmatrix} \sim \text{MVNormal}\left(\begin{bmatrix}\mu_{\text{mean},\Delta,1}\\ \mu_{\text{mean},\Delta,2}\end{bmatrix}, \mathbf{S}_{\mu}\right)$$
$$\begin{bmatrix}\sigma_{i,\Delta,1}\\ \sigma_{i,\Delta,2}\end{bmatrix} \sim \text{MVNormal}\left(\begin{bmatrix}\sigma_{\text{mean},\Delta,1}\\ \sigma_{\text{mean},\Delta,2}\end{bmatrix}, \mathbf{S}_{\sigma}\right) \quad (8)$$

Using a multivariate normal distribution allows us to estimate covariances ($\mathbf{S}_{\mu}$ and $\mathbf{S}_{\sigma}$ matrices) between individual-level parameters across sessions that can be decomposed into group-level parameter variances and the correlation between individual-level parameters across sessions—this correlation represents the test-retest reliability of the generative model parameters (see the online supplementary text for mathematical details). If the correlation is zero, then Equation 8 is equivalent to Equation 7 (i.e. the normal distributions are independent).

For the shifted lognormal model, we estimated a single shift parameter for each participant at each timepoint (assuming that shift is equivalent between task conditions). Details about the shift parameter specification and prior distributions for group-level parameters in equations 7-8 are available in the online supplementary text.

6.2.3.2 Delay Discounting Model

Extending the individual-level hyperbolic delay discounting model to a full generative model that can estimate test-retest reliability follows the same logic as outlined for response time models. We used the same multivariate normal distribution parameterization to estimate test-retest correlations between discounting rate (k) and choices sensitivity (c) parameters (for details, see online supplementary text).

6.2.4 Parameter Estimation

A benefit of Bayesian estimation is that after specifying a joint probability model (i.e. the full group- and individual-level generative model), it is possible to compute conditional probabilities that determine which parameter values are most credible given the observed data. This results in *posterior distributions* over model parameters that are directly interpretable as the probability that the parameter takes on a specific value given the model and data[9]. Because computing conditional probabilities analytically requires solving complex and often intractable integrals, Bayesian model parameters are typically estimated using numerical integration methods. We estimated parameters from all models using Stan (version 2.19.2), a probabilistic programming language that uses a variant of Markov Chain Monte Carlo to estimate posterior distributions for parameters within Bayesian models (Carpenter et al., 2016). Details are described in the online supplementary text.

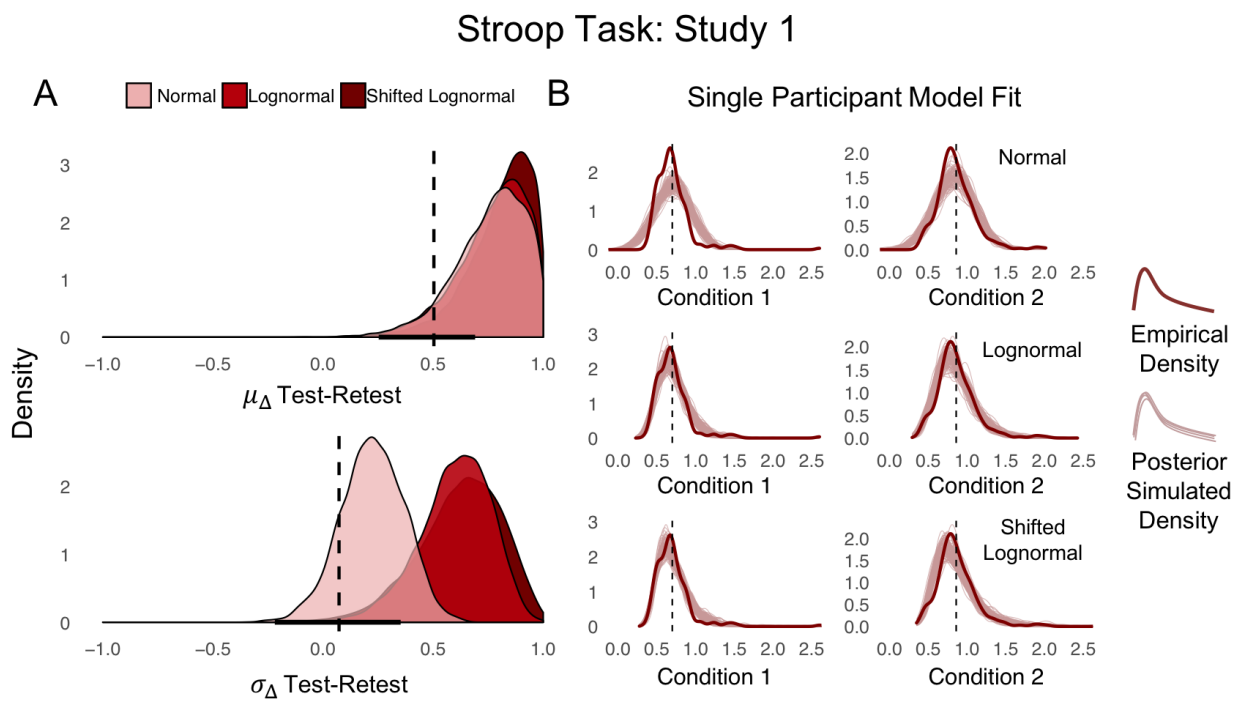
7. Results

[9] Posterior distributions therefore differ from frequentist confidence intervals, for which probability is a property of the long-run frequency of the confidence interval producing procedure rather than of the specific parameter value of interest.

To facilitate interpretation of our results, we provide a detailed interpretation of the data pertaining to the Stroop task, followed by a brief overview of all other tasks. Detailed results on each of the tasks are included in the online supplement.

The results for the Stroop task in Study 1 of Hedge et al. (2017) are shown in Figure 6. Panel A compares the estimated test-retest correlation for the two-stage approach versus each of the normal, lognormal, and shifted lognormal generative models. For the two-stage mean and standard deviation contrasts, the test-retest correlations were r = .5 (95% CI = [.25, .69]) and r = .07 (95% CI = [-.22, .35]), respectively. These estimates are consistent with the results originally obtained by Hedge et al. (2017), who reported a test-retest intraclass correlation for the mean contrast of ICC = .6 (95% CI = [.31, .78]). The discrepancy between their estimate and our own is due to both our inclusion of all trials and participants (i.e. no data pre-processing) and our use of the Pearson's as opposed to intraclass correlation. Regardless of the exact method, it is clear that the Stroop effect is indeed "unreliable" when estimated using the two-stage approach: with a test-retest reliability of r = .5 to r = .6, we would need well over 200 participants to detect (with adequate power) a simple correlation between the Stroop effect and an alternative individual difference measure with similar reliability (see Hedge et al., 2017). Such design constraints inherently limit the utility of the Stroop effect as a measure to advance theories of individual differences.

Figure 6. Test-retest correlations and model misfit for the Stroop task. (A) Posterior distributions for the test-retest correlations of each of the three generative models (red distributions) versus the two-stage sample mean/standard deviation approach (vertical dotted black line with corresponding horizontal 95% confidence interval) for the Stroop task in Study 1 of Hedge et al. (2017). (B) Posterior predictive simulations and sample means (vertical dotted black lines) for each of the generative models for a representative subject.



We now focus attention on the generative model estimates in Figure 6A, which take the form of *posterior probability distributions* rather than point estimates and confidence intervals. Note that the posterior distribution can be interpreted in a variety of ways depending on our goals. For example, if one is interested in the probability that the test-retest correlation of the normal generative model is greater than the two-stage estimate of $r = .5$, this quantity can be easily computed as the proportion of the posterior distribution greater than $r = .5$. Alternatively, if we are interested in the single most likely test-retest estimate, we can simply locate the mode (or the peak) of the posterior distribution. However, we are typically interested not only in a single value, such as the mode, but a range of likely values that help us convey uncertainty. Therefore, to facilitate interpretability of posterior distributions, we report the posterior mean (sometimes referred to as the posterior "expectation") along with the 95% *highest density interval* (HDI). An HDI is a generalization of the concept of the mode, but it is an interval rather than a single value. For example, a 20% HDI would contain 20% of the area of the entire posterior distribution, where every value within the interval is more likely than every value outside of the interval. We report 95% HDIs to maintain consistency with the 95% CIs reported for the two-stage approach, although we caution readers that HDIs and CIs are different concepts that have different interpretations. As has been a focus throughout this article, a mean and interval alone may do a poor job of summarizing a skewed distribution, so we recommend that readers interpret the posterior distributions holistically to fully appreciate the generative model estimates.

For the generative models, the posterior distributions for the mean/difficulty contrast parameters ($\mu_{i,\Delta}$) across models were concentrated above the two-stage estimates (posterior mean test-retest ranging from $r = .76$ to $r = .81$). Further, the 95% HDIs for the difficulty parameter in each of the

normal (95% HDI = [.46, 1.00]), lognormal (95% HDI = [.47, 1.00]), and shifted-lognormal (95% HDI = [.53, 1.00]) models included $r = 1.00$, indicating that we cannot rule out the possibility that there is in fact a perfect correlation in the mean/difficulty parameter contrast between retest sessions. This can be observed in the posterior distributions, which are concentrated against the upper limit of the correlation at $r = 1.00$. Posterior distributions for the standard deviation/dispersion parameters ($\sigma_{i,\Delta}$) were also concentrated above the two-stage estimates, although primarily for the lognormal and shifted lognormal models (posterior mean test-retest ranging from $r = .23$ to $r = .62$). In fact, the test-retest estimate for the standard deviation/dispersion parameters were much higher for the lognormal (95% HDI = [.26, .89]) and shifted-lognormal (95% HDI = [.25, .96]) models relative to the normal model (95% HDI = [-.05, .50]), which demonstrates the importance of our data-generating (distributional) assumptions when making inference on individual differences.

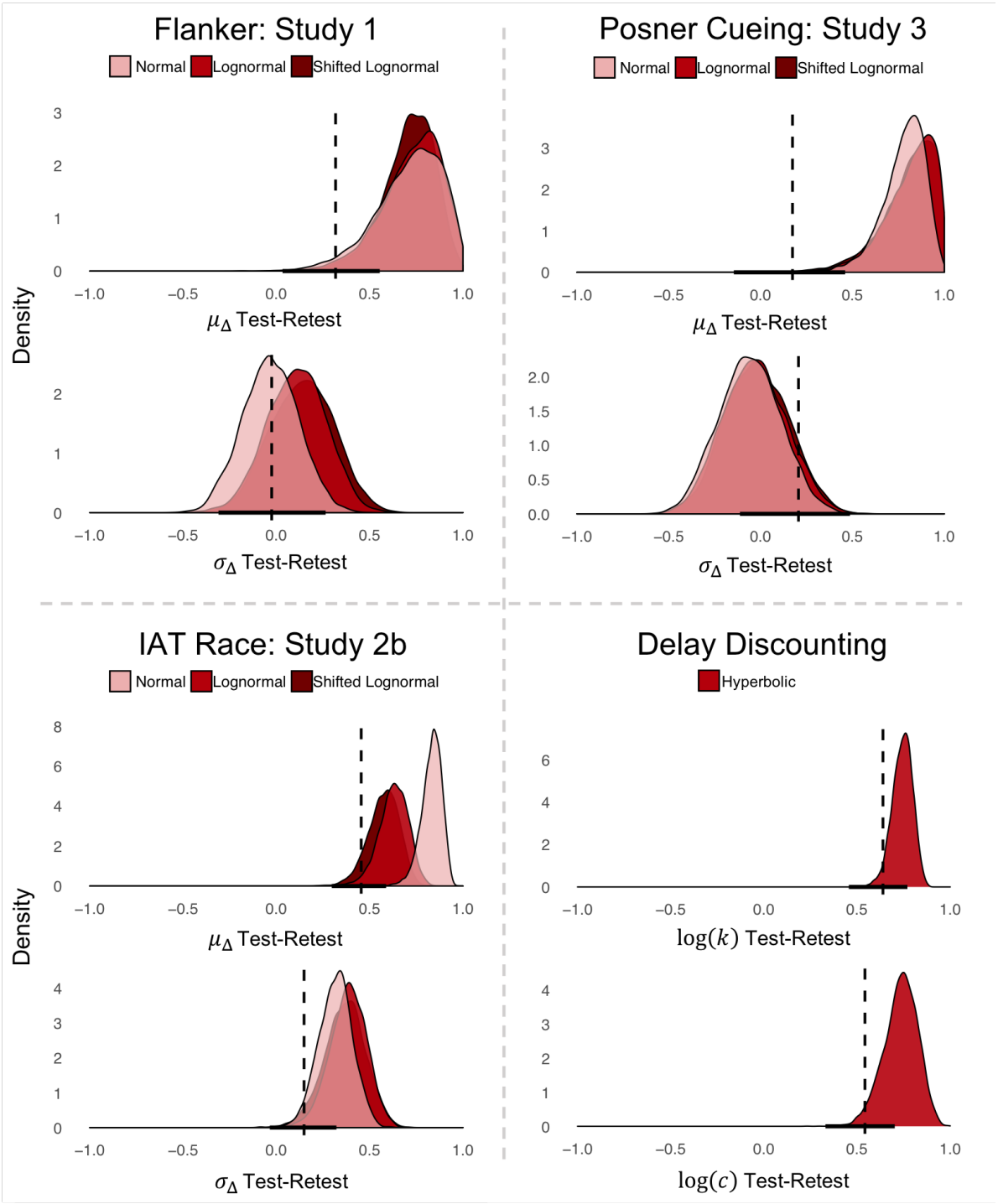
We can also compare the individual-level parameters across models to determine if the models produce different mechanistic inferences. For example, we may be interested in the proportion of participants who show a "Stroop effect" for each model. For demonstration, here we define an effect as when 95% or more of the individual-level posterior distribution on the contrast parameter of interest is greater than 0. We can then identify the proportion of participants meeting this criterion for each of the $\mu_{i,\Delta}$ and $\sigma_{i,\Delta}$ parameters. Across all generative models, all 47 participants showed evidence for an increase in $\mu_{i,\Delta}$ in the incongruent condition. However, for $\sigma_{i,\Delta}$, 36, 31, and 24 participants showed evidence for an increase in the incongruent condition according to the normal, lognormal, and shifted-lognormal models, respectively. This pattern of results suggests that changes in response times across conditions within participants

may be attributable primarily to changes in $\mu_{i,\Delta}$ (difficulty) rather than $\sigma_{i,\Delta}$ (dispersion) —an inference facilitated by the lognormal models.

Figure 6B shows the fitted model predictions compared to the observed response times for a random, representative participant. The two-stage approach is represented simply as the mean response time within each of the congruent and incongruent conditions, whereas the generative model predictions are represented by the light red curves. The light red curves are response time distributions simulated from this participant's estimated individual-level normal, lognormal, and shifted-lognormal model parameters, where variation between lines indicates uncertainty in the underlying parameters. With these simulated response times, we can compare how well each model can reproduce the observed response times. For this particular participant, the normal generative model reveals many shortcomings, the most obvious being the inability to capture right-skew along with the over-prediction of rapid response times. In contrast, the lognormal model in the middle panel provides a much better reproduction of the observed data, capturing both right-skew and the concentration of response times around the mean. The improvement offered by the shifted-lognormal model is more subtle in this example—it better captures the onset of the response time distribution (i.e. the most rapid response times) relative to the lognormal model due to the small shift, but otherwise performs similarly. We provide examples in the online supplement of where the shift makes a more noticeable difference (see Figures S2-S5). Note that the improvement in model fit is accompanied by an increase in expected test-retest reliability for the lognormal models over the normal model, particularly for the dispersion parameters.

Figure 7 visualizes the test-retest correlations for a subset of the remaining tasks, and Table 1 contains descriptive results of the two-stage approach versus generative models for both Study 1 and 2 of the Stroop task from Hedge et al. (2017), along with results for the Flanker and Posner Cueing tasks, the Self-Concept (introversion/extraversion) and Race (Black/White) versions of the IAT, and the delay discounting task. We include detailed results and figures (akin to Figure 6) for each of these tasks in the online supplement (see Figures S2-S6).

Figure 7. Test-retest correlations for the IAT and Flanker, Posner, and delay discounting tasks. The distributions and intervals have the same interpretation as in Figure 6. See Table 1 and the online supplement for more detailed figures and description of each task.



There are three main take-aways from the results presented in Figure 7 and Table 1. First, the generative models consistently inferred higher test-retest correlations relative to the two-stage approach, and in many cases the changes are quite substantial. For example, in study 2 of the Flanker task, the two-stage sample mean contrast test-retest correlation was non-significant at r = -.13, whereas the normal generative model inferred r = .64. For the IAT Race version, the two-stage sample mean contrast test-retest correlation was r = .45, whereas the normal generative model inferred r = .83. Such large differences have strong implications for testing and developing theories of individual differences within each paradigm. Indeed, low test-retest correlations at the individual level in the face of high group-level stability is the central paradox behind a recent influential theoretical advance within social psychology known as the "bias of crowds" (Payne, Vuletich, & Lundberg, 2017; see also Rivers et al., 2017). Attempting to solve this inconsistency led to the argument that IAT scores could be reliably caused by contexts, but do not exist within individual minds (absent specific eliciting contexts). As a result, the IAT is in the midst of a movement from its original conception as a measure of a construct with presumed trait-like qualities (e.g., unchanging) to one that picks up on whatever context an individual mind is currently embedded within (see Jost, 2019). Of note, others have argued that measurement error in the IAT is a more parsimonious solution to the apparent puzzle (e.g., Connor & Evans, 2020). This latter viewpoint is partially supported by our generative model estimates, although there is still variation after accounting for measurement error that could be attributed to state effects or other changes in the underlying construct over time.

Second, the generative model estimates are highly consistent across replications of the same task, whereas the two-stage approach estimates sometimes vary considerably (e.g., compare the two-

stage and generative model estimates for Flanker Study 1 versus Study 2). For example, for the Stroop task, the two-stage standard deviation contrast is significant in study 2 but not in study 1. Similarly, for the Flanker task, the two-stage mean contrast is significant in study 1 but not in study 2. By contrast, the more theoretically informed generative model (i.e. the lognormal models) parameters replicated consistently across studies.

Third, there is variation among the generative models themselves, indicating that test-retest reliability varies—sometimes quite substantially (e.g., compare the normal versus lognormal models for the Stroop task and IAT Race version)—depending on our assumed behavioral model. The variability across models suggests that we should make efforts not to overgeneralize the failings (or successes) of a single behavioral model to the attributes of the behavioral task itself. In other words, we should be explicit in acknowledging that inferences are conditional on a data-generating model and not the task per se.

Table 1. Test-retest results for all tasks and models

Task/Study	Model	Parameter	Estimate	95% Interval
Stroop Study 1	Two-stage Approach	Sample Mean	.50	[.25, .69]
		Sample SD	.07	[-.22, .35]
	Normal	μ_Δ	.76	[.46, 1.00]
		σ_Δ	.23	[-.06, .50]
	Lognormal	μ_Δ	.77	[.47. 1.00]
		σ_Δ	.60	[.26, .89]
	Shifted-Lognormal	μ_Δ	.81	[.53, 1.00]
		σ_Δ	.62	[.25, .96]
Stroop Study 2	Two-stage Approach	Sample Mean	.63	[.45, .76]
		Sample SD	.34	[.10, .55]
	Normal	μ_Δ	.84	[.67, .98]
		σ_Δ	.37	[.15, .60]
	Lognormal	μ_Δ	.82	[.65, 1.00]
		σ_Δ	.48	[.16, .76]
	Shifted-Lognormal	μ_Δ	.75	[.53, .93]
		σ_Δ	.54	[.15, .91]
Flanker Study 1	Two-stage Approach	Sample Mean	.32	[.03, .55]
		Sample SD	-.02	[-.31, .26]
	Normal	μ_Δ	.71	[.38, 1.00]
		σ_Δ	-.03	[-.33, .25]
	Lognormal	μ_Δ	.73	[.42, 1.00]
		σ_Δ	.11	[-.19, .41]
	Shifted-Lognormal	μ_Δ	.71	[.44, .95]
		σ_Δ	.14	[-.18, 47]
Flanker Study 2	Two-stage Approach	Sample Mean	-.13	[-.37, .13]
		Sample SD	.12	[-.14, .36]
	Normal	μ_Δ	.64	[.35, .89]
		σ_Δ	.09	[-.16, 35]
	Lognormal	μ_Δ	.73	[.48, .96]

	Shifted-Lognormal	σ_Δ	.07	[-.22, .37]
		μ_Δ	.74	[.54, .92]
		σ_Δ	.20	[-.13, .51]
Posner Study 3	Two-stage Approach	Sample Mean	.17	[-.15, .46]
		Sample SD	.21	[-.11, .49]
	Normal	μ_Δ	.78	[.55, .98]
		σ_Δ	-.06	[-.39, .26]
	Lognormal	μ_Δ	.81	[.54, 1.00]
		σ_Δ	-.03	[-.36, .31]
	Shifted-Lognormal	μ_Δ	.80	[.52, 1.00]
		σ_Δ	-.01	[-.35, .32]
IAT Self-Concept	Two-stage Approach	Sample Mean	.60	[.49, .69]
		Sample SD	.39	[.25, .52]
	Normal	μ_Δ	.73	[.63, .82]
		σ_Δ	.53	[.42, .65]
	Lognormal	μ_Δ	.69	[.59, .78]
		σ_Δ	.60	[.47, .71]
	Shifted-Lognormal	μ_Δ	.67	[.56, .76]
		σ_Δ	.40	[.21, .58]
IAT Race	Two-stage Approach	Sample Mean	.45	[.30, .59]
		Sample SD	.15	[-.03, .32]
	Normal	μ_Δ	.83	[.73, .93]
		σ_Δ	.32	[.15, .50]
	Lognormal	μ_Δ	.63	[.47, .78]
		σ_Δ	.39	[.19, .58]
	Shifted-Lognormal	μ_Δ	.57	[.42, .74]
		σ_Δ	.37	[.14, .57]
Delay Discounting	Two-stage MLE with Hyperbolic Model	k	.64	[.46, .77]
		c	.54	[.33, .70]
	Hierarchical Bayesian with Hyperbolic Model	k	.74	[.63, .84]
		c	.73	[.55, .90]

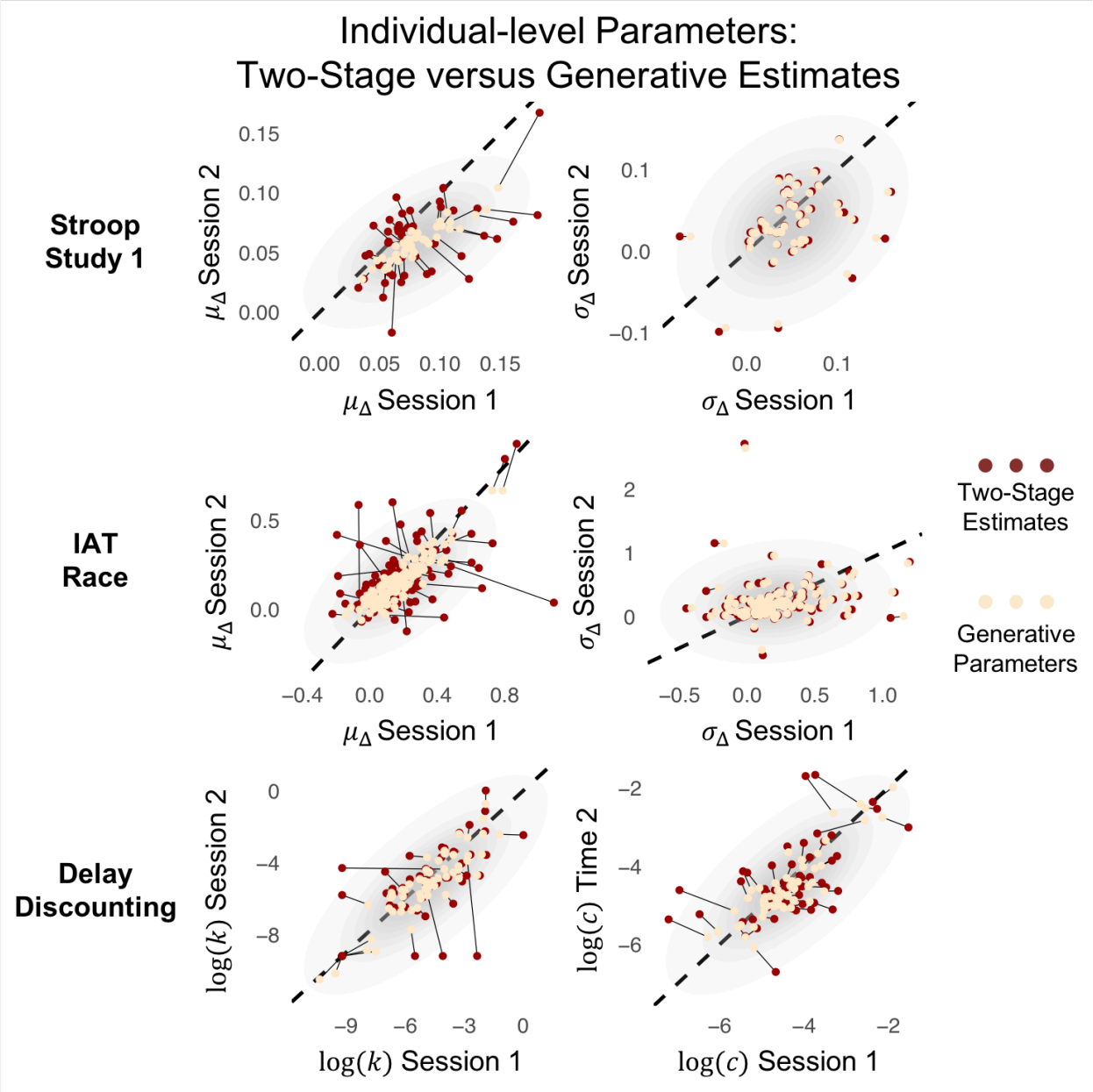
Note. This table contains descriptions of the test-retest correlations for all the tasks analyzed in the current study. 95% intervals indicate the 95% highest density interval for generative models, and the 95% confidence interval for traditional two-stage summary statistic or MLE approaches.

MLE = maximum likelihood estimation. Detailed results on each task and model are presented in the online supplement.

7.6 Comparing Summary Statistics to Generative Model Parameters

It is useful to compare the individual-level estimates of the two-stage approach to those of the generative models to develop an intuition for why test-retest is higher in the generative models. Figure 8 illustrates the differences between approaches. We chose these examples to demonstrate how the *hierarchical pooling* within the generative models affects individual-level parameter estimates in different circumstances. Hierarchical pooling refers to the regression of individual-level parameters toward the group-level mean, which results from Equation's 7-8. In Study 1 of the Stroop task, mean contrast estimates ($\mu_{i,\Delta}$) that would ordinarily be considered outliers in the two-stage approach are pooled toward the group-mean, which produces higher expected test-retest reliability (see Figure 6A and Table 1). The generative model parameter $\mu_{i,\Delta}$ estimates also reveal potential practice effects, whereby almost every participant's expected mean contrast is lower at Session 2 relative to Session 1. Conversely, standard deviation contrast estimates ($\sigma_{i,\Delta}$) show weak pooling in addition to poor expected test-retest reliability, which is reflected in the posterior distribution on the test-retest correlation for the normal model being centered around 0 in Figure 6A. The same general pattern holds in the IAT (Black/White Race version), where $\mu_{i,\Delta}$ and $\sigma_{i,\Delta}$ exhibit strong and weak pooling, respectively. However, pooled $\mu_{i,\Delta}$ estimates for the IAT show regression toward the mean rather than potential practice effects. For the delay discounting task, both discounting rate (k_i) and choice sensitivity (c_i) parameters show moderate pooling (see Figure 7). Taken together, results demonstrate that hierarchical models do not automatically confer higher test-retest reliability—instead, pooling only occurs to the extent it is warranted by data (see also our test-retest parameter recovery results in the online supplement).

Figure 8. Relationship between two-stage estimates and generative model parameters. For the response time models (Stroop & IAT tasks), two-stage estimates are the sample mean and standard deviation contrasts for each participant and retest session (i.e. estimates from the summary statistic approach); generative model parameters are means of the individual-level posterior distributions (i.e., posterior expectations) for each participant. Black lines connect the two-stage estimates and generative model parameters for each participant to demonstrate how the hierarchical model induces regression to the group-level mean. To help visualize the low correlation for the Stroop study, the standard deviation panel is zoomed in and two participants are not shown. For the delay discounting task, two-stage estimates reflect maximum likelihood estimates for each participant's discounting rate (k_i) and choice sensitivity (c_i) parameters; generative model parameters are means of the individual-level posterior distributions for each participant given by the full generative hyperbolic model.


Individual-level Parameters:
Two-Stage versus Generative Estimates
Stroop Study 1
IAT Race
Delay Discounting
μ_Δ Session 2
μ_Δ Session 1
σ_Δ Session 2
σ_Δ Session 1
$\log(k)$ Session 2
$\log(k)$ Session 1
$\log(c)$ Time 2
$\log(c)$ Session 1
Two-Stage Estimates
Generative Parameters

8. Discussion

Generative modeling is a framework that allows for researchers to use background knowledge to inform their statistical models, which, as we have demonstrated, allows for more precise characterization of individual differences in behavior and thereby facilitates theory development in a way not afforded by traditional methods. Our results run counter to mounting claims that behavioral tasks are poorly suited for developing theories of individual differences, which has been (erroneously) attributed to the low test-retest reliability of behavioral measures (e.g., Dang et al., 2020; Enkavi et al., 2019; Gawronski et al., 2017; Hedge et al., 2017; Wennerhold & Friese, 2020). By attending to the data-generating processes underlying behavior, generative modeling offers a solution not only to problems of low reliability (and predictive validity by extension), but also to problems with theory-description gaps arising from the use and overinterpretation of statistical models that fail to instantiate our substantive theories. In contrast, traditional methods of analyzing behavioral data are largely atheoretical—they make implicit data-generating assumptions that researchers seem not to be aware of, and these same assumptions lead to attenuated individual difference correlations and an overall impoverished view of behavioral data. This attenuation occurs through two primary sources. First, researchers rely on behavioral models that do not encode our substantive knowledge and therefore fail to capture important individual-level characteristics (i.e. distributions) of observed behavior. Second, researchers average response times (or other task behaviors) within participants before entering summary scores into secondary statistical models—this two-stage approach inappropriately assumes that the resulting individual-level summary measures contain no measurement error.

8.1 Further Improvements

Several authors have promoted *computational*/*cognitive models of behavior* that are more complex than the models we described. Like the shifted-lognormal model, cognitive models have parameters with theoretically informed interpretations, which makes them ideal for reducing theory-description gaps in social, behavioral, and brain research. However, these models are rarely used due to their complexity and barriers to implementation. Albeit simpler, the generative models of response times and delay discounting that we presented are nonetheless still powerful. We acknowledge, however, that more advanced computational/cognitive models can offer even more advantages toward understanding of mechanisms of behavior and provide further insight into how the models we used could be extended and refined. Readers interested in an extended tutorial can refer elsewhere for descriptions of such models (Guest & Martin, 2020; Heathcote, Poniel, & Mewhort, 1991; Klauer, Voss, Zchmitz, & Teige-Mocigemba, 2007; Jepma, Wagemakers, & Nieuwenhuis, 2012; Johnson, Hopwood, Cesario, & Pleskac, 2017; Voss, Nagler, & Lerche, 2013; White, Ratcliff, & Starns, 2011).

One extension of the response time models presented here is to add mechanisms to account for not only response times but also response accuracy. Estimates of individual differences related to paradigms such as the IAT can be informed by also quantifying joint distributions of correct or incorrect responses and corresponding response times (e.g., Conrey et al, 2005; Klauer et al., 2007). The diffusion decision model has been leveraged to this end (Klauer et al, 2007; Ratcliff et al, 2016), but precisely estimating the full model requires far more data than is ordinarily collected in the IAT. One way to sidestep this problem for practical applications is to use a simpler model, such as the EZ diffusion model (Wagenmakers et al, 2007), and then compare

parameters between conditions (congruent, incongruent). Another more natural extension of the shifted lognormal model is the lognormal race model (Rouder et al., 2014). This model jointly describes choices and response times as arising from competition among independent shifted lognormal accumulation processes for each possible response option.

Another potentially fruitful extension is to directly model cognitive mechanisms underlying the effects of condition manipulations on changes in response time distributions. For example, we modeled condition effects in the Stroop task as simple differences in generative model parameters between congruent and incongruent trials (i.e. $\mu_{i,\Delta,t}$ and $\sigma_{i,\Delta,t}$). However, each stimulus in the task consists of a specific word and color feature, where only one feature should be used to make a response. Presumably, competition between each stimulus feature and corresponding correct responses give rise to observed changes in response times (Cohen et al., 1990). This competition can be modeled with vector space semantic models of cognition, wherein different response options are represented as a mental association between concepts (i.e. psychological similarity). Such models, despite being much more complex than those presented here, offer many potential benefits. For example, they can be used to predict the effect of condition manipulations (e.g., different sets of colors in the Stroop task) on accuracy and response times in decision tasks (Bhatia, 2017; Kvam, 2019b), which makes them well suited to identify correspondence between different behavioral tasks (e.g., Stroop, Flanker). Indeed, many generative and cognitive models are developed to jointly capture phenomena across paradigms—a process that often produces mechanistic insights that are easily obscured when using summary statistics (e.g., Kellen et al., 2016; Luckman et al., 2018; Turner et al., 2018).

It is important to reemphasize that throughout the iterative process of generative model development, model parameters can be assessed to determine their psychometric properties. Although we focused on test-retest reliability, there are many other properties worth exploring including parameter identifiability (e.g., Spektor & Kellen, 2018), parameter recovery (e.g., Ahn et al., 2011; Haines et al., 2018; Miletić, Turner et al., 2017), tests of selective influence (a form of construct validity where experimental manipulations cause expected changes in parameter values; Criss, 2010), and parameter convergence between behavioral models and models derived at other levels of analysis (e.g., with trait or neural models; Haines et al., 2020; Turner et al., 2017). Bayesian analysis facilitates joint estimation of all model parameters and their hypothesized relations, thus allowing for proper calibration of uncertainty in key parameters (e.g., test-retest reliability).

8.2 Benefits of Building Better Explanations

We hope the previous section has made it clear that the landscape for building and refining generative models is vast, whereas typical summary statistics approaches are inherently limited. Generative modeling is thus especially appealing for improving mechanistic inferences about complex human behaviors across social, behavioral, and brain sciences. To summarize, generative models offer several key advantages over the two-stage summary statistics approach:

1. Generative models require explicit mechanistic assumptions, minimizing the theory-description gap. This facilitates theory development and principled abduction of competing hypotheses.
2. Generative models use all available data, increasing precision of individual-level (person-specific) parameter estimates when we have limited data at the individual level.

3. Generative models appropriately calibrate uncertainty in parameters (e.g., test-retest reliability) regardless of sample size, thus allowing results to be interpreted more confidently.

Although this list is non-exhaustive, it shows that generative modeling offers solutions to many recent critiques set forth regarding theory development and research as typically practiced in the social, behavioral, and brain sciences, including both (1) low measurement reliability (Chen et al., 2015; Elliott et al., 2020; Enkavi et al., 2019; Gawronski et al., 2017; Hedge et al., 2017; Noble et al., 2019), and (2) theory-description gaps arising from the mis-specification, mis-application, and mis-interpretation of statistical models, concepts, and effects (Corneille & Hütter, 2020; Devezer et al., 2019; 2020; Muthukrishna, & Henrich, 2019; Regenwetter & Robinson, 2017; Ross et al., 2020; Rotello et al., 2014; Szollosi & Donkin, 2019).

Advances in computational statistics have only recently made generative modeling widely accessible. We anticipate that generative modeling will proliferate as scientists from all backgrounds recognize their utility for rigorous theory development and testing. There are now many accessible resources and software packages available to help researchers gain a deeper understanding of generative modeling to apply it to their own work. These include introductions to the philosophy and utility of generative or computational modeling for theory development (e.g., Guest & Martin, 2020; van Rooij & Baggio, 2020), tutorials on building your own generative models from first principles (e.g., Wilson & Collins, 2019; van Rooij & Blokpoel, 2020), practical textbooks that combine introductions to both behavioral model development and hierarchical Bayesian modeling (Farrell & Lewandowsky, 2018; Lee & Wagenmakers, 2013), tutorials and case examples on developing joint generative models of behavior and brain activity

(Palestro et al., 2018; Turner et al., 2019), and open source R and Python software packages that allow beginners and advanced users alike to apply popular generative models of behavioral to their own data using hierarchical Bayesian modeling (e.g., Ahn et al, 2017; Mathys et al., 2014; Wiecki et al., 2013).

We recognize that a shift toward generative modeling requires an investment in resources and statistics training not typical of the social, behavioral, and brain sciences. However, even for those who do not take up generative modeling themselves, there are two actionable steps that any researcher can take to facilitate generative modeling:

1. Make raw behavioral data (i.e. trial- and item-level choices and response times) openly available, and
2. Make an effort to not overgeneralize results obtained from behavioral data that are not analyzed with an underlying generative model.

Although the benefits of (1) are quite straightforward, we emphasize (2) here because many of the criticisms regarding the utility of behavioral tasks for individual differences researchers relied on summary statistics alone, and as we have shown here these criticisms are overly general.

8.3 Future Directions

A final benefit of generative modeling is that it facilitates development of adaptive experimental designs that maximize informativeness of behavioral tasks. Staircasing procedures for behavioral measures (Cornsweet, 1962) and computer adaptive testing for self-report questionnaires or tests are examples, where stimuli are selected from a set of candidates based on behavioral responses

collected during the experiment. For staircasing, stimuli might vary along a particular dimension (e.g., delay until reward in a delay discounting task), and stimulus choice on the next trial follows a simple rule determined by a participant's prior response: increase the delay if the they choose the larger later reward; otherwise, decrease the delay. Computer adaptive testing, on the other hand, first quantifies properties of stimuli (e.g., item difficulty), and then proposes items based on a test taker's performance. For example, if a student answers the first few questions correctly on an exam, following up with more difficult questions is more informative for estimating the student's ability.

Adaptive design optimization is a powerful extension of traditional adaptive designs (ADO; Cavagnaro et al, 2011; Myung et al., 2013). With ADO, the simple adaptive rule used in staircasing is replaced by an algorithm that includes all possible stimuli that can be presented in the experiment *and* a generative model of the decision process (most often a cognitive model). Rather than presenting all stimuli equally often, or following some other heuristic method, the model-based algorithm in ADO guides selection of experimental stimuli toward regions in the design space that are the most informative for reducing uncertainty about parameters of the generative model. In the context of the generative models developed above, the objective of ADO is to optimize parameter estimation, which can substantially improve reliability. For example, in delay discounting tasks, ADO can identify combinations of rewards and delays that optimize estimation of discounting rates, achieving test-retest reliabilities greater than $r = .95$ in fewer than 20 trials (Ahn et al., 2020). ADO is currently underutilized in most areas of behavioral science, but user-friendly software packages are now available (Yang et al., in press).

9. Concluding Remarks

Collecting data with high precision throughout the social, behavioral, and brain sciences is often laborious, expensive, and time consuming—particularly in areas of research centered on at-risk, difficult to study populations or those that rely on expensive technologies such as MRI. It is therefore surprising that the most common analysis methods *throw away* useful information, often at the expense of theory. The two-stage summary statistic approach, which reduces behavioral data to point estimates and then uses these estimates in a second statistical model, produces parameters that are imprecise and often difficult to interpret vis-à-vis substantive theory. Consequences include overconfidence in estimates that are attenuated by measurement error (e.g., confidence intervals too narrow due to ignoring individual-level measurement error), incorrect inferences, and failures to replicate—all of which decrease the informativeness of research on individual differences. Although the two-stage approach has historically sufficed for making some group-level inferences (e.g., when comparing group means), it is problematic when used to make inferences about individual differences.

By contrast, generative modeling facilitates development of theory-informed models of behavior and takes advantage of all the information available, thereby improving accuracy of inference even when using smaller samples of behavioral (or neural) data. Our results question conclusions drawn from previous studies on reliability of various behavioral and neural measures—particularly those that relied on two-stage approaches (e.g., Chen et al., 2015; Elliott et al., 2020; Enkavi et al., 2019; Gawronski et al., 2017; Hedge et al., 2017; Klein, 2020; Noble et al., 2019). Future work could both extend the models presented here to explore idiosyncrasies in behavioral tasks in addition to relations between behavioral constructs and variables specified at other levels

of analysis (e.g., trait and neural measures), thus minimizing theory-description gaps. Sufficiently refined generative models can begin to take advantage of methods such as ADO that can improve the informativeness and efficiency of experiments even further.

We end with a cautionary yet hopeful note: as history has revealed, heuristic use of summary statistics absent a generative model can and will lead us astray. Although our generative models may be wrong or mis-specified, they are at least explicit, forcing us to specify our assumptions regarding how behavior arises. By embracing their incompleteness, we can strive to build generative models that are precise and thus meaningfully incorrect, rather than relying on heuristic models that are ambiguous and only circumstantially interpretable. Knowing where our assumptions are wrong then provides a natural path toward deepening our understanding of the mechanisms underlying behavior.

Funding

T.P.B. was supported in part by Grants UH2DE025980 and UL1TR002733 from the National Institutes of Health, and B.M.T. was supported by a CAREER award from the National Science Foundation.

Conflicts of Interest Statement

The authors declare no conflicts of interest.

Availability of data and materials

All de-identified data along with the R and Stan codes used to reproduce our results and figures are available on our GitHub repository (https://github.com/Nathaniel-Haines/Reliability_2020).

References

Ahn, W.-Y., & Busemeyer, J. R. (2016). Challenges and promises for translating computational tools into clinical practice. *Current Opinion in Behavioral Sciences* 11, 1-7: doi:10.1016/j.cobeha.2016.02.001

Ahn, W.-Y., Gu, H., Shen, Y., Haines, N., Teater, J. E., Myung, J. I., & Pitt, M. A. (2020). Rapid, precise, and reliable measurement of delay discounting using a Bayesian learning algorithm. *Scientific Reports.* Manuscript accepted for publication.

Ahn, W.-Y., Haines, N., & Zhang, L. (2017). Revealing neuro-computational mechanisms of reinforcement learning and decision-making with the hBayesDM package. *Computational Psychiatry*, 1: 24-57. doi:10.1162/CPSY_a_00002

Ahn, W.-Y., Krawitz, A., Kim, W., Busemeyer, J. R., & Brown, J. W. (2011). A model-based fMRI analysis with hierarchical Bayesian parameter estimation. *Journal of Neuroscience, Psychology, and Economics* 4: 95-110. doi:10.1037/a0020684

Bahg, G., Evans, D., Galdo, M., and Turner, B. M. (in press). Gaussian process linking functions for mind, brain, and behavior. In press at Proceedings of the National Academy of Sciences.

Beauchaine, T. P., & Hinshaw, S. P. (2020). RDoC and psychopathology among youth: Misplaced assumptions and an agenda for future research. *Journal of Clinical Child and Adolescent Psychology,* 49: 322-340. doi:10.1080/15374416.2020.1750022

Bhatia, S. (2017). Associative judgment and vector space semantics. *Psychological Review*, 124: 1-20. doi:10.1037/rev0000047

Bosson, J. K., Swann, W. B., & Pennebaker, J. W. (2000). Stalking the perfect measure of implicit self-esteem: The blind men and the elephant revisited? *Journal of Personality*

and Social Psychology, 79: 631–643. doi:10.1037/0022-3514.79.4.631

Busemeyer, J. R., Gluth, S., Rieskamp, J., & Turner, B. M. (2019). Cognitive and neural bases of multi-attribute, multi-alternative, value-based decisions. *Trends in Cognitive Sciences*, 23: 251-263. doi:10.1016/j.tics.2018.12.003

Carpenter, B., Gelman, A., Hoffman, M., & Lee, D. (2016). Stan: A probabilistic programming language. *Journal of Statistical Software* 76: doi:10.18637/jss.v076.i01.

Cavagnaro, D. R., Pitt, M. A., & Myung, J. I. (2011). Model discrimination through adaptive experimentation. *Psychonomic Bulletin and Review* 18: 204-210. doi:10.3758/s13423-010-0030-4

Chen, B., Xu, T., Zhou, C., Wang, L., Yang, N., Wang, Z., …Weng, X.-C. (2015). Individual variability and test-retest reliability revealed by ten repeated resting-state brain scans over one month. *PLoS ONE* 10: e0144963. doi:10.1371/journal.pone.0144963

Cohen, J. D., Dunbar, K. & McClelland, J. L. (1990). On the control of automatic processes: A parallel-distributed processing account of the Stroop effect. *Psychological Review* 97: 332-361.

Connor, P., & Evers, E. R. (2020). The Bias of Individuals (in Crowds): Why Implicit Bias Is Probably a Noisily Measured Individual-Level Construct. *Perspectives on Psychological Science*. doi:10.1177/1745691620931492

Conrey, F. R., Sherman, J. W., Gawronski, B., Hugenberg, K., & Groom, C. J. (2005). Separating multiple processes in implicit social cognition: The quad model of implicit task performance. *Journal of Personality and Social Psychology* 89: 469-487. doi:10.1037/0022-3514.89.4.469

Corneille, O., & Hütter, M. (2020). Implicit? What do you mean? A comprehensive review of the delusive implicitness construct in attitude research. *Personality and Social Psychology Review* 24: 212-232. doi:10.1177/1088868320911325

Cornsweet, T. N. (1962). The staircase-method in psychophysics. *American Journal of Psychology* 75: 485-491. doi:10.2307/1419876

Craigmile, P.F., Peruggia, M., & Van Zandt, T. (2010). Hierarchical Bayes models for response time data. *Psychometrika* 75: 613-632. doi:10.1007/s11336-010-9172-6

Criss, A. H. (2010). Differentiation and response bias in episodic memory: Evidence from reaction time distributions. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 36: 484-499. doi:10.1037/a0018435

Cyders, M., & Coskunpinar, A. (2011). Measurement of constructs using self-report and behavioral lab tasks: Is there overlap in nomothetic span and construct representation for impulsivity? *Clinical Psychology Review* 31: 965-982. doi:10.1016/j.cpr.2011.06.001

Dang, J., King, K. M., & Inzlicht, M. (2020). Why Are self-report and behavioral measures weakly correlated? *Trends in Cognitive Sciences* 1-3. doi:10.1016/j.tics.2020.01.007

Davis-Stober, C. P., Park, S., Brown, N., & Regenwetter, M. (2016). Reported violations of rationality may be aggregation artifacts. *Proceedings of the National Academy of Sciences* 113: E4761–E4763. doi:10.1073/pnas.1606997113

Devezer, B., Nardin, L. G., Baumgaertner, B., & Buzbas, E. O. (2019). Scientific discovery in a model-centric framework: Reproducibility, innovation, and epistemic diversity. *PLoS ONE*, 14: e0216125. doi:10.1371/journal.pone.0216125

Devezer, B., Navarro, D. J., Vandekerckhove, J., & Buzbas, E. O. (2020). The case for formal methodology in scientific reform. *bioRxiv.* Manuscript under review.

doi:10.1101/2020.04.26.048306

Duckworth, A. L., & Kern, M. L. (2011). A meta-analysis of the convergent validity of self-control measures. *Journal of Research in Personality* 45: 259-268. doi:10.1016/j.jrp.2011.02.004

Elliott, M. L., Knodt, A. R., Ireland, D., Morris, M. L., Ramrakha, S., Sison, M. L., …Hariri, A. R. (2020). Poor test-retest reliability of task-fMRI: New empirical evidence and a meta-analysis. *Psychological Science* 919: 1-31. doi:10.1177/0956797620916786

Enkavi, A. Z., Eisenberg, I. W., Bissett, P. G., Mazza, G. L., MacKinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2019). Large-scale analysis of test–retest reliabilities of self-regulation measures. *Proceedings of the National Academy of Sciences* 116: 5472-5477. doi:10.1073/pnas.1818430116

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics* 16: 143-149. doi:10.3758/BF03203267

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin* 53: 134-140. doi:10.1037/h0045156

Farrell, S., & Lewandowsky, S. (2018). *Computational modeling of cognition and behavior*. New York, NY: Cambridge University Press.

Friston, K. J., Stephan, K. E., Montague, R., & Dolan, R. J. (2014). Computational psychiatry: The brain as a phantastic organ. *The Lancet Psychiatry* 1: 148-158. doi:10.1016/S2215-0366(14)70275-5

Gawronski, B., Morrison, M., Phills, C. E., & Galdi, S. (2017). Temporal stability of implicit and explicit measures. *Personality and Social Psychology Bulletin* 43: 300-312.

doi:10.1177/0146167216684131

Gelman, A., & Carlin, J. (2014). Beyond power calculations assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science* 9: 641-651. doi:10.1177/1745691614551642

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd). London, UK: Chapman Hall.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press.

Gomez, P., Ratcliff, R., & Childers, R. (2015). Pointing, looking at, and pressing keys: A diffusion model account of response modality. *Journal of Experimental Psychology: Human Perception and Performance* 41: 1515-1523. doi:10.1037/a0039653

Green, L., & Myerson, J. (2004). A discounting framework for choice with delayed and probabilistic rewards. *Psychological Bulletin* 130: 769-792. doi:10.1037/0033-2909.130.5.769

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology* 74: 1464-1480. doi:10.1037/0022-3514.74.6.1464

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology* 85: 197-216. doi:10.1037/0022-3514.85.2.197

Guest, O., & Martin, A. E. (2020). How computational modeling can force theory building in psychological science. *PsyArXiv preprint*: 1-13. doi:10.31234/osf.io/rybh9

Haaf, J. M., & Rouder, J. N. (2017). Developing constraint in Bayesian mixed models.

Psychological Methods 22: 779-798. doi:10.1037/met0000156

Haines, N., Beauchaine, T. P., Galdo, M., Rogers, A. H., Hahn, H., Pitt, M. A., ... & Ahn, W. Y. (2020). Anxiety modulates preference for immediate rewards among trait-impulsive individuals: A hierarchical Bayesian analysis. *Clinical Psychological Science*. manuscript accepted for publication.

Haines, N., Vassileva, J., & Ahn, W.-Y. (2018). The outcome-representation learning model: A novel reinforcement learning model of the Iowa Gambling Task. *Cognitive Science* 47: 1-28. doi:10.1111/cogs.12688

Hand, D. J. (1996). Statistics and the theory of measurement. *Journal of the Royal Statistical Society Series A (Statistics in Society)* 159: 445-473. doi:10.2307/2983326

Heathcote, A., & Love, J. (2012). Linear deterministic accumulator models of simple choice. *Frontiers in Cognitive Science* 3: 292. doi:10.3389/fpsyg.2012.0029.

Heathcote, A., Popiel, S. J., & Mewhort, D. J. (1991). Analysis of response time distributions: an example using the Stroop task. *Psychological Bulletin* 109: 340-347. doi:10.1037/0033-2909.109.2.340

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin and Review* 7: 185-207. doi:10.3758/BF03212979

Hedge, C., Powell, G., & Sumner, P. (2017). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods* 103: 1-21. doi:10.3758/s13428-017-0935-1

Hempel, C., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science* 15: 135-175. doi:10.1086/286983

Hockley, W. E., & Corballis, M. C. (1982). Tests of serial scanning in item recognition. *Canadian Journal of Psychology/Revue Canadienne de Psychologie* 36: 189-212. doi:10.1037/h0080637

Huys, Q. J., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature neuroscience* 19: 404-413. doi:10.1038/nn.4238

Jarecki, J., Tan, J. H., & Jenny, M. A. (2020). A framework for building cognitive process models. *Psychonomic Bulletin and Review*. advance online publication. doi:10.3758/s13423-020-01747-2

Jepma, M., Wagenmakers, E. J., & Nieuwenhuis, S. (2012). Temporal expectation and information processing: A model-based analysis. *Cognition* 122: 426-441. doi:10.1016/j.cognition.2011.11.014

Johnson, J. G., & Busemeyer, J. R. (2005). A dynamic, stochastic, computational model of preference reversal phenomena. *Psychological Review* 112: 841-861. doi:10.1037/0033-295X.112.4.841

Johnson, D. J., Hopwood, C. J., Cesario, J., & Pleskac, T. J. (2017). Advancing research on cognitive processes in social and personality psychology: A hierarchical drift diffusion model primer. *Social Psychological and Personality Science* 8: 413-423. doi:10.1177/1948550617703174

Jost, J. T. (2019). The IAT is dead, long live the IAT: Context-sensitive measures of implicit attitudes are indispensable to social and political psychology. *Current Directions in Psychological Science* 28: 10-19. doi:10.1177/0963721418797309

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in

social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology* 103: 54-69. doi:10.1037/a0028347

Kellen, D. (2019). A model hierarchy for psychological science. *Computational Brain and Behavior* 2: 160–165. doi:10.1007/s42113-019-00037-y

Kellen, D., Pachur, T., & Hertwig, R. (2016). How (in)variant are subjective representations of described and experienced risk and rewards? 157: 126-138. doi:10.1016/j.cognition.2016.08.020

Kennedy, L., & Gelman, A. (2019). Know your population and know your model: Using model-based regression and poststratification to generalize findings beyond the observed sample. *arXiv preprint*.

Klauer, K. C., Voss, A., Schmitz, F., & Teige-Mocigemba, S. (2007). Process components of the Implicit Association Test: A diffusion-model analysis. *Journal of Personality and Social Psychology* 93: 353-368. doi:10.1037/0022-3514.93.3.353

Klein, C. (2020). Confidence intervals on Implicit Association test scores are really rather large. *PsyArXiv preprint*. doi:10.31234/osf.io/5djkh

Kruschke, J. K. (2015). *Doing Bayesian data analysis* (2nd ed.). New York, NY: Academic Press.

Kurdi, B., Seitchik, A. E., Axt, J. R., Carroll, T. J., Karapetyan, A., Kaushik, N., Tomezsko, D., Greenwald, A. G., & Banaji, M. R. (2019). Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis. *American Psychologist* 74: 569-586. doi:10.1037/amp0000364

Kvam, P. D. (2019a). Modeling accuracy, response time, and bias in continuous orientation

judgments. *Journal of Experimental Psychology: Human Perception and Performance* 45: 301-318. doi:10.1037/xhp0000606

Kvam, P. D. (2019b). A geometric framework for modeling dynamic decisions among arbitrarily many alternatives. *Journal of Mathematical Psychology* 91: 14-37. doi:10.1016/j.jmp.2019.03.001

Kvam, P. D., & Busemeyer, J. R. (2020). A distributional and dynamic theory of pricing and preference. *Psychological Review.* Advance online publication. doi:10.1037/rev0000215

Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology* 55: 1-7. doi:10.1016/j.jmp.2010.08.013

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling*. Cambridge, UK: Cambridge University Press.

Leth-Steensen, C., Elbaz, Z. K., & Douglas, V. I. (2000). Mean response times, variability, and skew in the responding of ADHD children: A response time distributional approach. *Acta Psychologica* 104: 167-190. doi:10.1016/s0001-6918(00)00019-6

Liew, S. X., Howe, P. D. L., & Little, D. R. (2016). The appropriacy of averaging in the study of context effects. *Psychonomic Bulletin and Review* 23: 1639-1646. doi:10.3758/s13423-016-1032-7

Luckman, A., Donkin, C., & Ben R Newell. (2017). Can a single model account for both risky choices and inter-temporal choices? Testing the assumptions underlying models of risky inter-temporal choice. *Psychonomic Bulletin and Review* 25: 785-792. doi:10.3758/s13423-017-1330-8

Ly, A., Boehm, U., Heathcote, A., Turner, B. M., Forstmann, B., Marsman, M., & Matzke, D. (2017). A flexible and efficient hierarchical Bayesian approach to the exploration of

individual differences in cognitive-model-based neuroscience. In A. A. Moustafa (Ed.), *Computational models of brain and behavior* (pp. 467-480). Hoboken, NJ: Wiley Blackwell.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin* 109: 163-203. doi:10.1037/0033-2909.109.2.163

Mathys, C. D., Lomakina, E. I., Daunizeau, J., Iglesias, S., Brodersen, K. H., Friston, K. J., & Stephan, K. E. (2014). Uncertainty in perception and the Hierarchical Gaussian Filter. *Frontiers in Human Neuroscience* 8: 825. doi:10.3389/fnhum.2014.00825

Mazur, J. E. (1987). *An adjusting procedure for studying delayed reinforcement.* In M. L. Commons, J. E. Mazur, J. A. Nevin, & H. Rachlin (Eds.), *Quantitative analyses of behavior, Vol. 5. The effect of delay and of intervening events on reinforcement value* (p. 55–73). Lawrence Erlbaum Associates, Inc.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science* 34: 103-115. doi.org/10.1086/288135

Michell, J. (2008). Is psychometrics pathological science? *Measurement: Interdisciplinary Research and Perspective* 6: 7-24. doi:10.1080/15366360802035489

Miletić, S., Turner, B. M., Forstmann, B. U., & Van Maanen, L. (2017). Parameter recovery for the leaky competing accumulator model. *Journal of Mathematical Psychology* 76: 25-50. doi:10.1016/j.jmp.2016.12.001

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in cognitive sciences* 16: 72-80. doi:10.1016/j.tics.2011.11.018

Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour* 3: 221-229. doi:10.1038/s41562-018-0522-1

Myung, J. I., Cavagnaro, D. R., & Pitt, M. A. (2013). A tutorial on adaptive design optimization. *Journal of Mathematical Psychology* 57: 53-67. doi:10.1016/j.jmp.2013.05.005.

Navarro, D. J. (2018). Between the devil and the deep blue sea: Tensions between scientific judgement and statistical model selection. *Computational Brain and Behavior* 2: 28-34. doi:10.1007/s42113-018-0019-z

Navarro, D. J. (2020). If mathematical psychology did not exist we would need to invent it: A case study in cumulative theoretical development. *PsyArXiv preprint*. doi:10.31234/osf.io/ygbjp

Noble, S., Scheinost, D., & Constable, R. T. (2019). A decade of test-retest reliability of functional connectivity: A systematic review and meta-analysis. *Neuroimage* 203: 116157. doi:10.1016/j.neuroimage.2019.116157

Pagan, A. (1984). Econometric issues in the analysis of regressions with generated regressors. *International Economic Review* 25: 221-247. doi:10.2307/2648877

Palestro, J. J., Bahg, G., Sederberg, P. B., Lu, Z.-L., Steyvers, M., & Turner, B. M. (2018). A tutorial on joint models of neural and behavioral measures of cognition. *Journal of Mathematical Psychology* 84: 20-48. doi:10.1016/j.jmp.2018.03.003

Parsons, S. (2020). Exploring reliability heterogeneity with multiverse analyses: Data processing decisions unpredictably influence measurement reliability. *PsyArXiv preprint*. doi:10.31234/osf.io/y6tcz

Payne, B. K., Vuletich, H. A., & Lundberg, K. B. (2017). The bias of crowds: How implicit bias bridges personal and systemic prejudice. *Psychological Inquiry* 28: 233-248. doi:10.1080/1047840X.2017.1335568

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology* 32: 3-25. doi:10.1080/00335558008248231

Rangel, A., Camerer, C. F., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience* 9: 545-556. doi:10.1038/nrn2357

Regenwetter, M., & Robinson, M. M. (2017). The construct–behavior gap in behavioral decision research: A challenge beyond replicability. *Psychological Review* 124: 533-550. doi:10.1037/rev0000067

Rivers, A. M., Rees, H. R., Calanchini, J., & Sherman, J. W. (2017). Implicit bias reflects the personal and the social. *Psychological Inquiry* 28: 301-305. doi:10.1080/1047840X.2017.1373549

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review* 107: 358-367. doi:10.1037/0033-295X.107.2.358

Romeu, R. J., Haines, N., Ahn, W.-Y., Busemeyer, J. R., & Vassileva, J. (2019). A computational model of the Cambridge Gambling Task with applications to substance use disorders. *Drug and Alcohol Dependence* 206: 107711. doi:10.1016/j.drugalcdep.2019.107711

Rotello, C. M., Heit, E., & Dubé, C. (2014). When more data steer us wrong: Replications with the wrong dependent measure perpetuate erroneous conclusions. *Psychonomic Bulletin and Review* 22: 944-954. doi:10.3758/s13423-014-0759-2

Rouder, J.N., & Haaf, J.M. (2019). A psychometrics of individual differences in experimental tasks. *Psychonomic Bulletin and Review* 26: 452-467. doi:10.3758/s13423-018-1558-y

Rouder, J., Kumar, A., & Haaf, J. M. (2019). Why most studies of individual differences with inhibition tasks are bound to fail. *PsyArXiv preprint*. doi:10.31234/osf.io/3cjr5

Rouder, J.N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin and Review* 12: 573-604. doi:10.3758/BF03196750

Rouder, J. N., Province, J. M., Morey, R. D., Gomez, P., & Heathcote, A. (2014). The lognormal race: A cognitive-process model of choice and latency with desirable psychometric properties. *Psychometrika* 80: 491-513. doi:10.1007/s11336-013-9396-3

Schimmack, U. (2019). The Implicit Association Test: A method in search of a construct. *Perspectives on Psychological Science*. Advance online publication. doi:10.1177/1745691619863798

Shiffrin, R.M., Lee, M.D., Kim, W., & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science* 32: 1248-1284. doi:10.1080/03640210802414826

Spektor, M. S., & Kellen, D. (2018). The relative merit of empirical priors in non-identifiable and sloppy models: Applications to models of learning and decision-making. *Psychonomic Bulletin and Review* 4: 1-22. doi:10.3758/s13423-018-1446-5

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology* 18: 643-662. doi:10.1037/h0054651

Suppes, P. (1966). Models of data. *Studies in Logic and the Foundations of Mathematics* 44: 252-261. doi:10.1016/S0049-237X(09)70592-0

Szollosi, A., & Donkin, C. (2019). Neglected sources of flexibility in psychological theories: From replicability to good explanations. *Computational Brain and Behavior* 2: 190-192.

doi:10.1007/s42113-019-00045-y

Tong, C. (2019). Statistical inference enables bad science; Statistical thinking enables good science. *The American Statistician* 73: 246–261. doi:10.1080/00031305.2018.1518264

Townsend, J. T. (2008). Mathematical psychology: Prospects for the 21st century: A guest editorial. *Journal of Mathematical Psychology* 52: 269-280. doi:10.1016/j.jmp.2008.05.001

Turner, B. M., Forstmann, B. U., Wagenmakers, E. J., Brown, S. D., Sederberg, P. B., and Steyvers, M. (2013). A Bayesian framework for simultaneously modeling neural and behavioral data. NeuroImage. 72, 193-206.

Turner, B. M., Forstmann, B. U., Love, B. C., Palmeri, T. J., & Van Maanen, L. (2017). Approaches to analysis in model-based cognitive neuroscience. *Journal of Mathematical Psychology* 76: 65-79. doi:10.1016/j.jmp.2016.01.001

Turner, B. M., Forstmann, B. U., & Steyvers, M. (2019). *Joint models of neural and behavioral data*. Springer International Publishing.

Turner, B. M., Schley, D. R., Muller, C., & Tsetsos, K. (2018). Competing theories of multialternative, multiattribute preferential choice. *Psychological Review* 125: 329-362. doi:10.1037/rev0000089

van Rooij, I., & Baggio, G. (2020). Theory before the test: How to build high-verisimilitude explanatory theories in psychological science. *PsyArXiv preprint*. doi:10.31234/osf.io/7qbpr

van Rooij, I., & Blokpoel, M. (2020). Formalizing verbal theories: A tutorial by dialogue. *PsyArXiv preprint*. doi:10.31234/osf.io/r2zqy

Vandekerckhove, J. (2014). A cognitive latent variable model for the simultaneous analysis

of behavioral and personality data. *Journal of Mathematical Psychology* 60: 58-71. doi:10.1016/j.jmp.2014.06.004

Voss, A., Nagler, M., & Lerche, V. (2013). Diffusion models in experimental psychology: A practical introduction. *Experimental Psychology* 60: 385-402. doi:10.1027/1618-3169/a000218

Wagenmakers, E.-J., & Brown, S. (2007). On the linear relation between the mean and the standard deviation of a response time distribution. *Psychological Review* 114: 830-841. doi:10.1037/0033-295X.114.3.830

Wagenmakers, E. J., Van Der Maas, H. L., & Grasman, R. P. (2007). An EZ-diffusion model for response time and accuracy. *Psychonomic Bulletin and Review* 14: 3-22. doi:10.3758/BF03194023

Wennerhold, L., & Friese, M. (2020). Why self-report measures of self-control and inhibition tasks do not substantially correlate. *Collabra: Psychology* 6: 9. doi:10.1525/collabra.276

Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs Is harder than you think. *PLOS ONE* 11: e0152719. doi:10.1371/journal.pone.0152719

Wiecki, T. V., Poland, J., & Frank, M. J. (2015). Model-based cognitive neuroscience approaches to computational psychiatry: Clustering and classification. *Clinical Psychological Science*, 3: 378-399. doi:10.1177/2167702614565359

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. *Frontiers in Neuroinformatics* 7: 14. doi:10.3389/fninf.2013.00014

Williams, D. R., Zimprich, D. R. & Rast, P. (2019) A Bayesian nonlinear mixed-effects

location scale model for learning. *Behavioral Research Methods* 51: 1968-1986. doi:10.3758/s13428-019-01255-9

Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *eLife* 8: 558. doi:10.7554/eLife.49547

Whelan, R. (2008). Effective analysis of reaction time data. *Psychological Record* 58: 475-482. doi:10.1007/BF03395630

White, C. N., Ratcliff, R., & Starns, J. J. (2011). Diffusion models of the Flanker Task: Discrete versus gradual attentional selection. *Cognitive Psychology* 63: 210-238. doi:10.1016/j.cogpsych.2011.08.001

Wolsiefer, K., Westfall, J., & Judd, C. M. (2017). Modeling stimulus variation in three common implicit attitude tasks. *Behavior Research Methods* 49: 1193-1209. doi:10.3758/s13428-016-0779-0

Yang, J., Pitt, M. A., Ahn, W.-Y., & Myung, J. I. (2020). ADOpy: A Python package for adaptive design optimization. *In press at Behavior Research Methods*. doi:10.31234/osf.io/mdu23